Transactions on Graph Data and Knowledge

Volume 3 | Issue 2 | October, 2025

ISSN 2942-7517

ACM Classification 2012
Computing methodologies → Reasoning about belief and knowledge; Mathematics of computing → Discrete mathematics; Information systems → Specialized information retrieval

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany.
Online available at
https://www.dagstuhl.de/dagpub/2942-7517.

Publication date
October, 2025

Digital Object Identifier
10.4230/TGDK.3.2.0

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at https://dnb.d-nb.de.

Aims and Scope
Transactions on Graph Data and Knowledge (TGDK) is an Open Access journal that publishes original research articles and survey articles on graph-based abstractions for data and knowledge, and the techniques that such abstractions enable with respect to integration, querying, reasoning and learning. The scope of the journal thus intersects with areas such as Graph Algorithms, Graph Databases, Graph Representation Learning, Knowledge Graphs, Knowledge Representation, Linked Data and the Semantic Web. Also in-scope for the journal is research investigating graph-based abstractions of data and knowledge in the context of Data Integration, Data Science, Information Extraction, Information Retrieval, Machine Learning, Natural Language Processing, and the Web.

The journal is Open Access without fees for readers or for authors (also known as Diamond Open Access).

Editorial Office
Schloss Dagstuhl – Leibniz-Zentrum für Informatik
TGDK, Editorial Office
Oktavie-Allee, 66687 Wadern, Germany
tgdk@dagstuhl.de
https://www.dagstuhl.de/tgdk

Contents

List of Authors

Alexandre Bazin (2)
LIRMM, CNRS, Université de Montpellier, Montpellier, France

Melinda Hodkiewicz (3)
School of Mechanical Engineering, The University of Western Australia, Australia

Giacomo Kahn (2)
Université Lumière Lyon 2, INSA Lyon, Université Claude Bernard Lyon 1, France; Université Jean Monnet Saint-Etienne, DISP UR4570, Bron, France

N'Dah Jean Kouagou (1)
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

Axel-Cyrille Ngonga Ngomo (1)
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

Nicole Roocke (3)
Minerals Research Institute of Western Australia, Australia

Henri Scaffidi (3)
School of Computer Science and Software Engineering, The University of Western Australia, Australia

Arnab Sharma (1)
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

Caitlin Woods (3)
School of Computer Science and Software Engineering, The University of Western Australia, Australia

Resilience in Knowledge Graph Embeddings

Arnab Sharma
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

N'Dah Jean Kouagou
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

Axel-Cyrille Ngonga Ngomo
Data Science Group (DICE), Heinz Nixdorf Institute, Paderborn University, Germany

Abstract

In recent years, knowledge graphs have gained interest and witnessed widespread applications in various domains, such as information retrieval, question-answering, recommendation systems, amongst others. Large-scale knowledge graphs to this end have demonstrated their utility in effectively representing structured knowledge. To further facilitate the application of machine learning techniques, knowledge graph embedding models have been developed. Such models can transform entities and relationships within knowledge graphs into vectors. However, these embedding models often face challenges related to noise, missing information, distribution shift, adversarial attacks, etc. This can lead to sub-optimal embeddings and incorrect inferences, thereby negatively impacting downstream applications. While the existing literature has focused so far on adversarial attacks on KGE models, the challenges related to the other critical aspects remain unexplored. In this paper, we, first of all, give a unified definition of *resilience*, encompassing several factors such as generalisation, in-distribution generalization, distribution adaption, and robustness. After formalizing these concepts for machine learning in general, we define them in the context of knowledge graphs. To find the gap in the existing works on resilience in the context of knowledge graphs, we perform a systematic survey, taking into account all these aspects mentioned previously. Our survey results show that most of the existing works focus on a specific aspect of resilience, namely robustness. After categorizing such works based on their respective aspects of resilience, we discuss the challenges and future research directions.

2012 ACM Subject Classification Computing methodologies → Reasoning about belief and knowledge

Keywords and phrases Knowledge graphs, Resilience, Robustness

Digital Object Identifier 10.4230/TGDK.3.2.1

Category Survey

Funding This work has been supported by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) within the project SAIL under the grant no NW21-059D, the project WHALE (LFN 1-04) funded under the Lamarr Fellow Network programme by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW), the European Union's Horizon Europe research and innovation programme under grant agreement No 101070305, and by the German Federal Ministry of Research, Technology and Space (BMFTR) within the project KI-OWL under the grant no 01IS24057B.
Arnab Sharma: SAIL under the grant no NW21-059D
N'Dah Jean Kouagou: European Union's Horizon Europe research and innovation programme under grant agreement No 101070305, WHALE (LFN 1-04) funded under the Lamarr Fellow Network programme
Axel-Cyrille Ngonga Ngomo: KI-OWL under the grant no 01IS24057B

Received 2024-09-24 **Accepted** 2025-03-11 **Published** 2025-10-15

1 Introduction

In recent years, there has been significant progress in the construction and application of knowledge graphs (KGs). Many KGs, including Freebase [14], DBpedia [4], YAGO [91], and NELL [20], have been developed and successfully implemented in various real-world applications. Due to

their effectiveness in knowledge representation, KGs now find applications in domains such as information retrieval [32], question answering [42], and recommendation systems [99], amongst others. A KG serves as a structured depiction of knowledge, organized as a multi-relational graph where nodes denote entities or concepts, and edges signify relationships between them [53]. Knowledge therein is represented using assertions – model statements (which could in some cases be real-world facts) – in the form of triples denoted as (h, r, t), where h and t correspond to the head and tail entities respectively, and r represents the relationship between them. For instance, the fact "Biden is the president of USA" can be represented in a KG as $(\texttt{Biden}, \texttt{presidentOf}, \texttt{USA})$.

Knowledge Graph Embedding (KGE) involves transforming the entities and relations within a KG into vectors [15, 35, 103, 121, 124]. This transformation makes computational operations more feasible, allowing machine learning and deep learning techniques to be applied to extract insights from the KG. Consequently, an effective KGE model should aim to preserve the properties and semantics inherent in the original KG. Based on the type of KGE models, entities and relations are commonly embedded in d-dimensional vector spaces $\mathbb{V}$ such as $\mathbb{R}^d$ (real numbers) [15], $\mathbb{C}^d$ (complex numbers) [97], or even $\mathbb{H}^d$ (quaternions) [21].

Despite their effectiveness in capturing complex relationships between entities of KGs and facilitating various downstream tasks, KGE models can be vulnerable to *adversarial manipulations* [11,12,80,108,119,122]. Since these models rely heavily on the observed connections in a given graph, noise or missing information can lead to sub-optimal embeddings and potentially incorrect inferences. For instance, the presence of incorrect triples (in the sense of non-conformity with an ontology, or wrong assertions) might lead to poorly performing KGE models on certain downstream tasks [57, 129]. To this end, a deliberate attack or the presence of noise can both equally degrade the performance. KGE models might also struggle to generalize to out-of-distribution or unseen data, e.g., when the underlying data distribution changes or when encountering new entities or relations. Since KGs often contain sensitive and critical information pertaining to individuals or organizations, this might give rise to potential security vulnerabilities. For instance, an attacker might subtly alter the relation between entities or introduce fictitious entities and relationships that distort the model's understanding of the graph and make the KGE model learn *poisoned* embeddings. Such adversarial attacks on KGE models can take various forms, such as adding, deleting, or modifying triples within the knowledge graph, where such perturbations are often minimal and crafted to exploit vulnerabilities in the embedding process. Due to the usage of KGE models in various downstream tasks, such adversarial attacks can cause potential disruptions in these tasks, for instance, in

1. **Question answering**, adversarial modifications can cause KGE models to produce incorrect or manipulated answers or fail to retrieve relevant information,
2. **Recommendation systems**, the embeddings can be poisoned to promote certain items unfairly, leading to biased or irrelevant recommendations,
3. **Information extraction**, adversarial perturbations can result in inaccurate extractions of facts, affecting downstream applications like content summarization or data integration, amongst others.

Therefore, to reliably use KGE models in downstream tasks, there is a need to develop models that can function without any potential disruption even in the presence of such adversarial conditions.

Although the aforementioned challenges pose potential threats to the use of KGE models in critical downstream tasks, current efforts to deal with these challenges still remain infancy. The existing literature mostly contains works addressing challenges related to noisy data, distribution shifts, and adversarial attacks in the context of graph neural networks [24, 40, 127]. So far, works considering KGE models mostly focused on performing adversarial attacks on them [11,12,119,122]. The core idea behind these attacks is to target specific facts and modify the KGE model to either

increase or decrease their plausibility scores. These scores reflect the likelihood of a fact being true: higher scores imply higher probability, while lower scores imply lower probability. For instance, if (`Biden`, `PresidentOf`, `USA`) is selected as the target triple, one type of adversarial attack would be to make the underlying KGE model assign a low plausibility score to it. In this case, such attacks are typically dealt with via a min-max optimization function, where the objective is to minimize the inclusion/deletion of adversarial/existing triples in/from the underlying KG [119]. Simultaneously, the attacker aims to maximize the objective function, which involves either increasing or decreasing the plausibility of a targeted fact being true.

Since KGs are used in many safety-critical environments, safeguarding sensitive information and preserving user privacy are paramount considerations in deploying KGE models in real-world settings. Furthermore, we need to enable KGE models to adapt to dynamic environments and evolving data distributions to enhance their resilience to concept drift and temporal changes. Therefore, in this work, we first of all propose the concept of *resilience* in the context of ML, and further extend the definition for KGE models. We aim to bridge the gap in resilience literature on KGE from a holistic perspective that considers the diverse facets of robustness, adaptability, distribution shift, and consistency, amongst others. By addressing these aspects comprehensively, researchers can propel the development of resilient KGE models that not only excel in performance metrics but also demonstrate stability and reliability in real-world applications. Note that, our resilience definition is quite generic, i.e., it does not depend on any specific application domain. Precisely, we give a generic formal definition of resilience in ML models considering (i) generalization consistency, (ii) distribution adaptation, (iii) in-distribution generalization, (iv) robustness, and (v) missing entry handling. We then discuss these aspects of resilience in the context of KGE models. To this end, we survey the works on KGE models considering the aforementioned aspects of resilience. Specifically, we provide a survey of works studying the resilience of KGE models in any of the aspects from (i)–(v). After discussing these works, we highlight possible challenges and suggest future work directions.

Note that this paper provides two-faceted contributions. After exploring existing literature on KGE models, we recognize the need for a holistic definition of resilience in embedding models. Therefore, in this work, we first introduce a formal definition of resilience, considering five aspects. Thereafter, we discuss the related works in this context. In this sense, our paper is not purely a survey, rather, it combines a conceptual framework that defines and evaluates resilience in KGE models with a survey of different notions of resilience. Additionally, we propose a comparative analysis of existing methods and explore potential challenges and future works.

This paper is organized as follows. Section 2 formalizes the notions of KGs and KGE models. The definition of resilience is given in Section 3. Section 4 describes the methodology regarding the collection of papers. Existing works discussing aspects of resilience are presented in Section 5. Section 6 presents and discusses different aspects of robustness. Section 10 highlights existing challenges and potential future work directions, and Section 11 concludes the paper.

2 Foundations

A knowledge graph is a collection of assertions that describe a domain of interest. In this paper, we consider knowledge graphs composed of *triples* $(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where $\mathcal{E}$ is a discrete set of entities and $\mathcal{R}$ is a discrete set of relations. Therefore, KGs are representations of information in a discrete space. More formally, a KG is defined as a set of triples $\mathcal{G} := \{(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}$, where $\mathcal{E}$ and $\mathcal{R}$ stand for a set of entities and a set of relations [7,35]. To facilitate downstream applications, KGE algorithms have been developed to represent a KG in a continuous, low-dimensional vector space. Given a KG $\mathcal{G} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, the goal of a KGE model is to learn continuous vector

representations for entities and relation types in $\mathcal{G}$ such that these representations can be used to recover all the facts in $\mathcal{G}$. Most KGE approaches are tailored towards link prediction [21, 53], i.e., their scoring function is $\phi_\Theta : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$, where Θ denotes parameters and often comprises $\mathbf{E}$, $\mathbf{R}$, and additional parameters (e.g., affine transformations, batch normalizations, convolutions). Given an assertion in the form of a triple $(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, a prediction $\hat{y} := \phi_\Theta(h, r, t)$ signals the likelihood of (h, r, t) being true [35, 103, 121]. Therefore, KGE models are learned in such a way that the scoring function assigns a higher score to the triples that exist in the KG compared to the non-existing ones.

Let $\mathbb{V}$ denote a normed-division algebra, e.g. $\mathbb{R}, \mathbb{C}, \mathbb{H}$, or $\mathbb{O}$ [6, 34, 97, 116, 125]. A KGE model of a KG comprises entity embeddings $\mathbf{E} \in \mathbb{V}^{|\mathcal{E}| \times d_e}$ and relation embeddings $\mathbf{R} \in \mathbb{V}^{|\mathcal{R}| \times d_r}$, where d_e and d_r are the size of the embedding vectors. Note that some KGE models represent entities or relations as matrices or higher-dimensional tensors, e.g., RESCAL [65, 76]. Throughout this paper, we will focus on vector representations for entities and relations and denote embedding vectors with bold fonts, for instance, the embedding of h, r, and t will be denoted as $\mathbf{h}$, $\mathbf{r}$, and $\mathbf{t}$, respectively. Since KGs contain triples which represent the existing facts only, to learn a KGE model effectively, non-existing facts, i.e., *negative* facts often need to be incorporated into the learning process. For that, a technique called negative sampling is used to generate a number of false facts or negative triples. To this end, Bordes et al. [15] proposed a negative sampling technique by perturbing an entity in a randomly sampled triple from the KG. In this setting, a triple $(h, r, t) \in \mathcal{G}$ is considered as a positive example, whilst $\{(h, r, x) | x \in \mathcal{E} \wedge (h, r, x) \notin \mathcal{G}\} \cup \{(x, r, t) | x \in \mathcal{E} \wedge (x, r, t) \notin \mathcal{G}\}$ is regarded as the set of possible candidate negative triples corresponding to (h, r, t). During training, k negative triples are constructed for every correct triple.

3 Resilience

As mentioned beforehand, resilience is a term that is frequently used when engineering systems, more specifically in the context of building fault-tolerant systems [90]. In those systems, *resilience refers to the ability of a system to maintain its functionality and performance in the face of faults, failures, disruptions, or adverse conditions.* In other words, a resilient system is capable of detecting, mitigating, and recovering from faults or failures, ensuring continuous operation and minimal impact on its overall performance and availability. Therefore, in [10], the authors defined the resilience of a system using

1. **availability**, i.e., the readiness for correct service,
2. **reliability**, i.e., the probability of performing correctly for a period of time,
3. **safety**, i.e., the robustness against adversarial manipulations,
4. **integrity**, i.e., the absence of improper system altercation, and
5. **maintainability**, i.e., the ability to undergo modifications and repairs.

Note that, due to the lack of resilience definition in the ML literature, we use this as our starting point. While the typical definition of resilience in fault-tolerant systems provides a useful help for understanding resilience in the machine learning domain, it needs to be extended and adapted to account for the unique characteristics, challenges, and considerations inherent in machine learning models and systems. More specifically, for ML models, resilience cannot be defined by using these parameters directly since they do not capture the typical data-driven workflow that is used in ML. For this, we need to consider other factors such as *consistent* performance in a distribution, or when a *distribution shift* occurs, robustness, stability, amongst others.

This paper makes a two-fold contribution in addressing this gap. First, after analyzing existing works on KGE models, we identify the necessity for a holistic definition of resilience tailored to embedding models. To this end, we propose a formal definition of resilience, considering aspects

that encapsulate the ability of a model to generalize, adapt, and maintain stable performance under varying conditions. Second, we provide a structured discussion of related works that align with these resilience principles, offering insights into existing approaches and their limitations. As mentioned beforehand, our work is not purely a survey; rather, it combines theoretical formalization with a comprehensive review, bridging the gap between conceptual understanding and practical advancements in resilient knowledge representation.

To define resilience, we start with some basic formalization. Let us consider an ML model as a function f which takes as input x coming from a specific distribution $\mathcal{D}$. We define two types of distributions from where the data might come, the source distribution which is defined as $\mathcal{D}_s$ from where the training data comes, and the target distribution $\mathcal{D}_t$ on which the model would typically operate. The sets of values corresponding to the distributions $\mathcal{D}_s$ and $\mathcal{D}_t$ can be defined as $\mathcal{X}_s$ and $\mathcal{X}_t$, respectively. $\mathcal{H}(\mathcal{D}_s, \mathcal{D}_t)$ defines a divergence measure between the two distributions $\mathcal{D}_s$ and $\mathcal{D}_t$. Furthermore, we define $\mathcal{L}_f$ as the *loss function* of the model measuring the model's performance on a set of data instances. Note that, the loss function can be of any type, however, our definition is independent of it. An ML model f is said to be resilient if it conforms to the following constraints:

Generalization consistency corresponds to the ability of the model to generalize consistently across different distributions of data. This can be formally defined as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ \mathcal{D}_s, \mathcal{D}_t,\ |\mathbb{E}_{\mathcal{D}_s}(\mathcal{L}_f) - \mathbb{E}_{\mathcal{D}_t}(\mathcal{L}_f)| \leq \epsilon, \tag{1}$$

where ϵ defines a threshold that basically bounds the difference between the average losses on the training data distribution $\mathcal{D}_s$ and the target data distribution $\mathcal{D}_t$. Xu et al. [112] defined this as the robustness property of the learning algorithms where they argued that a robust algorithm should achieve similar performance on the training and testing data that are *close* in some sense; which basically corresponds to the robust optimization problem. However, in connection to resilience, we define this as the *consistency* property over the generalization of the model f. To this end, we simply say that the loss occurring on the data instances taken from the target distribution might differ only by a threshold ϵ from the loss occurring on the instances of the source distribution. Note that, in that sense, this definition could also be termed as *out of distribution generalization*, since this captures how well a model can perform when the inputs are out-of-distribution compared to the training dataset. Furthermore, in this definition, we are not concerned with whether a model f achieves high accuracy or low loss on the training data; with generalization consistency, we aim to signify that the model's performance should not vary drastically between the training and test data distributions.
In the context of knowledge graph embedding models, generalization consistency refers to the model's ability to meaningfully construct embeddings for unseen entities or relations, and accurately predict missing links between entities based on the learned patterns from the training data. Assuming $\mathcal{L}_{\phi_\Theta}$ is a loss function which can be used to train the parameterized embedding model ϕ_Θ, generalization consistency can be defined as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ \mathcal{D}_\mathcal{G}, \mathcal{D}_{\mathcal{G}'},\ |\mathbb{E}_\mathcal{G}(\mathcal{L}_{\phi_\Theta}) - \mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta})| \leq \epsilon, \tag{2}$$

where $\mathcal{D}_\mathcal{G}$ and $\mathcal{D}_{\mathcal{G}'}$ refer to the distribution of the training knowledge graph's data and that of the test knowledge graph's data, respectively.

Distribution adaption corresponds to the model's ability to adapt to a target domain (i.e., test data distribution) without significantly compromising its performance as achieved on the source domain (i.e., training data distribution). This can be defined as follows

$$\forall\ \epsilon > 0, \exists\ \delta > 0 \text{ s.t. } \forall\ \mathcal{D}_s, \mathcal{D}_t,\ \mathcal{H}(\mathcal{D}_s, \mathcal{D}_t) \leq \delta \Rightarrow |\mathbb{E}_{\mathcal{D}_s}(\mathcal{L}_f) - \mathbb{E}_{\mathcal{D}_t}(\mathcal{L}_f)| \leq \epsilon, \tag{3}$$

where $\mathcal{H}(\mathcal{D}_s, \mathcal{D}_t)$ defines any divergence measure such as maximum mean discrepancy (MMD) [89], Kullback-Leibler (KL) divergence [59], or Wasserstein distance [106]. Informally, if the distributions $\mathcal{D}_s$ and $\mathcal{D}_t$ are different with a bound δ, then the average prediction losses on the data instances in these distributions must not differ more than ϵ. Note that, the distributional mismatch between the training and test data has been studied in many settings, for instance, in [13, 39, 54, 82, 88, 115] and as pointed out by the authors in [1] most of these works assume the *covariate shift* where only the distribution of class labels differs between the training and test distributions. There exist some works such as [9, 29, 41] which consider shift of generic data distributions, however, none of them consider this as part of the resilience of ML models. For KGE models, distribution adaptation refers to a model's ability to adjust its parameters to account for changes in a given knowledge graph. When new entities, relation types, or new links are added to (or removed from) a given knowledge graph, the resulting graph data distribution might deviate from the initial one. In this case, the KGE model's adaptation to this distribution change can be formally defined as follows.

$$\forall\ \epsilon > 0, \exists\ \delta > 0 \text{ s.t. } \forall\ \mathcal{D}_{\mathcal{G}}, \mathcal{D}_{\mathcal{G}'},\ \mathcal{H}(\mathcal{D}_{\mathcal{G}}, \mathcal{D}_{\mathcal{G}'}) \leq \delta \Rightarrow |\mathbb{E}_{\mathcal{D}_{\mathcal{G}}}(\mathcal{L}_{\phi_\Theta}) - \mathbb{E}_{\mathcal{D}_{\mathcal{G}'}}(\mathcal{L}_{\phi_\Theta})| \leq \epsilon. \tag{4}$$

Unlike generalization consistency, which assumes stable data conditions, distribution adaptation ensures that the model can adjust to new distributions without significant performance degradation. In other words, generalization consistency ensures stability across similar distributions, while distribution adaptation guarantees stability on dynamic or shifted distributions.
Note that, in the context of graphs, a distribution shift refers to a change in the statistical distribution of the graph data. This can manifest in different ways, such as

1. **Node feature distribution shift** which occurs when the distribution of node attributes or features changes over time or across different subsets of the graph. For example, in a knowledge graph representing entities and their attributes (e.g., people and their professions), a node attribute shift could involve changes in the distribution of professions among individuals over time or across different subsets of the graph. Nodes may furthermore be added to or removed from the knowledge graph, leading to changes in the overall node distribution. This could happen, for instance, when new entities are discovered or when outdated entities are removed from the knowledge graph.
2. **Node degree shift** which happens when some relationships between entities are removed (e.g., two entities that were previously friends are no longer friends) or added, e.g., (an entity gets married to another entity). It could also be the case that new entities are introduced but with little to zero links to other entities in the graph. When such changes in relationships between entities are significant, the average degree of nodes in the considered knowledge graph might also shift.
3. **Edge feature distribution shift** which refers to changes in the properties or attributes associated with the relationships (edges) between nodes in the knowledge graph. For example, in a knowledge graph representing relationships between entities (e.g., co-authorship relationships between researchers), an edge attribute shift could involve changes in the publication venues or collaboration patterns over time. New relationships may further be established or existing relationships may be removed from the knowledge graph, leading to changes in the edge distribution. This could occur due to the emergence of new relationships or the obsolescence of existing ones.
4. **Graph structure shift** which involves alterations in the overall structure or topology of the knowledge graph, including changes in connectivity patterns between nodes, changes in node/edge attributes (e.g., many entities and relationships in the reference knowledge graph now have textual descriptions), and changes in entity type hierarchies. For example, in a

knowledge graph representing hierarchical relationships (e.g., taxonomy or ontology), changes in the hierarchy or the addition of new branches can lead to structural shifts. Changes to the schema or ontology of the knowledge graph, such as the addition, modification, or removal of entity types, relationship types, or property types, can also constitute graph structure shifts. These changes may reflect updates in domain knowledge or evolving data modeling requirements.

To make the distinction between generalization consistency and distribution adaption more concrete in the context of KGE models, we consider the following example. Consider a recommendation system based on a knowledge graph. Generalization consistency would ensure that the embeddings trained on historical data remain effective for predicting new links in the same dataset. However, distribution adaptation would be required if the dataset undergoes significant changes, such as the inclusion of new user demographics or shifting product categories.

In-distribution generalization corresponds to the model's ability to perform consistently across different instances or subsets of a data distribution. Typically, this distribution could be the target distribution $\mathcal{D}_t$ where the data instances come from the model deployment phase. Formally, consistency can be defined as follows,

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ \mathcal{S}, \mathcal{S}' \subseteq \mathcal{D}_{\mathcal{G}},\ |\mathbb{E}_{\mathcal{S}}(\mathcal{L}_f) - \mathbb{E}_{\mathcal{S}'}(\mathcal{L}_f)| \leq \epsilon. \tag{5}$$

Here we enforce that, for any two non-empty subsets $\mathcal{S}$ and $\mathcal{S}'$ from the distribution $\mathcal{D}_{\mathcal{G}}$, the expected losses achieved on the two sets differ at most only by some parameter ϵ [1]. If the observed differences between the losses are statistically significant (e.g., greater than ϵ), it indicates that the model's performance varies consistently across different subsets of the data, suggesting potential limitations or biases in the model. On the other hand, if the observed differences are not statistically significant, it suggests that the model's performance remains consistent across subsets, providing greater confidence in its resilience. Furthermore, this measure of consistency is different from the generalization consistency in the sense that herein we consider uniform performance across different sub-spaces of the same distribution space, whereas in case of generalization consistency, two different distributions are considered.

In-distribution generalization for the KGE models refers to the model's ability to maintain consistent performance across different instances or subsets of the knowledge graph data distribution, particularly when deployed in real-world applications where the distribution of incoming triples may vary. In other words, the KGE model should demonstrate resilience to variations in the distribution of knowledge graph data encountered during deployment, ensuring that its performance remains reliable and predictable across different scenarios. This consistency is crucial for maintaining the effectiveness and reliability of the model in real-world applications where the knowledge graph would evolve over time or across different contexts.

Robustness focuses on the model's stability with respect to some small changes in the input. In the literature, two versions of robustness are generally considered, namely *local* and *global* robustness [47, 85]. Informally, local robustness corresponds to a single point x, and requires any points within a specific distance Δ to x to be classified as the same as the former. More formally, this can be defined with respect to a data point x as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ x',\ ||x - x'||_p \leq \Delta \Rightarrow ||f(x) - f(x')||_p \leq \epsilon. \tag{6}$$

[1] Note that, here we have considered a strong notion of consistency, however, a weaker notion can also be chosen where the subsets must follow some specific rules.

On the other hand, Seshia et al. [85] defined global robustness considering all the points within a specific distribution $\mathcal{D}$. In other words, for every point x within a considered distribution, any other point x' which is within Δ distance from x should be classified as the same class as x. This can be formally defined as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ x, x' \in \mathcal{D},\ ||x - x'||_p \leq \Delta \Rightarrow ||f(x) - f(x')|| \leq \epsilon. \tag{7}$$

Note that, in the literature, robustness is more often associated with the idea of local robustness for a single point or a set of points. Thus, in defining resilience, we would primarily consider the local robustness property of ML models. Note that, herein, $f(x)$ could be a single integer or could also be a probability. This would depend on the type of the underlying model.
In the context of KGE models, we can adapt the concept of local robustness to refer to the model's ability to produce consistent embeddings for entities or relations that are similar in the graph structure. Informally, local robustness in this context would correspond to: any entity (respectively, relation) within a specific neighborhood of an entity h (respectively a relation r), defined by a distance metric, should have an embedding that is similar to that of h (respectively r). More formally, for an entity or a relation x in the knowledge graph, and for any other entity or relation x' within a specific distance Δ of x in the graph structure, the embeddings produced by the KGE model, say $\mathbf{x}$ and $\mathbf{x}'$ should be similar, with their distance in the embedding space bounded by ϵ. Given a knowledge graph $\mathcal{G}$, this idea of robustness can be formally defined as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ x, x' \in \mathcal{G}, d_{\mathcal{G}}(x, x') \leq \Delta \Rightarrow d_{Emb}(\mathbf{x}, \mathbf{x}') \leq \epsilon, \tag{8}$$

where $d_{\mathcal{G}} : \mathcal{G} \times \mathcal{G} \rightarrow \mathbb{R}_+$ is a distance on the graph $\mathcal{G}$, e.g., Adamic-Adar index, Katz similarity, or Common Neighbors, and d_{Emb} a distance function in the embedding space, e.g., the Euclidean distance. ϵ is a threshold that limits the allowable difference between embeddings to ensure local robustness.
The above definition of robustness concerns the functionality of the embedding models in generating robust embeddings. However, we require further robustness notion encompassing the KGE model as well as the scoring function together. To this end, first of all, we define the adversarial robustness.
Adversarial robustness for KGE models refers to a model's ability to maintain its performance and produce reliable predictions in the presence of worst-case perturbations intentionally crafted to degrade its functionality. Herein, we define the robustness property considering the KGE model+scoring function whereas the previous robustness definition (Equation 8) considers solely the KGE models. The perturbations considered can be applied to the symbolic KG or directly to the embedding space and are designed to maximize the model's predictive errors. Formally, adversarial robustness by considering the *symbolic KG* can be described as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ \mathcal{G}, \mathcal{G}', \Psi(\mathcal{G}, \mathcal{G}') \leq \Delta \implies \frac{\mathbb{E}_{\mathcal{G}}(\mathcal{L}_{\phi_\Theta})}{\mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta}) + \eta} \approx \epsilon, \tag{9}$$

where $\Psi(\mathcal{G}, \mathcal{G}')$ denotes the structural similarity of two graphs, $\mathbb{E}_{\mathcal{G}}(\mathcal{L}_{\phi_\Theta})$, $\mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta})$ denote the expected loss of the embedding model ϕ_Θ on the graphs $\mathcal{G}$ and $\mathcal{G}'$, respectively, and η is an infinitesimal number (e.g., $\eta = 10^{-8}$, i.e., a small but non-zero scalar value). ϵ in the ideal case should be close to 1. $\Psi(\mathcal{G}, \mathcal{G}')$ could be defined using graph isomorphism [49], or sub-graph similarity [81] matching technique such as graphlet similarity, frequent subgraph mining, or global graph similarity techniques. Such a measure could be decided based on the specific domain. Note that, the above definition considers changes in the KG, however, this could be extended considering perturbations performed on the embedding. Formally, adversarial robustness by considering the embedding can be described as

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ \mathbf{E}, \mathbf{E}' : \Psi(\mathbf{E}, \mathbf{E}') \leq \Delta \implies \frac{\mathbb{E}_{\mathbf{E}}(\mathcal{L}_{\phi_\Theta})}{\mathbb{E}_{\mathbf{E}'}(\mathcal{L}_{\phi_\Theta}) + \eta} \approx \epsilon. \tag{10}$$

Here $\mathbf{E}$ and $\mathbf{E}'$ represent the original and perturbed embeddings, respectively. $\Psi(\mathbf{E}, \mathbf{E}')$ measures the similarity or distance between $\mathbf{E}$ and $\mathbf{E}'$. This can be defined as $||\mathbf{E} - \mathbf{E}'||_p$ (e.g., $p = 1$ for the L1 norm and $p = 2$ for the L2 norm) or $\frac{\mathbf{E} \cdot \mathbf{E}'}{||\mathbf{E}||, ||\mathbf{E}'||}$ (cosine similarity) or $\sum_i \mathbf{E}_i \log\left(\frac{\mathbf{E}_i}{\mathbf{E}'_i}\right)$ (KL divergence).

Note that Equations (8)–(10) give a generic notion of adversarial robustness which can be extended by considering the case where the aim is to degrade the score of a specific triple, i.e., $\phi_\theta(h, r, t)$ by doing δ changes on the KG $\mathcal{G}$ or on the embedding space $\mathbf{E}$. The existing works on adversarial robustness of KGE models, while lacking a formal definition, focus on this specific notion [75, 86, 87, 111, 128].

Non-adversarial robustness corresponds to the ability of a KGE model (including its scoring function) to be invariant to a certain level of noise present in a KG. More specifically, the performance of a robust KGE model should not degrade considerably when noise is prevalent in KG. Consider $\mathcal{G}$ as a clean KG and $\mathcal{G}'$ as a noisy KG, where the latter is obtained by adding δ amount of noise to the KG $\mathcal{G}$, i.e., $\mathcal{G}' = \mathcal{G} + \delta$. [2] Then the robustness can be defined as

$$\frac{\mathbb{E}_{\mathcal{G}}(\mathcal{L}_{\phi_\Theta})}{\mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta}) + \eta} \approx 1, \tag{11}$$

where $\mathbb{E}_{\mathcal{G}}(\mathcal{L}_{\phi_\Theta})$, $\mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta})$ denote the expected loss of the embedding model ϕ_Θ on the graphs $\mathcal{G}$ and $\mathcal{G}$, respectively, and η is an infinitesimal number. This implies that the performance of the KGE models should remain almost the same even when δ amount of noise is present in $\mathcal{G}$. Note that we assume the expected loss not to be zero, as it is often the case in most machine learning tasks. Note that the primary difference between adversarial and non-adversarial robustness lies in the nature of the perturbations. Adversarial perturbations are crafted with intent and precision, targeting the model's weaknesses, while non-adversarial perturbations are accidental and random, reflecting real-world data imperfections. Additionally, adversarial robustness is critical for security-focused applications to protect against malicious attacks, whereas non-adversarial robustness is essential for ensuring reliability in a real-world environment.

Stability w.r.t. incomplete input deals with the model's ability to handle missing values, more specifically, maintain accurate predictions despite the presence of missing values in the input features. We can express this as follows:

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ x, x^*, |x| > |x^*| \ \wedge\ \text{Sim}(x, x^*) \leq \delta \Rightarrow ||f(x) - f(x')|| \geq \epsilon. \tag{12}$$

Herein $\text{Sim}(x, x^*)$ measures the similarity between two vectors x and x^* with unequal numbers of elements (i.e., $|x| > |x^*|$). One such similarity measure could be the cosine similarity, which is often used for comparing the similarity between vectors in high-dimensional spaces. The cosine similarity measures the cosine of the angle between two vectors and is defined as the dot product of the vectors divided by the product of their magnitudes. Elements that are missing in one vector but present in the other are effectively treated as zeros in the dot product. Another approach is to use measures that explicitly handle missing values, such as the Jaccard similarity or the Pearson correlation coefficient, with the imputation of missing values.

[2] Here, + is not the usual addition, but a perturbation operator instead.

In the context of KGs, this aspect of resilience deals with the ability of a KGE model to maintain stable predictions despite missing nodes, edges, or attributes in the input knowledge graph. To this end, we define stability to incomplete input formally as follows:

$$\exists\ \epsilon > 0 \text{ s.t. } \forall\ \mathcal{G}, \mathcal{G}^*, |\mathcal{G}| > |\mathcal{G}^*| \wedge \text{Sim}(\mathcal{G}, \mathcal{G}^*) \leq \delta \Rightarrow ||\mathbb{E}_{\mathcal{G}}(\mathcal{L}_{\phi_\Theta}) - \mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta})|| \leq \epsilon. \tag{13}$$

where $\mathcal{G}$ is the original, complete knowledge graph, and $\mathcal{G}'$ is the incomplete KG with missing nodes, edges, or attributes. $\text{Sim}(\mathcal{G}, \mathcal{G}^*)$ measures the structural similarity between the two graphs. For KGs, Sim() could be graph edit distance, i.e., number of node/edge insertions, deletions, or modifications required to transform $\mathcal{G}$ into $\mathcal{G}'$, or Jaccard similarity over entity/relation sets, amongst others. Note that, this definition implies that the expected loss of the embedding model ϕ_Θ on the graphs $\mathcal{G}$ and $\mathcal{G}$, $\mathbb{E}_{\mathcal{G}}(\mathcal{L}_{\phi_\Theta})$, $\mathbb{E}_{\mathcal{G}'}(\mathcal{L}_{\phi_\Theta})$ respectively, should not change more than a fixed threshold ϵ. The similarity measure can be quite flexible and will potentially depend on the domain (for e.g., image classification, graph data, and others), and hence, we do not fix the Sim() function. Depending on this function and the domain of the application, the bound δ will also change, however, not drastically.

4 Paper Collection Methodology

To discuss resilience in KGs and KGE models, in this work, we further review existing works in this domain. While doing a literature survey of such works, we adhere to specific inclusion criteria for compiling papers for our review. If a paper satisfies any or many of the following criteria, it is considered for inclusion:

1. the paper introduces or discusses the overarching concept of any related aspect outlined in Section 3.
2. the paper proposes an approach, study, or tool/framework aimed at developing resilient or robust KGE models.
3. the paper introduces a set of measurement criteria applicable for defining resilience of KGE models or KGs.

We briefly discuss some papers focusing solely on using KGs to make resilient systems, however, we do not delve into detail on this. To comprehensively gather papers across various research domains, we initiated our search process by employing precise keyword queries on prominent scientific databases such as Google Scholar, DBLP, and arXiv. The keywords that we searched for are detailed in the **Keywords** column in Table 1. We conducted searches across the three repositories until 23.09.2024, aiming to encompass a broad spectrum of literature. The specifics of the paper collection outcomes are outlined in Table 1. It is observed that the papers obtained from Google Scholar and arXiv were subsets of those gathered from DBLP. Therefore, we solely present the results obtained from DBLP. Furthermore note that apart from the papers that discuss resilience in KGE models, or in KGs, we also report the results here where any of the aspects of resilience (as described in Section 3) are discussed in the body of the paper.

Note that there exist a number of surveys discussing primarily two aspects of resilience proposed in this paper, such as robustness of deep learning models [28, 74, 113], language models [27, 48]; distribution adaption [64]. However, none of them give a definition of resilience considering the notions that we describe above and discuss the related works encompassing these aspects individually. Moreover, such a study is not done considering the KGE models. The closest work to ours are the works related to GNNs. Existing literature considering GNNs is quite vast and furthermore, there already exist surveys discussing the robustness [31, 114] and some other aspects of resilience of graph neural networks [123, 131]. Discussing the works related to the resilience of GNNs would extend this paper to a much greater extend. Therefore, we do not consider the GNNs in this paper.

Table 1 Paper query results. Here, "Body" represents the main content of a paper. Numbers correspond to the number of articles where the keyword occurred more than once.

Keywords	Title	Body
resilience in knowledge graphs	4	0
resilience in knowledge graph embedding models	2	6
generalization consistency in/of knowledge graphs	0	0
generalization consistency in/of knowledge graph embedding models	0	0
domain adaption in/of knowledge graph embedding models	0	2
distribution shift of knowledge graphs	1	0
in-distribution generalization of knowledge graph embedding models	1	0
robustness of knowledge graph embedding models	8	18

5 Resilience in KGE

The existing works considering resilience in KGE models mainly focus on a specific aspect, that is building KGE models that are resilient against the noise present in the KGs. To this end, there exist a number of such contributions [75, 86, 87, 111, 128]. These works consider if the performance of the model does not degrade with noise present in the KG, then the underlying KGE model is said to be resilient. This, however, is not resilience based on the definition provided in Section 3 where we defined resilience as a multi-faceted term that takes into account many aspects. Based on our definition of resilience, the work on resilience to safeguard against *noises* in the KGs mostly aligns with the definition of robustness, more specifically, the non-adversarial robustness (as defined in Equation 11). Therefore, we categorize this line of work related to resilience against noise as part of the non-adversarial robustness.

Moreover, there exist some works which concentrate on constructing resilient systems leveraging KGs [3, 30, 58, 117]. For instance, the works in [117] focus on building a KG-based risk assessment framework to improve the resiliency of supply chain management. A KG is built in [58] from the natural disaster data to improve the disaster management department's resilience towards such incidents. A similar sort of study is done in [3] to employ a resilient management system in case a crisis happens in a city. To assess the resiliency of the cyber-physical system for a water management system, Dagnas et al. [30] utilized the KG as a modeling graph.

Apart from the works mentioned earlier, no other works could be found that consider resilience (as we defined in this work) as part of KGs or KGE models, therefore, in this work, we further survey the existing literature by considering individual aspects of resilience as described in Equation (1)–(13). Figure 1 shows the categorization of the works found corresponding to the aspects of resilience that we described in Section 3. There exist works that focus on improving the generalizability of the KGE models [50, 55, 62, 107], or focusing on the logical consistency of the ontological rules [33, 37, 38, 51, 84], however, no work exists discussing generalization consistency, distribution adaption, and stability to incomplete input of KGE models. A very recent work by Zhu et al. [133] discusses the in-distribution generalization aspect and this is the only work that we could find related to this aspect of resilience.

More importantly, all the works found in regards to resilience in KGE can be distributed along two fields of robustness, namely adversarial and non-adversarial. There exists work such as by Zhu et al. [132] where they proposed to use KG to tackle the distribution shift problem for the *few-shot* learning approach. More specifically, by using KGs, the aim is to capture the semantic relationship between different categories of instances. Despite the data samples originating from

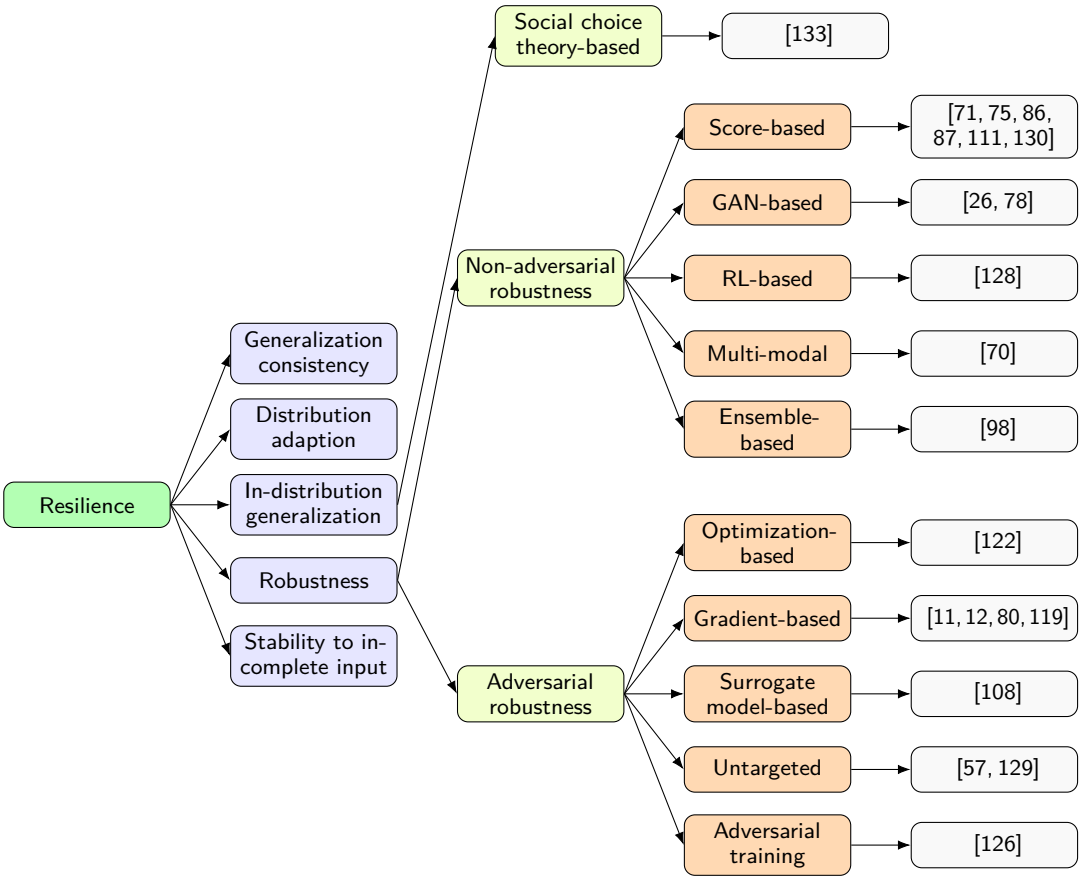


Figure 1 Categorisation of different works based on the underlying approaches in the context of resilience for knowledge graph embedding models.

diverse distributions, they frequently possess shared auxiliary knowledge, along with prior semantic relationships between classes. For this, KG can be used to find out when such a distribution shift occurs and help the underlying model to adapt.

Note that there exists *temporal* KG which dynamically evolves over time [56]. The idea therein is to model the temporal information in KGs to keep track of how different assertions/facts evolve over time. For this, the KG is defined as $\mathcal{G}_t := \{(h, r, t, t') \in \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}\}$, where t' basically points to the timestamp for a specific fact. For instance, "(Barack Obama, President_of, USA, [2009-2017])" is an assertion associated with a timestamp [2009-2017] for which it is true. Temporal KGs are called dynamic KGs since they are not static and evolve with the addition of new timestamps corresponding to assertions, and there are KGE models that attempt to learn embeddings for such graphs, such as [2, 52, 92, 94, 120]. Although such KGE models learn to map entities and relations into an embedding space that changes over time, we do not expect the learned embeddings to adapt to the distribution shift of the KG. This is because such KGE approaches assume that, alike typical KGs, the distribution of temporal KGs does not change and only a new timestamp is added by replacing the old timestamp, for instance, (h, r, t, t') is replaced by (h, r, t, t^*). The underlying KGE models therefore only need to adapt based on this newly added timestamp. Therefore, we do not consider such works under distribution adaption aspect of resilience, and a survey on temporal KGE models can be found here [101].

Since there do not exist any works considering the aspects of resilience apart from robustness and in-distribution generalization, in the following, we first describe the related works encompassing the two areas that exist regarding the resilience of KGE.

6 Robustness

The concept of robustness in KGE models, as discussed earlier, can be divided into two main types: adversarial and non-adversarial robustness. As mentioned in Section 3, adversarial robustness concerns the model's ability to withstand intentional attacks, where malicious entities modify the knowledge graph (KG) to compromise the KGE model's performance. In contrast, non-adversarial robustness deals with the model's resilience against noise and inconsistencies naturally present in KGs, without any malicious intent. In the following subsections, we provide a detailed survey of the current research related to these two areas of robustness in KGE models. We begin with adversarial robustness, followed by an exploration of non-adversarial robustness techniques.

6.1 Adversarial Robustness

Despite its significance, the existing works on adversarial robustness in KGE models are in their infancy. The majority of studies focus on generating adversarial examples to deliberately manipulate the knowledge graph and assess the vulnerability of KGE models [11,12,80,108,119,122]. These works proposed methods for *attacking* the KGE models by generating adversarial examples to study the robustness of the existing KGE models. Below we categorize them based on the approaches used to perform such attacks.

6.1.1 Optimization Approach

One of the earlier works in this area [122] introduced a data poisoning attack strategy, aiming to alter the score of a target triple (h_t, r_t, t_t) by modifying the KG. To achieve the poisoning goal, they assumed the attacker had a fixed budget (for instance, like Δ in Equation 8) in terms of the number of changes that could be made on the KG. To this end, they have given two attack strategies, namely direct and indirect attack.

- **Direct attack.** The direct attack involves identifying the necessary perturbations, such as adding or removing triples, to achieve the attacker's objective – for example, reducing the likelihood of a target fact (h_t, r_t, t_t) being true. This process starts by determining the embedding shift ϵ required for either the head entity $\mathbf{h}_t$ or the tail entity $\mathbf{t}_t$ of the target triple to ensure that the new score $\phi'_\Theta(h_t, r_t, t_t)$, learned on the adversarially modified KG, is lower than the original score $\phi_\Theta(h_t, r_t, t_t)$. Potential perturbations are evaluated and ranked based on a scoring metric, guiding the selection of the most effective changes. The top M perturbations are then chosen using an optimization technique, taking into account the attacker's budget and constraints.
- **Indirect attack.** Performing a direct attack that involves shifting the embeddings of the target triple might be detected by using some kind of sanity check. Hence, to make the attack stealthy, the authors in [122] proposed indirect attack which involves shifting the embedding of the entities which are some k-hops away from the target triple (h_t, r_t, t_t). The changes would then propagate to the required embedding shifting of the target triple.

The adversarial attacks described above are performed considering the KG itself, and therefore, it adheres to Definition 9. While direct and indirect adversarial attacks on knowledge graph embeddings leverage optimization techniques to identify the most impactful perturbations, there are drawbacks when using such an approach in this process. One key limitation is that KGs have a highly discrete and complex structure, making it difficult to navigate the search space effectively using straightforward optimization.

6.1.2 Gradient-based and Attribution Attacks

Gradient-based approaches have emerged as a more effective alternative to performing adversarial attacks on the KGE models compared to simple optimization approaches. Note that the adversarial attack performed herein follows the definition of Equation 10. By leveraging the continuous embedding space of KGE models, gradient-based methods allow for a more effective exploration of potential perturbations. These approaches identify influential triples or paths in the KG by analyzing the gradient of the model's loss function with respect to the embeddings, enabling targeted modifications that maximize the attack's impact. Unlike optimization-based methods, gradient-based approaches offer computational advantages by operating in a lower-dimensional, continuous space, albeit with limitations in their applicability to specific types of KGE models.

Building on the ideas of direct and indirect attacks that target specific triples or entities in the knowledge graph, Pezeshkpour et al. [80] followed a typical gradient-based approach to find out the most influential neighboring triple (h'_t, r'_t, t_t) of the target triple (h_t, r_t, t_t), the removal $(\mathcal{G} \setminus \{(h'_t, r'_t, t_t)\})$ or addition $(\mathcal{G} \cup \{(h'_t, r'_t, t_t)\})$ of which would maximize the attack objective which can be defined as

$$\underset{(h'_t, r'_t)}{\operatorname{argmax}} \phi_\Theta(h_t, r_t, t_t) - \phi'_\Theta(h_t, r_t, t_t),$$

where $\phi'_\Theta(h_t, r_t, t_t)$ defines the score when trained on either $\mathcal{G} \setminus \{(h'_t, r'_t, t_t)\}$ or $\mathcal{G} \cup \{(h'_t, r'_t, t_t)\}$. However, searching for such a h'_t, r'_t is computationally expensive since the size of the search space is $|\mathcal{E}| \times |\mathcal{R}|$ (*number of entities in* $\mathcal{G}$ $\times$ *number of relations in* $\mathcal{G}$). Therefore, unlike the previous work [122], the authors herein modified the objective function by performing the search in the embedding domain, i.e., in the continuous space which gives the embedding for the optimal head and relation as $\mathbf{h}'_t, \mathbf{r}'_t$. Thereafter, an autoencoder is used to get h'_t, r'_t from $\mathbf{h}'_t, \mathbf{r}'_t$. However, one of the drawbacks of this approach is that it could only be used for multiplicative KGE models and moreover it does not take into account the nature of the KGE model being attacked.

Bhardwaj et al. [12] proposed a poisoning attack on KGE models by leveraging the inductive capabilities of these models, encapsulated through relationship patterns such as symmetry, inversion, and composition within a knowledge graph. Their approach aims to either decrease or increase the model's confidence in predicting a target triple h_t, r_t, t_t. For instance, if the attacker's goal is to decrease the score, they aim to ensure that $\phi_\Theta(h_t, r_t, t_t) > \phi'_\Theta(h, r, t)$, where ϕ'_Θ is the model learned on the KG modified with the addition of adversarial triples, referred to as *decoy triples.* These decoy triples are selected based on the inductive relation patterns that the KGE model captures. For example, if there exists a target triple h, r, t composed of $h_t, r_1, \bar{t}$ and $\bar{t}, r_2, t_t$, an additive model that captures the symmetry relationship can be exploited, such that $\mathbf{r}_1 + \mathbf{r}_2 = \mathbf{r}$. The model then selects a relation $\mathbf{r}_t$ as the target relation, minimizing the Euclidean distance $|\mathbf{r}_t - (\mathbf{r}_1 + \mathbf{r}_2)|$. By doing so, the method identifies the relation that strongly captures the symmetry. Once the target relation is chosen, two decoy triples are added in the form of h, r_1, t^* and t^*, r_2, t'. These added triples manipulate the inductive properties of the KGE model, indirectly decreasing the score of the original target triple h, r, t. By exploiting the underlying inductive patterns that KGE models learn, such as symmetry and composition, this approach makes the target triple less likely to be predicted as true.

Bhardwaj et al. [11] further extended their approach by employing instance attribution methods from the domain of interpretable machine learning to carry out data poisoning attacks on KGE models. The aim of these attacks remains similar to their previous work: reducing the likelihood of the target triple (h_t, r_t, t_t) being correctly predicted by the KGE model. They specifically defined the attacker's capability as the ability to make a single change (either by removing or adding a triple) within the neighborhood of the target triple. The neighborhood is constructed

based on triples that share either the subject or object of the target triple, formally defined as $\mathcal{H} = (h_n, r_n, t_n) \mid h_n \in h_t, t_t \vee t_n \in h_t, t_t$. To identify which triple should be manipulated, they introduced an *influence score* $\mathcal{I}((h_t, r_t, t_t), (h, r, t))$. This score measures the effect that a particular training triple (h, r, t) has on the model's prediction for the target triple (h_t, r_t, t_t). A larger influence score indicates that removing the triple (h, r, t) would significantly reduce the predicted score for (h_t, r_t, t_t). However, directly retraining the KGE model for each triple removal is computationally expensive. To tackle this, the authors adopted techniques from interpretable machine learning, specifically using similarity metrics in the embedding space.

You et al. [119] recently proposed a model-agnostic, semantic, and stealthy data poisoning attack on KGE models, addressing several aspects: black-box attack, semantically preserving poisoning, and stealthiness by ensuring good performance for clean triples. Unlike previous works, their approach focuses on inserting *indicative paths* rather than individual triples to maximize the prediction probability of a target poisoned triple. The attack goal can be formalized as

$$\max_{\hat{T}} \phi_{\Theta}(h_t, r_t, t_t),$$

where $\hat{T}$ is the set of triples in the indicative path. In their approach, the key idea is to add *indicative paths* that comprise more than one triple, which encourages the KGE model to predict the malicious fact as true. They translate the relation of the malicious fact into a sequence of relations using a *path template*. For example, a path template $p_{h_t \to t_t}$ could be $h_t \xrightarrow{r_1} e \xrightarrow{r_2} t_t$, where r_1, r_2 is a relation template, and e is an entity satisfying certain semantic constraints. The steps involve using the Path Ranking Algorithm (PRA) to generate candidate relation paths. Next, they leverage *semantic constraints* by selecting entities for the indicative paths that adhere to the domain and range constraints of the relations involved. The selection is carried out using a gradient-based search technique to find the indicative paths that maximize the prediction score for the target triple h_t, r_t, t_t. By ensuring that the added paths align with semantic constraints and maximize the plausibility of the malicious triple, their approach not only remains stealthy but also effectively biases the model's predictions towards the attacker's objective. This method is validated through extensive evaluations on benchmark datasets, demonstrating its effectiveness in achieving a high attack success rate under various opaque-box settings.

6.1.3 Surrogate Model-based Attack

Building on the approaches discussed in the previous sections, where gradient-based and attribution-based methods target specific triples or entities in the knowledge graph, surrogate model-based attacks introduce an alternative perspective. Instead of directly manipulating the embeddings or leveraging inductive patterns, these attacks employ an intermediate surrogate model to simulate the behavior of the original KGE model. By doing so, they enable the attacker to optimize adversarial manipulations in a more computationally efficient manner, particularly for downstream tasks where KGE models are used to answer user queries.

Xi et al. [108] introduced ROAR, an attack strategy designed to attack KGE models through both knowledge graph poisoning and *query misguiding*. ROAR particularly focuses on downstream applications where KGEs provide answers to user queries. The goal of the attack is to manipulate the response to a specific query by poisoning the knowledge graph in a manner that maximizes the probability of the targeted fact being true. The attack begins by generating a *surrogate knowledge graph* $\mathcal{G}'$ from the original one. This surrogate graph is used to build a *surrogate knowledge graph reasoner*, which consists of a surrogate embedding function ϕ' and a transformation function ψ. These functions are trained on a set of question-answer pairs sampled from $\mathcal{G}'$. The challenge here is that directly searching for poisoning facts that make the targeted fact true in the discrete space of the knowledge graph is computationally expensive.

To overcome this, the authors first employ *latent space optimization*. They search for an *anchor entity* connected to the target fact and identify facts in the embedding space which, when added, increase the probability of the targeted fact. These potential additions to the graph are gathered in a set of embeddings $\{\mathbf{h}_i, \mathbf{r}_i, \mathbf{t}_i\}_{i=1}^{N}$. Next, the effectiveness of adding each potential fact is assessed using a *fitness score*, which indicates how much each fact's addition would increase the plausibility of the target fact. Based on this score, the top n_g facts are selected for addition to the knowledge graph. This selection process ensures that only the most influential facts are included in the poisoning attack, thereby maximizing the impact on the targeted queries. This two-step process of latent space optimization followed by fitness-based selection makes ROAR a highly adaptable and effective adversarial attack against KGEs, especially in scenarios where downstream applications rely on the knowledge graph for query resolution.

6.1.4 Untargeted Attack

Apart from the adversarial attacks primarily focusing on making the KGE model perform badly on a specific triple, there exists a type of attack aiming to downgrade the overall accuracy of KGE models. This is referred to as untargeted attacks [57, 129], and so far only a few works have considered this. To this end, Zhao et al. [129] proposed a logic-rule-driven framework for conducting untargeted adversarial attacks on knowledge graph embeddings. The key idea herein is to perform adversarial additions or deletions that can systematically degrade overall model performance. To achieve this, the authors exploit logic rules that summarize global structural patterns in a KG. First, they use NCRL, a neural rule learning method [25], to extract high- and low-confidence rules from the graph. Based on these rules, they design two attack strategies, namely adversarial deletion and addition. In adversarial deletion, triples that strongly support high-confidence rules are removed, breaking reliable structural dependencies and preventing the model from learning accurate regularities. In contrast, in the case of addition, low-confidence rules are deliberately corrupted into non-existing rules which are then used to generate noisy triples. This then distorts the KG's semantics and encourages the model to capture misleading patterns. Therefore, the attacks do not focus on a specific target fact or triple, rather aim to disrupt the overall performance of the KGE models in the underlying tasks.

Based on a similar idea of performing untargeted attacks, Kapoor et al. [57] studied the robustness of KGE models considering three different *attack surfaces*, namely graph, parameter, and the labels. To this end, they first consider the knowledge graph perturbation, wherein a subset of triples from the KG is randomly modified by replacing either the head entity or the relation with another from the graph. This introduces structural inconsistencies without introducing new entities or relations. In parameter perturbation, embedding vectors are considered where the noise vectors are added directly to a subset of entity or relation embeddings during training. This is similar to the adversarial attacks like [5, 60, 67] where an attacker gains limited access to model parameters and subtly corrupts the representation space. Finally, in label perturbation, the label vectors used in training are inverted, flipping positives to negatives and vice versa.

6.1.5 Adversarial Training

Most of the works studied adversarial attack approaches for knowledge graph embedding models. To this end, we could find only the work by Zhang et al. [126] that focused on developing a defence approach against such attacks. They proposed a two-fold approach to improve the robustness of KGE models against adversarial perturbations. Firstly, by considering the adversarial training approach using GAN, the approach uses a generator–discriminator setup where the generator proposes adversarial perturbation triples and the discriminator learns to distinguish true from

perturbed triples. This forces the KGE model to become more resilient by directly training on adversarially crafted negatives. In the second step, to filter malicious triples from the graph, the authors propose subgraph-based detection methods. They focus on subgraphs around target triples, apply link prediction scores, and compare outputs of models trained on different subgraph partitions. This approach generates candidate completions from clean subgraphs to identify likely adversarial additions.

6.2 Non-adversarial Robustness

While adversarial robustness focuses on defending against malicious attacks, non-adversarial robustness concerns the model's resilience to naturally occurring noise and inconsistencies in KGs as defined in Equation 11. Real-world KGs are often incomplete, contain errors, and exhibit conflicting information due to the diverse sources from which they are constructed. A robust KGE model should be able to handle these imperfections without significantly compromising its performance. Several approaches have been proposed to improve the robustness of KGE models under noisy KGs, ranging from confidence score-based methods to GAN-based frameworks, reinforcement learning techniques, multi-modal approach, and ensemble approach [26, 70, 71, 75, 86, 87, 98, 111, 128, 130]. Below we discuss the existing works considering these approaches.

6.2.1 Confidence Score-based Approaches

Confidence score-based approaches have been proposed to enhance the robustness of the KGE models by quantifying the reliability of each triple within the KG. These methods assign a confidence score, trustworthiness value, or distance-based measure to each triple, allowing the model to prioritize more reliable data during training [71, 75, 86, 87, 111, 130]. The confidence scores guide the learning process, helping the model to distinguish between correct and noisy triples, thus reducing the impact of inaccuracies present in real-world KGs. In this section, we discuss several works that introduce different mechanisms for computing and utilizing confidence scores to improve the robustness of KGE models. These mechanisms range from local and global confidence scores to trustworthiness evaluations and distance-based assessments.

6.2.1.1 Local and Global Confidence Score

Xie et al. [111] introduced one of the earliest methods to address noise in knowledge graphs by developing KGE models that are robust to such noise. They proposed a novel approach known as the confidence-aware knowledge representation learning (CKRL) framework, which assigns a confidence score to each triple in the KG. This score indicates the correctness and significance of each triple, allowing the model to prioritize more reliable triples during learning. Their model builds upon the translation-based KGE approach, specifically utilizing TransE [16], as the scoring function ϕ_Θ. The standard margin-based ranking loss function [22] was modified to incorporate the confidence scores of triples. The revised objective function aims to minimize the impact of noisy triples by giving higher importance to more reliable triples. Specifically, they introduced the confidence-aware loss function:

$$\sum_{(h,r,t)\in S^+} \sum_{(h,r,x)\in S^-} [\gamma + \phi_\Theta(h,r,t) - \phi_\Theta(h,r,x)] \cdot C(h,r,t),$$

where γ is the margin, and S^+, S^- are the sets of positive and negative triples, respectively. Here, $C(h,r,t)$ is the confidence score for the triple h,r,t. A higher confidence score signals that the model should prioritize this triple during training. In essence, triples with lower scores are weighted less, which are essentially considered as noisy.

The computation of the confidence score $C(h, r, t)$ involves two components: local and global confidence scores as described below.

- **Local confidence score.** This score evaluates how well a triple conforms to the translation assumption within the KGE model. The triple's quality is updated iteratively during training. If a triple does not align with the translation rule, its confidence decreases by a geometric rate α. Conversely, if it does align, the confidence increases at a constant rate β. This iterative adjustment ensures that the confidence scores reflect the quality of triples over time.
- **Global confidence score.** Global confidence scores assess a triple's reliability by analyzing its broader structural context in the knowledge graph (KG). It consists of prior path confidence (PP), and adaptive path confidence (AP). PP measures how often a relation co-occurs with multi-step paths connecting the same entities. If similar paths frequently support the relation, PP is high. AP learns semantic similarity between a relation and its multi-step paths using embeddings. If a path relates to the target relation, AP is computed as high.

By combining these scores, CKRL effectively learns embeddings while simultaneously detecting and mitigating the influence of noise in the KG. This pioneering work laid the foundation for later developments in confidence-aware KGE models.

In a later work, Shan et al. [86] argued that the confidence score mechanism proposed by Xie et al. [111] could lead to the *zero loss problem.* This issue occurs when the negative triples sampled during training quickly fall outside the margin in the ranking loss function, resulting in zero loss. When this happens, the negative triples cease to contribute to refining the model's embeddings, leading to slow convergence, reduced accuracy, and diminished effectiveness in detecting noise within the knowledge graph. To address this problem, Shan et al. introduced a novel *confidence-aware negative sampling method.* They proposed a mechanism to assign a confidence score not just to positive triples but also to negative triples, with the goal of identifying high-quality negative triples that could contribute more significantly to the model's learning process. The key idea is to incorporate the confidence scores of negative triples into the training process.

Shao et al. [87] extended the confidence score-based methods by introducing a novel framework called DSKRL (Dissimilarity-Support-Aware Knowledge Representation Learning) to handle noise in KGs more effectively. Their approach incorporates two main components: triple dissimilarity and triple support, leveraging both structural and auxiliary information in KGs. While the former measures how well the entities and relations in a triple match, using entity hierarchical types and relation paths, the latter combines local and dynamic path support to assess a triple's credibility. After computing both the dissimilarity estimator and triple support, they are combined to improve the noise resilience in KGE models.

6.2.1.2 Trustworthiness Score

While confidence score-based methods focus on quantifying the reliability of individual triples through local and global assessments, trustworthiness score approaches extend this concept by leveraging semantic information and structural properties of entities and relations within the Knowledge Graph (KG). These methods aim to evaluate entities' inherent credibility and associations, refining the training process to prioritize trustworthy information. More specifically, such approaches differ from traditional confidence scores by incorporating additional semantic and contextual cues, such as entity types, descriptions, and path-based correlations. This integration allows for a more nuanced understanding of the data, enabling KGE models to better handle noise and inconsistencies in real-world KGs.

In [130], Zhao et al. proposed TransT, a method to compute the *trustworthiness value* of a triple by leveraging *entity types* and *descriptions*. The key idea is that certain entity types are more credible for specific relations. For example, a living entity (e.g., `/people/person`) is a more suitable subject for `was_born_in` than a non-living one (e.g., `/book/written_work`). TransT quantifies trustworthiness using two components:

- **Entity type trustiness (TT)** measures type compatibility for a relation:

$$TT(h,r,t) = \frac{1}{Z} \sum_{(h_i,t_i)\in\mathcal{T}(r)} \exp(-d(h_i,r,t_i)),$$

  where $\mathcal{T}(r)$ is the set of valid type pairs, and d measures alignment between types and relations.
- **Entity description trustiness (DT)** captures semantic consistency using cosine similarity:

$$DT(h,r,t) = \cos(\mathbf{d}_h + \mathbf{r}, \mathbf{d}_t),$$

where $\mathbf{d}_h$ and $\mathbf{d}_t$ are entity description embeddings, and $\mathbf{r}$ is the relation vector. The final trustworthiness score is a weighted combination:

$$T(h,r,t) = \alpha \cdot TT(h,r,t) + \beta \cdot DT(h,r,t),$$

where α and β control the contributions of the two factors. This trustworthiness score is integrated into the knowledge graph embedding model, prioritizing reliable triples during training.

While TransT focuses on assessing trustworthiness at the entity level, leveraging type compatibility and semantic descriptions, Ma et al. [71] take a structural approach with PTrustE by evaluating path trustworthiness and triple embeddings. Instead of relying solely on entity-level attributes, PTrustE incorporates path-based reasoning to detect noisy triples, capturing both local and global structural features within the knowledge graph. More specifically, given a triple (h,r,t), PTrustE first searches all paths between the head entity h and the tail entity t. Each path consists of a series of intermediate entities and relations, which are then used to compute both local and global trustworthiness scores. Specifically, two types of trustworthiness are introduced. local triple trustworthiness and global triple trustworthiness. In the absence of connecting paths, the confidence score of the triple relies more heavily on the local trustworthiness score derived from triple embeddings rather than path-based features. More specifically, PTrustE evaluates whether h and t are structurally disconnected or if they exist in separate KG components. The triple is likely to be erroneous if the entities belong to isolated graph fragments. In such cases, embedding-based similarity and logical constraints from the KG are used to assess plausibility, rather than path-based reasoning.

PTrustE focuses on detecting noise in KGs by leveraging path trustworthiness and probabilistic logic, since it primarily aims to filter out incorrect triples before embedding learning. An alternative approach to handling noisy triples is to directly modify the training objective rather than discarding them outright. Nayyeri et al. [75] introduced a modification to the marginal ranking loss function to handle noisy data in knowledge graphs (KGs), particularly focusing on incorrect triples. Their approach does not build on the previous confidence score-based works but instead introduces a distance-based strategy to identify and manage noisy triples effectively. In their method, the authors define separate objective functions for positive and negative triples and then combine them into a unified loss function. One key component of their approach is a distance function, which intuitively measures the likelihood of a triple being correct or noisy. During the optimization process, this distance is constrained to lie within the range $[0,\gamma]$, where γ serves as a discriminator that separates positive and negative triples. A probability function is employed to assign a score

based on the computed distance. A high probability indicates a high likelihood of the triple being incorrect (noisy), whereas a lower probability suggests a higher confidence in the triple's correctness. The objective is to minimize the overall loss by maximizing the likelihood of correct triples and minimizing the likelihood of noisy ones.

To summarize, confidence score-based methods improve the performance of KGE models by adjusting the influence of noisy triples during training. These methods vary in how they compute and integrate confidence scores, leveraging different aspects of local consistency, global structural reasoning, and adaptive loss functions. Early approaches like CKRL [111] introduced confidence-aware learning by assigning local and global confidence scores to triples, refining embeddings iteratively. However, CKRL suffered from the zero-loss problem, where negative triples quickly became uninformative. To mitigate this, Shan et al. [86] proposed a confidence-aware negative sampling strategy, dynamically selecting high-quality negative triples to improve training effectiveness. Expanding beyond embeddings, DSKRL [87] integrated semantic knowledge, such as entity types and relation paths, to compute confidence scores, improving robustness against inconsistencies but requiring additional structured information. Alternatively, PTrustE [71] introduced a path-based trustworthiness framework, assessing global and local triple reliability through correlation networks and probabilistic logic. While effective in structured graphs, PTrustE is computationally expensive and less applicable to sparse KGs. A distinct approach was taken by Nayyeri et al. [75], who modified the ranking loss function to incorporate a distance-based confidence score, adjusting training weights dynamically instead of explicitly assigning confidence scores. This method avoids reliance on heuristic scoring functions but requires fine-tuning distance thresholds for optimal performance. Embedding-based approaches are computationally efficient but can struggle with noisy negatives, whereas semantic-aware models (DSKRL) improve interpretability but depend on auxiliary knowledge. Path-based trustworthiness methods (PTrustE) enhance global reasoning but introduce high complexity, and distance-based confidence models (Nayyeri et al.) provide a principled alternative at the cost of hyperparameter sensitivity. The optimal choice, therefore, depends on dataset characteristics, noise levels, and computational constraints.

6.2.2 GAN-based Approaches

While confidence score and trustworthiness score approaches address noise by quantifying the reliability of triples or entities based on structural and semantic properties, Generative Adversarial Network (GAN)-based approaches adopt a more dynamic mechanism. These methods introduce an adversarial framework to detect and mitigate noise in Knowledge Graphs (KGs) by simultaneously learning to generate and classify noisy triples. More specifically, GAN-based approaches leverage adversarial training to refine the embeddings by detecting and mitigating the impact of noisy triples during training. By generating synthetic noisy triples and training the model to differentiate between true and noisy triples, these methods ensure that the learned embeddings remain robust.

NoiGAN [26] extends the idea of confidence score proposed in Section 6.2.1. They argued, similar to the previously described approaches, that using only the confidence score as an indication of how well a triple fits to the KGE model might lead to bias and uncertainty. Therefore, the confidence score $C(h, r, t)$ in this work is learned by using a *generator* and *discriminator* as a generative adversarial network (GAN). More specifically, they proposed a learning framework inspired by the adversarial training [8, 63, 73] methods. In the GAN framework, NoiGAN consists of two main components: a generator and a discriminator. The generator is designed to generate noisy triples, while the discriminator is trained to distinguish between true and noisy triples, ultimately computing the confidence score for each triple. During training, the KGE model uses this confidence score as a guiding signal to eliminate noisy data.

Given a true triple (h, r, t), the generator generates a noisy triple (h', r, t') from an initially generated negative sample candidate set $\mathcal{N}(h, r, t)$. This is achieved through a neural network that takes as input the embedding vectors of the triple (h', r, t') and outputs a probability indicating the plausibility of the triple being noisy. More formally, the generator aims to maximize the expected reward:

$$R_G = \sum_{(h,r,t)} \mathbb{E}_{(h',r,t') \sim G(\cdot|(h,r,t);\Theta_G)}[\log f_D(h', r, t')],$$

where $f_D(h', r, t')$ is the probability predicted by the discriminator that the generated triple (h', r, t') is true. The generator uses reinforcement learning to generate triples that can effectively fool the discriminator. The discriminator, on the other hand, acts as a noisy triple classifier. It aims to distinguish between true triples and noisy triples generated by the generator.

Apart from the embedding models, the GAN-based approach is also used in KG-based systems such as in the entity-alignment approach. To this end, Pie et al. [78] propose an approach to make robust cross-lingual entity alignment between KGs by incorporating noise detection into the alignment process using a generative adversarial network (GAN)-based approach [46]. The model consists of a Graph Neural Network (GNN) for entity embedding and a Generative Adversarial Network (GAN) for noise detection. The GAN therein consists of a generator G and a discriminator D. The generator generates fake entity pairs, while the discriminator assigns a trust score $T(e_1, e_2)$ to distinguish correct and noisy pairs. To align entities across KGs, a margin-based ranking loss is used to bring correct entity pairs closer together and push noisy pairs further apart.

6.2.3 Reinforcement Learning Approaches

Reinforcement Learning (RL) approaches take a different perspective by formulating noise detection and triple selection as a decision-making problem. RL-based methods focus on improving the robustness of KGE models by systematically identifying and removing noisy triples before the training process begins. This proactive approach ensures that the KGE models are trained on cleaner datasets, leading to more reliable embeddings. A recent work by Zhang et al. [128] proposes a multi-task reinforcement learning (RL) framework to make the KGE models robust by identifying and removing noisy triples from the training dataset. Unlike previous approaches that directly train on noisy datasets, this method first cleans the dataset before the training process, ensuring that the KGE models are learned on a noise-free graph. The authors define the state, action, reward, and the objective of the RL framework in the following manner.

State. Each state in RL is represented as the set of triples that have already been selected as clean and the current triple that is under consideration. Mathematically, the state at time step t can be defined as $s_t = (T_{\text{selected}}, (h, r, t))$, where T_{selected} is the set of triples that have already been marked as clean up to time t, and (h, r, t) is the triple being evaluated.

Action. The RL agent takes an action to either select or reject the triple (h, r, t). The action space A consists of binary decisions, $A = \{0, 1\}$, where 1 indicates selecting the triple as clean, and 0 indicates rejecting it.

Reward. The reward function is designed based on the scoring functions of multiple KGE models like TransE, DistMult, ConvE, or RotatE, along with a heuristic term that encourages the model to select more triples. The reward R for a set of selected triples T_{selected} is calculated as

$$R = \frac{1}{|T_{\text{selected}}|} \sum_{(h,r,t) \in T_{\text{selected}}} \phi_\Theta(h, r, t) + \alpha \frac{|T_{\text{selected}}|}{|T_{\text{total}}|},$$

where α is a hyperparameter, and $|T_{\text{total}}|$ is the total number of triples in the KG.

Objective. The aim of the RL model is to maximize the expected reward by selecting those triples that exhibit higher plausibility.

The authors highlight that this approach has the potential drawback of filtering out a large number of triples, which could include some correct triples. However, the RL framework's use of scoring functions from different KGE models helps to mitigate this by making decisions based on the inferred relationships and plausibility scores.

6.2.4 Multi-modal Knowledge Representation

Multi-modal methods aim to combine information from different knowledge sources to better capture the semantics and context of entities and relations within the knowledge graph. This integration enables the model to mitigate the effects of noise in a single modality by relying on complementary information from other modalities. To this end, the work closest to the idea of robustness of KGE models is done by Lu et al. [70] where they propose multi-modal knowledge representation learning (MMKRL) to generate robust KGE models. The idea therein is to use several knowledge such as textual knowledge, entity description, visual knowledge to generate the embedding [95, 105, 110]. MMKRL essentially consists of two main modules: knowledge reconstruction and adversarial training, where the knowledge reconstruction module aligns and integrates various knowledge embeddings to reconstruct multi-modal knowledge graphs, while the training module enhances robustness and performance using adversarial strategies.

6.2.5 Ensemble Approaches

Ensemble-based approaches combine multiple models trained on diverse subgraphs of the Knowledge Graph (KG). This strategy leverages the principle that an ensemble of learners can outperform individual models, especially in the presence of noise or inconsistencies in the data. By aggregating predictions from multiple models, ensemble-based approaches mitigate the impact of errors or biases present in a single model. Wan et al. [98] proposed an ensemble-based approach to enhance the robustness of the KGE models. Their method involves generating a set of diverse subgraphs from a given KG $\mathcal{G}$ and training an individual base learner for each subgraph.

Due to the complexity of KGs, traditional graph sampling methods are not directly applicable. To address this, Wan et al. employ a random walk-based approach [69] to sample meaningful subgraphs. The random walk process starts by selecting an initial fact (h, r, t) uniformly at random from the KG $\mathcal{G}$. Then, the random walk samples a neighbor of the current node, following the relations in the KG. This sampling continues until a predefined boundary condition L (e.g., a maximum path length or number of nodes) is met. After executing multiple random walks, a set of subgraphs $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$ is generated. For each subgraph $\mathcal{G}_i$, a shallow KGE model ϕ_{Θ_i} is trained independently to obtain entity and relation embeddings. The model's goal is to learn an embedding function $\phi_{\Theta_i}(h, r, t)$ that maximizes the plausibility of triples in the subgraph. The final ensemble model combines the outputs of these n base learners. Let ϕ_{ensemble} represent the final embedding function, which is defined as a weighted combination of the individual models:

$$\phi_{\text{ensemble}}(h, r, t) = \sum_{i=1}^{n} \alpha_i \phi_{\Theta_i}(h, r, t),$$

where α_i is the weight assigned to model ϕ_{Θ_i} based on its prediction performance. The weights α_i are determined by an uncertainty measure, which reflects the predictive capability of each model on its corresponding subgraph. For example, the uncertainty can be calculated using entropy or variance in the predictions. The robustness of this ensemble approach is then evaluated by injecting noise into the KG $\mathcal{G}$. Wan et al. demonstrate that their ensemble model performs significantly better than individual KGE models in the presence of noisy triples.

6.3 Comparative Analysis and Future Directions

The previous sections explored various approaches to enhance the robustness of KGE models when KGs contain noisy triples. These approaches differ in how they detect and mitigate noise, with some focusing on explicit confidence estimation, others leveraging adversarial learning, and some incorporating external multimodal information. In this section, we contrast these methods, analyze their respective strengths and limitations, and propose potential hybrid strategies to further enhance resilience. To better understand how different robustness approaches complement or compete with one another, first of all, we categorize them based on key aspects such as information used for noise handling, adaptability to different types of noise, and computational complexity as summarized in Table 2.

Table 2 Comparison of different noise-robust KGE approaches.

Approach	Key Mechanism	Strengths	Limitations
Confidence score	Assigns confidence scores to triples based on local/global plausibility	Adaptive to structured noise, interpretable	Struggles with adversarial noise, requires careful calibration
Trust score	Uses entity type and path-based information to determine trustworthiness	Strong semantic reasoning, robust to inconsistencies	Relies on well-defined entity types, limited adaptability
GAN	Generator-discriminator model to filter noise iteratively	Dynamically adapts to different noise patterns	Training instability, risk of mode collapse
RL-based	Reinforcement learning selects reliable triples pretraining	Generalizes well, avoids overfitting to noise	Filtering errors can lead to knowledge loss
Multi-modal	Uses text and images to supplement KG information	Effective for missing or ambiguous data	Requires external data sources, computationally expensive
Ensemble	Aggregates predictions from multiple KGE models	Improves generalization and robustness	Computational overhead, limited effect in adversarial settings

Furthermore, given the strengths and limitations of individual methods, a promising direction is to develop hybrid approaches that integrate complementary techniques. Below, we suggest three strategies to enhance robustness by combining different noise-handling mechanisms.

Confidence score with GAN-based noise correction. Confidence-based methods provide an effective first step in detecting structured noise, while GAN-based filtering adapts dynamically to unstructured noise. A potential hybrid model could first use a confidence estimator, such as CKRL, to assign preliminary confidence scores to triples. These confidence scores help distinguish between highly reliable triples and those suspected as being noisy. Once the confidence scores are assigned, the high-confidence triples can be fed directly into a standard KGE training process to learn robust embeddings from cleaner data. Meanwhile, the low-confidence triples, which are more likely to contain noise, are passed into a GAN-based filtering mechanism. The GAN consists of a generator that produces synthetic noise and a discriminator that learns to distinguish between correct and incorrect triples. During training, the discriminator iteratively refines its decision boundary by learning from both real and generated noisy triples.

RL-based filtering with multi-modal learning. RL-based models excel at identifying and removing highly noisy triples, making them suitable to use as an initial data-cleaning step before training a multi-modal KGE model. The combination of RL and multi-modal learning allows for more effective noise reduction while leveraging complementary knowledge sources to enhance embeddings. For instance, an RL agent could be trained to evaluate triples based on a reward function that incorporates multiple KGE scoring functions. The agent iteratively selects high-confidence triples while discarding unreliable ones. Once the RL agent filters out noisy triples, a multi-modal KGE model is trained exclusively on the cleaned dataset. This model integrates information from textual descriptions, entity attributes, and visual embeddings to improve the quality of entity and relation representations. The multi-modal embeddings can further be used to refine the RL-based filtering in a feedback loop. If a previously filtered triple gains support from external modalities (e.g., a missing relation is inferred via textual descriptions), it may be reintroduced into the knowledge graph.

Ensemble learning with path-based trustworthiness scores. Ensemble learning enhances robustness by aggregating predictions from multiple KGE models, while path-based trustworthiness scoring ensures that models are weighted based on their reliability in capturing meaningful entity-relation patterns. Instead of training a single KGE model, multiple models are trained on different subgraphs generated through random walks, clustering-based sampling, or relation-specific partitions. Thereafter, each entity pair in the KG is evaluated based on the reliability of intermediate paths connecting them. Approaches like PTrustE [71] could be used to score paths based on semantic consistency, redundancy, and coherence with established entity-type constraints. The final embedding for a given entity or relation is determined by aggregating the predictions from the ensemble models, weighted according to their path-based trustworthiness scores. Models that perform better on structurally supported paths contribute more to the final representation. These hybrid approaches offer promising directions for improving the robustness of KGE models.

7 Robustness of KG-based Systems

There are some works that do not directly discuss the robustness of the KGE models; however, they consider the KG-driven systems, such as entity linking [72], cross-lingual entity alignment [78], knowledge-grounded dialogue system [100], improving the robustness of the facts of KG [109].

Mao et al. [72] propose a robust entity linking method that tackles 3 aspects, namely, inefficient graph encoders, the need for negative sampling, and catastrophic forgetting in semi-supervised learning. To improve the graph encoders therein, they use relational attention to update the entity features. Furthermore, the authors prove that negative samples are unnecessary in entity linking. It adopts a symmetric negative-free alignment loss to align entity pairs without generating negative samples thereby removing the need for negative samples, and aligning entity pairs with the loss function. Finally, to mitigate catastrophic forgetting, the approach stores previously learned embeddings and selectively reviews them during each training iteration. This approach allows the model to maintain alignment accuracy without retraining on all previous data. The evaluation has shown state-of-the-art results with improved robustness.

Pei et al. [78] propose REA (Robust Entity Alignment), a method for cross-lingual entity alignment between noisy knowledge graphs (KGs). Existing entity alignment models assume clean labeled data, but in real-world scenarios, labeled entity pairs often contain errors that degrade the alignment quality. REA first encodes the structure of knowledge graphs using a Graph Neural Network (GNN). The GNN processes entities and their relationships within each knowledge graph, learning meaningful embeddings that capture the structural similarities between

entities, even if they exist in different languages. For example, if "Eiffel Tower" in an English knowledge graph has the same connections as "Tour d'Eiffel" in a French knowledge graph, their embeddings should be similar. The approach further introduces a trust score for each labeled entity pair. The trust score acts as a measure of confidence, determining how reliable a given entity alignment is. REA also uses a margin-based ranking loss function. This function ensures that correctly aligned entity pairs have their embeddings placed closer together, while incorrect pairs are pushed further apart in the learned space. The noise detection module within REA, which operates using an adversarial training framework, continuously updates the trust scores based on newly identified errors. In turn, the noise-aware entity alignment module adjusts the entity embeddings based on these refined trust scores. This iterative learning process ensures that the model becomes increasingly accurate, filtering out noise while improving the quality of entity alignment. An extensive evaluation on real-world multilingual knowledge graph datasets such as DBP15K, DWY100K [93] show that REA outperforms state-of-the-art methods (e.g., GCN-Align [104], MuGNN [19]) in noisy settings. REA provides a robust approach for integrating multilingual knowledge graphs, ensuring high-quality entity alignment despite label noise.

Wang et al. [100] introduce an entity-based contrastive learning framework, named EnCo, to enhance the robustness of knowledge-grounded dialogue (KGD) systems. Given a dialogue context $C = \{u_1, u_2, \ldots, u_{n-1}\}$ consisting of utterances u_i and an external knowledge set $K = \{(h_1, r_1, t_1), \ldots, (h_m, r_m, t_m)\}$ comprising knowledge triples where h_i, r_i, and t_i represent the head entity, relation, and tail entity respectively, the goal of a KGD system is to generate a response u_n based on C and K. The authors aim to enhance the robustness of KGD models to handle real-world perturbations, including semantic-irrelevant (e.g., misspellings, paraphrasing) and semantic-relevant (e.g., incorrect entity replacements) perturbations. To this end, they leverage contrastive learning to improve robustness by constructing positive and negative samples and training the model to recognize semantic similarities and differences.

Xiao et al. [109] address the problem of evaluating the robustness of *outstanding* facts (OFs) derived from KGs. An OF is defined as a statement highlighting how an entity stands out based on specific attributes when compared to its peers. Consider a KG $\mathcal{G}$ containing information about universities and their employees, including attributes like gender. An OF from this KG might state: "At the American Council on Education (ACE), only 31% of the employees are male." This statement could suggest a notable gender disparity at ACE-affiliated institutions. However, the robustness of this fact needs to be evaluated by considering the broader context and possible data variations. To formalize this, Xiao et al. introduce the concept of robustness by analyzing how the "strikingness" of an OF changes under various perturbations. Let $\mathcal{S}(f)$ denote the strikingness of an OF f in a given context. The goal is to ensure that $\mathcal{S}(f)$ remains consistent even when the context or data changes slightly. The authors propose two types of perturbations to evaluate this:

- **Entity perturbation.** It assesses the robustness of an OF by replacing its context entity c with a similar entity c'. Formally, let c represent the context entity in the OF f. We replace c with c', where c' is chosen based on its similarity to c. The similarity between entities c and c' is computed as

$$\mathrm{Sim}(c, c') = \frac{|N(c) \cap N(c')|}{|N(c) \cup N(c')|},$$

  where $N(c)$ and $N(c')$ are the sets of neighbors of c and c', respectively.
- **Data perturbation.** It involves modifying the KG by adding or altering edges, thereby changing the peer entity set of the OF. Formally, the relevance of a data perturbation is quantified using a head-tail relevance function, which measures the semantic connection of the newly added edges to the original fact. Given an added edge (h', r', t'), the head-tail relevance function $\mathcal{R}(h', r', t')$ evaluates whether the modification preserves the context's semantic integrity.

The robustness of an OF is then defined by the expected strikingness $\mathbb{E}_{p(\mathcal{P})}[\mathcal{S}(f)]$ over a perturbation relevance distribution (PRD) $p(\mathcal{P})$. This method of evaluating robustness relates to earlier discussions in the literature on robustness, specifically, similar to our proposed robustness formalization in Equation 10. Much like ensuring that KGE models are resilient against adversarial attacks and noise (e.g., as described in works like Xie et al. [111] and Shan et al. [86]), evaluating OFs for robustness ensures that their interpretations remain valid across different contexts.

8 Robustness Improvement Using Knowledge Graphs

Note that, similar to using KGs to improve the resilience of several systems, there also exist a number of works that use KGs [83, 118] and KGE models [61] to improve the robustness of ML models. However, the notion of robustness therein pertains to the effectiveness of performing the underlying tasks. Below we describe some of them.

Multi-object detection. Lang et al. [61] propose the use of KGEs to develop more robust multi-object detection models. The main idea is to use KGEs to incorporate semantic knowledge into object detection, aiming to achieve more structured and semantically grounded predictions. Traditional object detection models often use a one-hot encoding approach, treating object classes as discrete and unrelated. This method maximizes inter-class distances but ignores the semantic relationships between different object types. The authors therein introduce a new formulation where they replace these learnable class prototypes with fixed object type embeddings derived from knowledge graphs. Specifically, the object detector learns to map visual features into a semantic embedding space, using either word embeddings (like GloVe) [79] or embeddings derived directly from knowledge graphs using any standard KGE models. In their evaluation, this approach demonstrated more semantically grounded misclassifications, meaning the errors made by the model were often more contextually appropriate. Additionally, their evaluation on benchmark datasets showed that KGE-based models matched or even outperformed traditional one-hot methods, particularly in challenging object detection benchmarks.

Deep learning. Radtke et al. [83] propose using KGs to enhance deep learning models for fault diagnostics in prognostics and health management (PHM). They introduce a KG-enhanced deep learning approach to incorporate domain-invariant knowledge, improving model robustness and generalization. The method leverages the structure of KGs to encode semantic information hierarchically and combines this with supervised contrastive learning to create a more stable feature representation. Experimental results demonstrate that this approach increases the model's ability to handle domain shifts, making fault diagnostics more resilient across varying conditions.

Recommender system. Yang et al. [118] propose knowledge graph contrastive learning (KGCL) to suppress noise and enhance item representations in recommender systems. Their approach addresses challenges such as *long-tail entity distributions* and *noisy, topic-irrelevant connections* in Knowledge Graphs (KGs). More specifically, to improve robustness, KGCL generates two perturbed views of the KG, $\mathcal{G}_1$ and $\mathcal{G}_2$, by randomly dropping edges. This introduces structural perturbations, allowing the model to learn robust embeddings by contrasting entity representations across different views. KGCL then employs contrastive learning to maximize agreement between the same entity's embeddings in different views while minimizing similarity with other entities that are not close.

9 In-distribution Generalization

The work by Zhu et al. [133] is the only work that contributes to this aspect of resilience. Therein, they define this as *predictive multiplicity*, a phenomenon where multiple models with similar accuracy make conflicting predictions for the same query. The authors conduct an empirical study on multiple KGE models and datasets to measure predictive multiplicity. For each KGE algorithm, they train multiple models with different random initializations and hyperparameters. They then select a set of "competing" models – those whose link prediction performance is virtually the same as a best baseline model (within a small tolerance ϵ, e.g. 1% difference in Hits@K). Using this set of models, the authors evaluate how often their predictions diverge. For each test query (a partially specified triple such as $(h, r, ?)$), they compare the model's top-ranked results. If one model's top answer is different from another's, that query is counted as a conflicting case. The ambiguity metric is computed as the percentage of test queries with any such conflict, and discrepancy reflects the maximum disagreement rate among the models. These metrics provide a quantitative measure of predictive multiplicity for the link prediction task.

After measuring the extent of conflicting predictions, the authors apply ensemble voting methods to combine model outputs. Each model in the competing set produces a ranked list of candidate entities for a query. The authors apply three voting schemes to aggregate these rankings into one result,

1. majority voting which picks the candidate that appears as the top choice for most models,
2. borda voting, which assigns points based on rank position (e.g., a candidate gets more points for being ranked 1st, slightly fewer for 2nd, and so on, across all models) and then selects the candidate with the highest total points, and
3. range voting, which uses the actual prediction scores from each model (rescaled to a common range) and sums them up for each candidate.

These methods generate an aggregated ranking intended to reflect a consensus. The impact of aggregation is assessed by recomputing the ambiguity and discrepancy metrics on the combined ranking, and by checking the standard accuracy metrics (Hits@K) to ensure that the ensemble prediction is still performing well.

10 Challenges and Future Works

Future research in the domain of resilience on knowledge graphs and KGE models presents a number of possibilities to improve different aspects of resilience that we defined in this work. We can envisage works aiming at developing KGE models considering generalization consistency, distribution adaptation, and in-distribution generalization, amongst others. We describe future work directions in more detail in the following.

Generalization under Distribution Shift. One promising avenue for future work is the development of resilient KGE models that can adaptively adjust to changes in the underlying data or graph structure. Traditional KGE models often assume static or stationary environments, which may not hold in dynamic or evolving KGs. Future research could explore dynamic embedding techniques that continuously update entity and relation embeddings to capture temporal or contextual changes, for instance, when new entities or relations are added to the KG. Additionally, integrating uncertainty modeling and probabilistic reasoning mechanisms into KGE models could enhance their resilience to noisy or uncertain data. Some existing works attempt to quantify uncertainty in KGE models, such as *probabilistic soft logic-based methods* in [23] and confidence-aware embedding techniques in [66]. The work in [23] employs probabilistic soft logic to generate confidence scores capturing structural and assertional

uncertainties, enabling the model to provide confidence-based predictions when new entities and relations are introduced. Similarly, [66] defines a KG as uncertain when each assertion is associated with a confidence score, which is integrated into the KGE learning process to adjust predictions dynamically. These studies establish an important foundation for making KGE models aware of distribution shifts by incorporating uncertainty estimation.

To further enhance robustness against distribution shifts, conformal prediction could be incorporated into KGE models. Conformal prediction provides a mathematically sound framework for quantifying the uncertainty of model predictions by constructing prediction sets that offer guaranteed coverage probabilities [44]. Instead of generating point estimates, KGE models could output prediction intervals for link prediction tasks, ensuring that the true answer is included within a certain confidence level. Note that there already exist some works incorporating this technique in KGE models, such as [134] where the authors apply conformal prediction theory, which enables uncertainty-aware answer set prediction by ensuring that the correct answer is included within a generated answer set with probabilistic guarantees. This shows the potential of utilizing conformal prediction in KGE models in dealing with resilience. For instance, adaptive conformal prediction techniques could be applied to KG completion tasks, where the KGE model dynamically updates its uncertainty estimates as new data arrives. When distribution shifts occur, such as new entities being introduced or relationships evolving, the conformal predictor could adjust its confidence intervals accordingly. This is particularly useful in real-world applications, where decision-making systems rely on KGE models and require calibrated confidence scores for each prediction.

Besides uncertainty-aware learning and conformal prediction, other techniques could be explored to increase the robustness of KGE models in dynamic environments such as:

Bayesian knowledge graph embeddings. Instead of learning fixed embeddings, a Bayesian approach would model entity and relation embeddings as probability distributions (e.g., using Gaussian distributions) [96], allowing the model to express uncertainty in predictions explicitly. This would be particularly effective in scenarios where distribution shifts occur.

Meta-learning for KGE adaptation. A meta-learning framework could be designed to quickly adapt KGE models when distribution shifts occur. Few-shot learning techniques, such as Model-Agnostic Meta-Learning (MAML) [43], could be used to train KGE models to generalize across different graph structures with minimal re-training. This would be beneficial in dynamic knowledge graphs, where new domains or unseen entities frequently appear.

Contrastive learning for distribution shift detection. Contrastive learning techniques could be integrated to detect and quantify shifts in graph structures [45]. By learning embedding distances between past and present snapshots of a KG, models can determine when a significant shift has occurred and retrain embeddings accordingly. This approach could also be combined with self-supervised learning, enabling KGE models to update embeddings without requiring extensive labeled data.

Incorporating conformal prediction, Bayesian embeddings, meta-learning, and contrastive learning into KGE models could significantly enhance their ability to handle distribution shifts and noisy data. Note that, the list of approaches mentioned here is not exhaustive, and there could be further techniques to tackle the aforementioned problems. Future research should delve into exploring more such approaches and find out explore how these techniques can be efficiently integrated into KGE pipelines while ensuring computational scalability and real-time adaptation capabilities.

Adversarial and non-adversarial robustness. Another possible research direction is that of resilience of KGE models against adversarial attacks and manipulations, i.e., developing KGE models that are adversarially robust. As mentioned beforehand, real-world KGs might suffer

from adversarial attacks where adversaries may attempt to exploit vulnerabilities in the KG or KGE models to inject false information, manipulate inference results, or disrupt system functionality. There have already been some works to this end, however, all of them focus on developing methods to perform targeted attacks, i.e., considering a specific fact to add or remove from the KG and thereby making the KGE model learn based on the attackers' goal. To this end, only the KG has been considered as a possible attack surface. However, there can be other possibilities, for instance, the parameters of the already trained KGE model can be attacked. Such kind of attacks are often prevalent in the ML domain and termed Trojan attacks [67, 68, 102] where the attacker aim to make the model learn their objective either by generating inputs with certain *triggers*, or by changing the already trained model's parameters. For KGE models, such Trojan attacks could correspond to the modification of the entries of learned embedding vectors so as to achieve a specific attacker's objective. Apart from considering targeted attacks by taking into account different attack surfaces, it would also be needed to consider performing non-targeted attacks [57], where the idea is to simply disrupt the performance of the underlying KGE models by introducing noise in the KGs or in the KGE models. As mentioned previously, there are already some works which considered such kind of attacks. However, more sophisticated attack approaches to this end could be explored.

Additionally, the targeted attacks so far have been considered only for a specific type of task, namely link prediction tasks. KGE models are used in many critical downstream application tasks [32, 42, 99], and hence, more research is needed to understand how to perform adversarial attacks on such KGE-based tasks. This basically opens up a number of different attack surfaces along with the need to explore different attack dimensions, including non-targeted attacks.

While several works considered adversarial attacks on KGE models, a much-needed direction to be focused on is the development of defence mechanisms against such attacks that can detect and mitigate adversarial attacks in real-time, thereby enhancing the overall robustness of KG-based applications. This would include developing graph-based anomaly detection algorithms to identify and mitigate adversarial attacks or abnormal patterns in the KG, performing adversarial training of KGE algorithms, developing certified guaranteed methods to build robust KGE models, and so on. Furthermore, defence mechanisms should be extended to combat non-targeted attacks, effectively addressing noise, or incompleteness inherent in KGs. This entails the creation of robust data integration and ensemble algorithms capable of handling diverse and noisy information from various sources. Moreover, exploring techniques for automated error detection, correction, and data validation within KGs could significantly enhance their quality and reliability over time.

Recently, the works to combine large language models (LLMs) with KGEs are gaining popularity [77]. There is a potential that by augmenting LLMs to KGEs, one could achieve improved robustness. Leveraging the semantic richness of natural language representations encoded in LLMs, such as BERT [36] or GPT [17], may enhance the understanding and representation of entities and relations in the KG. This integration could potentially mitigate the impact of noisy or incomplete KGs on downstream tasks.

Stability to incomplete inputs. Existing works primarily focus on handling missing data through imputation or data augmentation rather than explicitly ensuring robustness against missingness. The lack of standardized evaluation benchmarks and theoretical formulations of stability further hinders progress in this area. Therefore, to this end, we first of all require a benchmark to evaluate different approaches. Afterward, we require suitable graph similarity metrics depending on different domains, moving beyond standard metrics like graph edit distance or Jaccard similarity. Furthermore, we can envisage using adversarial training [18] where missing elements are simulated during training to improve resilience. To this end, *dropouts* for KGEs, randomly

removing nodes/edges during training could also help. Knowledge distillation technique is quite useful in learning where a model trained on complete KGs transfers knowledge to one dealing with incomplete KGs. Apart from the training techniques, adapting loss functions could also be considered. For instance, a stability-aware loss function could be designed that explicitly penalize drastic embedding changes due to missing data, ensuring bounded divergence within ϵ. To this end, furthermore, a consistency regularization technique, where models minimize differences in predictions from full vs. incomplete graphs could also be used.

Finally, resilience has already been vastly explored in fault-tolerant systems, therefore, interdisciplinary approaches that draw insights from fields such as network science, complex systems theory, and resilience engineering could provide valuable perspectives and methodologies for enhancing the resilience of KGs and KGE models. By leveraging principles from these domains, researchers can develop holistic, multi-faceted strategies for improving the reliability and robustness of KG-based systems in diverse application domains.

11 Conclusion

In this work, we explored the resilience of knowledge graph embedding models, addressing their ability to withstand and adapt to various challenges such as noise, adversarial attacks, and dynamic changes in the underlying knowledge graphs. While significant research has been conducted on robustness, particularly adversarial robustness, there is a pressing need to consider a more comprehensive notion of resilience. This broader understanding includes aspects such as generalization consistency, distribution adaption, and performance stability under diverse real-world conditions. A key finding of this survey is that while adversarial robustness has received considerable attention, with various strategies to perform attacks on KGE models, resilience in non-adversarial contexts is equally critical. Models must not only defend against malicious interventions but also maintain their reliability in the presence of natural noise and inconsistencies prevalent in real-world KGs. The surveyed works on non-adversarial robustness primarily focus on mitigating the effects of noise by incorporating confidence-aware learning and enhanced negative sampling strategies. However, these approaches often overlook the dynamic nature of KGs, particularly temporal and evolving KGs, where distribution shifts are inevitable. Addressing such shifts through adaptive retraining mechanisms remains an open challenge. Moreover, ensuring in-distribution generalization across diverse application domains is essential. KGE models must be able to operate effectively even with incomplete input data, which is a common scenario in real-world applications. Achieving this consistency demands that future research goes beyond traditional robustness frameworks, integrating novel methodologies from graph neural networks, reinforcement learning, and explainable AI to enhance both the adaptability and transparency of KGE models. In conclusion, while much progress has been made in improving the robustness of KGE models, a more holistic approach to resilience – incorporating adaptability, consistency, and robustness in the face of both adversarial and natural challenges – will be key to unlocking the full potential of these models in real-world, dynamic, and noisy environments.

References

1 Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 2704–2729. PMLR, 2022. URL: `https://proceedings.mlr.press/v178/agarwal22b.html`.

2 Saadullah Amin, Stalin Varanasi, Katherine Ann Dunfield, and Günter Neumann. Lowfer: Low-rank bilinear pooling for link prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July*

2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, pages 257–268. PMLR, 2020. URL: http://proceedings.mlr.press/v119/amin20a.html.

3 Amin Anjomshoaa, Hannah Schuster, Johannes Wachs, and Axel Polleres. From data to insights: constructing spatiotemporal knowledge graphs for city resilience use cases. In *Second International Workshop On Linked Data-driven Resilience Research 2023*, 2023.

4 Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007. doi:10.1007/978-3-540-76298-0_52.

5 Jiawang Bai, Baoyuan Wu, Yong Zhang, Yiming Li, Zhifeng Li, and Shu-Tao Xia. Targeted attack against deep neural networks via flipping limited weight bits. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: https://openreview.net/forum?id=iKQAk8a2kM0.

6 Ivana Balažević, Carl Allen, and Timothy M Hospedales. Hypernetwork knowledge graph embeddings. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings 28*, pages 553–565. Springer, 2019.

7 Ivana Balažević, Carl Allen, and Timothy M Hospedales. Tucker: Tensor factorization for knowledge graph completion. *arXiv preprint arXiv:1901.09590*, 2019.

8 Mislav Balunovic and Martin T. Vechev. Adversarial training and provable defenses: Bridging the gap. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=SJxSDxrKDr.

9 Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010. doi:10.1007/S10994-009-5152-4.

10 Christian Berger, Philipp Eichhammer, Hans P. Reiser, Jörg Domaschka, Franz J. Hauck, and Gerhard Habiger. A survey on resilience in the iot: Taxonomy, classification, and discussion of resilience mechanisms. *ACM Comput. Surv.*, 54(7):147:1–147:39, 2022. doi:10.1145/3462513.

11 Peru Bhardwaj, John D. Kelleher, Luca Costabello, and Declan O'Sullivan. Adversarial attacks on knowledge graph embeddings via instance attribution methods. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8225–8239. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.emnlp-main.648.

12 Peru Bhardwaj, John D. Kelleher, Luca Costabello, and Declan O'Sullivan. Poisoning knowledge graph embeddings via relation inference patterns. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1875–1888. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.acl-long.147.

13 Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 81–88. ACM, 2007. doi:10.1145/1273496.1273507.

14 Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008. doi:10.1145/1376616.1376746.

15 Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

16 Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795, 2013. URL: https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

17 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information*

Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

18 Liwei Cai and William Yang Wang. KBGAN: Adversarial learning for knowledge graph embeddings. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1470–1480, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1133.

19 Yixin Cao, Zhiyuan Liu, Chengjiang Li, Juanzi Li, and Tat-Seng Chua. Multi-channel graph neural network for entity alignment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1452–1461, 2019.

20 Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, pages 1306–1313. AAAI Press, 2010. doi:10.1609/AAAI.V24I1.7519.

21 Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545*, 2020. arXiv:2005.00545.

22 Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhiming Ma, and Hang Li. Ranking measures and loss functions in learning to rank. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 315–323. Curran Associates, Inc., 2009. URL: https://proceedings.neurips.cc/paper/2009/hash/2f55707d4193dc27118a0f19a1985716-Abstract.html.

23 Xuelu Chen, Muhao Chen, Weijia Shi, Yizhou Sun, and Carlo Zaniolo. Embedding uncertain knowledge graphs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3363–3370. AAAI Press, 2019. doi:10.1609/AAAI.V33I01.33013363.

24 Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, Kaili Ma, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/8b21a7ea42cbcd1c29a7a88c444cce45-Abstract-Conference.html.

25 Kewei Cheng, Nesreen K. Ahmed, and Yizhou Sun. Neural compositional rule learning for knowledge graph reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: https://openreview.net/forum?id=F8VKQyDgRVj.

26 Kewei Cheng, Yikai Zhu, Ming Zhang, and Yizhou Sun. Noigan: Noise aware knowledge graph embedding with adversarial learning. In *ICLR 2020 Conference*, 2020. URL: https://api.semanticscholar.org/CorpusID:226951634.

27 Arijit Ghosh Chowdhury, Md Mofijul Islam, Vaibhav Kumar, Faysal Hossain Shezan, Vaibhav Kumar, Vinija Jain, and Aman Chadha. Breaking down the defenses: A comparative survey of attacks on large language models. *CoRR*, abs/2403.04786, 2024. doi:10.48550/arXiv.2403.04786.

28 Joana C Costa, Tiago Roxo, Hugo Proença, and Pedro RM Inácio. How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*, 2024.

29 Gabriela Csurka. Deep visual domain adaptation. In *22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2020, Timisoara, Romania, September 1-4, 2020*, pages 1–8. IEEE, 2020. doi:10.1109/SYNASC51798.2020.00013.

30 Romain Dagnas, Michel Barbeau, Joaquin Garcia-Alfaro, and Reda Yaich. Resilience assessment of multi-layered cyber-physical systems. In *IFIP Networking 2024-IOCRCI*, 2024.

31 Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Mach. Intell. Res.*, 21(6):1011–1061, 2024. doi:10.1007/s11633-024-1510-8.

32 Jeffrey Dalton, Laura Dietz, and James Allan. Entity query feature expansion using knowledge base links. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 365–374. ACM, 2014. doi:10.1145/2600428.2609628.

33 Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. Lifted rule injection for relation embeddings. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1389–1399. The Association for Computational Linguistics, 2016. doi:10.18653/V1/D16-1146.

34 Caglar Demir, Diego Moussallem, Stefan Heindorf, and Axel-Cyrille Ngonga Ngomo. Convolutional hypercomplex embeddings for link prediction. In *Asian Conference on Machine Learning*, pages 656–671. PMLR, 2021. URL: https://proceedings.mlr.press/v157/demir21a.html.

35 Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

36 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi:10.18653/V1/N19-1423.

37 Jianfeng Du, Kunxun Qi, and Yuming Shen. Knowledge graph embedding with logical consistency. In Maosong Sun, Ting Liu, Xiaojie Wang, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018, Changsha, China, October 19-21, 2018, Proceedings*, volume 11221 of *Lecture Notes in Computer Science*, pages 123–135. Springer, 2018. doi:10.1007/978-3-030-01716-3_11.

38 Jianfeng Du, Kunxun Qi, Hai Wan, Bo Peng, Shengbin Lu, and Yuming Shen. Enhancing knowledge graph embedding from a logical perspective. In Zhe Wang, Anni-Yasmin Turhan, Kewen Wang, and Xiaowang Zhang, editors, *Semantic Technology - 7th Joint International Conference, JIST 2017, Gold Coast, QLD, Australia, November 10-12, 2017, Proceedings*, volume 10675 of *Lecture Notes in Computer Science*, pages 232–247. Springer, 2017. doi:10.1007/978-3-319-70682-5_15.

39 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In Shafi Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012. doi:10.1145/2090236.2090255.

40 Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, and Bai Wang. Generalizing graph neural networks on out-of-distribution graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1):322–337, 2024. doi:10.1109/TPAMI.2023.3321097.

41 Uriel Feige, Yishay Mansour, and Robert E. Schapire. Learning and inference in the presence of corrupted inputs. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 637–657. JMLR.org, 2015. URL: http://proceedings.mlr.press/v40/Feige15.html.

42 David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. Building watson: An overview of the deepqa project. *AI Mag.*, 31(3):59–79, 2010. doi:10.1609/AIMAG.V31I3.2303.

43 Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. URL: http://proceedings.mlr.press/v70/finn17a.html.

44 A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 148–155. Morgan Kaufmann Publishers Inc., 1998.

45 Saurabh Garg, Amrith Setlur, Zachary C. Lipton, Sivaraman Balakrishnan, Virginia Smith, and Aditi Raghunathan. Complementary benefits of contrastive learning and self-training under distribution shift. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/26f96550613971371c5d07f37f0e06c0-Abstract-Conference.html.

46 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. URL: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

47 Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL: http://arxiv.org/abs/1412.6572.

48 Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. A survey of adversarial defenses and robustness in nlp. *ACM Comput. Surv.*, 55(14s), July 2023. doi:10.1145/3593042.

49 Martin Grohe and Pascal Schweitzer. The graph isomorphism problem. *Commun. ACM*, 63(11):128–134, October 2020. doi:10.1145/3372123.

50 Shu Guo, Quan Wang, Bin Wang, Lihong Wang, and Li Guo. SSE: semantically smooth embedding for knowledge graphs. *IEEE Trans. Knowl. Data Eng.*, 29(4):884–897, 2017. doi:10.1109/TKDE.2016.2638425.

51 Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. Jointly embedding knowledge graphs and logical rules. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 192–202. The Association for Computational Linguistics, 2016. doi:10.18653/V1/D16-1019.

52 Peng He, Gang Zhou, Hongbo Liu, Yi Xia, and Ling Wang. Hyperplane-based time-aware knowledge graph embedding for temporal knowledge graph completion. *J. Intell. Fuzzy Syst.*, 42(6):5457–5469, 2022. doi:10.3233/JIFS-211950.

53 Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37, 2021. doi:10.1145/3447772.

54 Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In Bernhard Schölkopf, John C. Platt, and Thomas Hofmann, editors, *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 601–608. MIT Press, 2006. URL: https://proceedings.neurips.cc/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html.

55 Nitisha Jain, Trung-Kien Tran, Mohamed H. Gad-Elrab, and Daria Stepanova. Improving knowledge graph embeddings with ontological reasoning. In Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam M. Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani, editors, *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, volume 12922 of *Lecture Notes in Computer Science*, pages 410–426. Springer, 2021. doi:10.1007/978-3-030-88361-4_24.

56 Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021. doi:10.1109/TNNLS.2021.3070843.

57 Sourabh Kapoor, Arnab Sharma, Michael Röder, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. Robustness evaluation of knowledge graph embedding models under non-targeted attacks. In Edward Curry, Maribel Acosta, María Poveda-Villalón, Marieke van Erp, Adegboyega K. Ojo, Katja Hose, Cogan Shimizu, and Pasquale Lisena, editors, *The Semantic Web - 22nd European Semantic Web Conference, ESWC 2025, Portoroz, Slovenia, June 1-5, 2025, Proceedings, Part I*, volume 15718 of *Lecture Notes in Computer Science*, pages 264–281. Springer, 2025. doi:10.1007/978-3-031-94575-5_15.

58 Seonhyeong Kim and Young-Woo Kwon. Construction of disaster knowledge graphs to enhance disaster resilience. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 6721–6723, 2022. doi:10.1109/BigData55660.2022.10021017.

59 Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

60 Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2793–2806. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.249.

61 Christopher Lang, Alexander Braun, and Abhinav Valada. Robust object detection using knowledge graph embeddings. In Björn Andres, Florian Bernard, Daniel Cremers, Simone Frintrop, Bastian Goldlücke, and Ivo Ihrke, editors, *Pattern Recognition - 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27-30, 2022, Proceedings*, volume 13485 of *Lecture Notes in Computer Science*, pages 445–461. Springer, 2022. doi:10.1007/978-3-031-16788-1_27.

62 Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2723–2732. Association for Computational Linguistics, 2019. doi:10.18653/V1/D19-1274.

63 Jintang Li, Jiaying Peng, Liang Chen, Zibin Zheng, Tingting Liang, and Qing Ling. Spectral adversarial training for robust graph neural network. *IEEE Trans. Knowl. Data Eng.*, 35(9):9240–9253, 2023. doi:10.1109/TKDE.2022.3222207.

64 Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *Int. J. Comput. Vis.*, 133(1):31–64, 2025. doi:10.1007/S11263-024-02181-W.

65 Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

66 Qi Liu, Qinghua Zhang, Fan Zhao, and Guoyin Wang. Uncertain knowledge graph embedding: an effective method combining multi-relation and multi-path. *Frontiers Comput. Sci.*, 18(3):183311, 2024. doi:10.1007/S11704-023-2427-Z.

67 Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and

Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018. URL: https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-5_Liu_paper.pdf.

68 Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*, volume 12355 of *Lecture Notes in Computer Science*, pages 182–199. Springer, 2020. doi:10.1007/978-3-030-58607-2_11.

69 László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993.

70 Xinyu Lu, Lifang Wang, Zejun Jiang, Shichang He, and Shizhong Liu. MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Appl. Intell.*, 52(7):7480–7497, 2022. doi:10.1007/S10489-021-02693-9.

71 Jiangtao Ma, Chenyu Zhou, Yanjun Wang, Yifan Guo, Guangwu Hu, Yaqiong Qiao, and Yong Wang. Ptruste: A high-accuracy knowledge graph noise detection method based on path trustworthiness and triple embedding. *Knowl. Based Syst.*, 256:109688, 2022. doi:10.1016/J.KNOSYS.2022.109688.

72 Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. Are negative samples necessary in entity alignment?: An approach with high performance, scalability and robustness. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 1263–1273. ACM, 2021. doi:10.1145/3459637.3482232.

73 Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993, 2019. doi:10.1109/TPAMI.2018.2858821.

74 Awais Muhammad and Sung-Ho Bae. A survey on efficient methods for adversarial robustness. *IEEE Access*, 10:118815–118830, 2022. doi:10.1109/ACCESS.2022.3216291.

75 Mojtaba Nayyeri, Sahar Vahdati, Emanuel Sallinger, Mirza Mohtashim Alam, Hamed Shariat Yazdi, and Jens Lehmann. Pattern-aware and noise-resilient embedding models. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 483–496. Springer, 2021. doi:10.1007/978-3-030-72113-8_32.

76 Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 3104482–3104584, 2011.

77 Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large language models and knowledge graphs: Opportunities and challenges. *TGDK*, 1(1):2:1–2:38, 2023. doi:10.4230/TGDK.1.1.2.

78 Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. Rea: Robust cross-lingual entity alignment between knowledge graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2175–2184, 2020. doi:10.1145/3394486.3403268.

79 Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014. doi:10.3115/V1/D14-1162.

80 Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. Investigating robustness and interpretability of link prediction via adversarial modifications. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019*, 2019. URL: https://openreview.net/forum?id=Hkg7rbcp67.

81 Natasa Przulj. Biological network comparison using graphlet degree distribution. *Bioinform.*, 26(6):853–854, 2010. doi:10.1093/BIOINFORMATICS/BTQ091.

82 Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

83 Maximilian-Peter Radtke, Marco Huber, and Jürgen Bock. Increasing robustness of data-driven fault diagnostics with knowledge graphs. In *Proceedings of the Annual Conferenceof the PHM Society 2023*. PHM Society, 2023.

84 Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. Injecting logical background knowledge into embeddings for relation extraction. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1119–1129. The Association for Computational Linguistics, 2015. doi:10.3115/V1/N15-1118.

85 Sanjit A. Seshia, Ankush Desai, Tommaso Dreossi, Daniel J. Fremont, Shromona Ghosh, Edward Kim, Sumukh Shivakumar, Marcell Vazquez-Chanlatte, and Xiangyu Yue. Formal specification for deep neural networks. In Shuvendu K. Lahiri and Chao Wang, editors, *Automated Technology for Verification and Analysis - 16th International Symposium,*

ATVA 2018, Los Angeles, CA, USA, October 7-10, 2018, Proceedings, volume 11138 of *Lecture Notes in Computer Science*, pages 20–34. Springer, 2018. doi:10.1007/978-3-030-01090-4_2.

86 Yingchun Shan, Chenyang Bu, Xiaojian Liu, Shengwei Ji, and Lei Li. Confidence-aware negative sampling method for noisy knowledge graph embedding. In Xindong Wu, Yew-Soon Ong, Charu C. Aggarwal, and Huanhuan Chen, editors, *2018 IEEE International Conference on Big Knowledge, ICBK 2018, Singapore, November 17-18, 2018*, pages 33–40. IEEE Computer Society, 2018. doi:10.1109/ICBK.2018.00013.

87 Tianyang Shao, Xinyi Li, Xiang Zhao, Hao Xu, and Weidong Xiao. DSKRL: A dissimilarity-support-aware knowledge representation learning framework on noisy knowledge graph. *Neurocomputing*, 461:608–617, 2021. doi:10.1016/J.NEUCOM.2021.02.099.

88 Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

89 Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory, 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007, Proceedings*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer, 2007. doi:10.1007/978-3-540-75225-7_5.

90 Lorenzo Strigini. Fault tolerance and resilience: Meanings, measures and assessment. In Katinka Wolter, Alberto Avritzer, Marco Vieira, and Aad P. A. van Moorsel, editors, *Resilience Assessment and Evaluation of Computing Systems*, pages 3–24. Springer, 2012. doi:10.1007/978-3-642-29032-9_1.

91 Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007. doi:10.1145/1242572.1242667.

92 Jinze Sun, Yongpan Sheng, Ling Zhan, and Lirong He. TKGR-RHETNE: A new temporal knowledge graph reasoning model via jointly modeling relevant historical event and temporal neighborhood event context. In Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, and Chaojie Li, editors, *Neural Information Processing - 30th International Conference, ICONIP 2023, Changsha, China, November 20-23, 2023, Proceedings, Part V*, volume 14451 of *Lecture Notes in Computer Science*, pages 331–343. Springer, 2023. doi:10.1007/978-981-99-8073-4_26.

93 Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 13(11):2326–2340, 2020. URL: http://www.vldb.org/pvldb/vol13/p2326-sun.pdf.

94 Xiaoli Tang, Rui Yuan, Qianyu Li, Tengyun Wang, Haizhi Yang, Yundong Cai, and Hengjie Song. Timespan-aware dynamic knowledge graph embedding by incorporating temporal evolution. *IEEE Access*, 8:6849–6860, 2020. doi:10.1109/ACCESS.2020.2964028.

95 Xing Tang, Ling Chen, Jun Cui, and Baogang Wei. Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. *Inf. Process. Manag.*, 56(3):809–822, 2019. doi:10.1016/J.IPM.2019.01.005.

96 Armin Toroghi and Scott Sanner. Bayesian inference with complex knowledge graph evidence. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 20550–20558. AAAI Press, 2024. doi:10.1609/AAAI.V38I18.30040.

97 Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR, 2016. URL: http://proceedings.mlr.press/v48/trouillon16.html.

98 Guojia Wan, Bo Du, Shirui Pan, and Jia Wu. Adaptive knowledge subgraph ensemble for robust and trustworthy knowledge graph completion. *World Wide Web*, 23(1):471–490, 2020. doi:10.1007/S11280-019-00711-Y.

99 Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. Multi-task feature learning for knowledge graph enhanced recommendation. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2000–2010. ACM, 2019. doi:10.1145/3308558.3313411.

100 Jiaan Wang, Jianfeng Qu, Kexin Wang, Zhixu Li, Wen Hua, Ximing Li, and An Liu. Improving the robustness of knowledge-grounded dialogue via contrastive learning. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19135–19143. AAAI Press, 2024. doi:10.1609/AAAI.V38I17.29881.

101 Jiapu Wang, Boyue Wang, Meikang Qiu, Shirui Pan, Bo Xiong, Heng Liu, Linhao Luo, Tengfei Liu, Yongli Hu, Baocai Yin, and Wen Gao. A survey on temporal knowledge graph completion: Taxonomy, progress, and prospects. *CoRR*, abs/2308.02457, 2023. doi:10.48550/arXiv.2308.02457.

102 Jie Wang, Ghulam Mubashar Hassan, and Naveed Akhtar. A survey of neural trojan attacks and defenses in deep learning. *CoRR*, abs/2202.07183, 2022. arXiv:2202.07183.

103 Meihong Wang, Linling Qiu, and Xiaoli Wang. A survey on knowledge graph embeddings for

link prediction. *Symmetry*, 13(3):485, 2021. doi:10.3390/SYM13030485.

104 Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 349–357, 2018. doi:10.18653/V1/D18-1032.

105 Zikang Wang, Linjing Li, Qiudan Li, and Daniel Zeng. Multimodal data enhanced representation learning for knowledge graphs. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. doi:10.1109/IJCNN.2019.8852079.

106 L. N. Wasserstein. Markov processes over denumerable products of spaces describing large systems of automata. *Problems of Information Transmission*, 5:47–52, 1969.

107 Junkang Wu, Wentao Shi, Xuezhi Cao, Jiawei Chen, Wenqiang Lei, Fuzheng Zhang, Wei Wu, and Xiangnan He. Disenkgat: Knowledge graph embedding with disentangled graph attention network. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2140–2149. ACM, 2021. doi:10.1145/3459637.3482424.

108 Zhaohan Xi, Tianyu Du, Changjiang Li, Ren Pang, Shouling Ji, Xiapu Luo, Xusheng Xiao, Fenglong Ma, and Ting Wang. On the security risks of knowledge graph reasoning. In Joseph A. Calandrino and Carmela Troncoso, editors, *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 3259–3276. USENIX Association, 2023. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/xi.

109 Hanhua Xiao, Yuchen Li, Yanhao Wang, Panagiotis Karras, Kyriakos Mouratidis, and Natalia Rozalia Avlona. How to avoid jumping to conclusions: Measuring the robustness of outstanding facts in knowledge graphs. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 3539–3550. ACM, 2024. doi:10.1145/3637528.3671763.

110 Ruobing Xie, Stefan Heinrich, Zhiyuan Liu, Cornelius Weber, Yuan Yao, Stefan Wermter, and Maosong Sun. Integrating image-based and knowledge-based representation learning. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):169–178, 2019. doi:10.1109/TCDS.2019.2906685.

111 Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. Does william shakespeare REALLY write hamlet? knowledge representation learning with confidence. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4954–4961. AAAI Press, 2018. doi:10.1609/AAAI.V32I1.11924.

112 Huan Xu and Shie Mannor. Robustness and generalization. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 503–515. Omnipress, 2010. URL: http://colt2010.haifa.il.ibm.com/papers/COLT2010proceedings.pdf#page=511.

113 Jiarong Xu, Junru Chen, Siqi You, Zhiqing Xiao, Yang Yang, and Jiangang Lu. Robustness of deep learning models on graphs: A survey. *AI Open*, 2:69–78, 2021. doi:10.1016/j.aiopen.2021.05.002.

114 Jiarong Xu, Junru Chen, Siqi You, Zhiqing Xiao, Yang Yang, and Jiangang Lu. Robustness of deep learning models on graphs: A survey. *AI Open*, 2:69–78, 2021. doi:10.1016/J.AIOPEN.2021.05.002.

115 Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. *CoRR*, abs/2005.12914, 2020. arXiv:2005.12914.

116 Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

117 Yi Yang, Chen Peng, En-Zhi Cao, and Wenxuan Zou. Building resilience in supply chains: A knowledge graph-based risk management framework. *IEEE Transactions on Computational Social Systems*, pages 1–9, 2023. doi:10.1109/TCSS.2023.3334768.

118 Yuhao Yang, Chao Huang, Lianghao Xia, and Chenliang Li. Knowledge graph contrastive learning for recommendation. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1434–1443. ACM, 2022. doi:10.1145/3477495.3532009.

119 Xiaoyu You, Beina Sheng, Daizong Ding, Mi Zhang, Xudong Pan, Min Yang, and Fuli Feng. Mass: Model-agnostic, semantic and stealthy data poisoning attack on knowledge graph embedding. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2000–2010. ACM, 2023. doi:10.1145/3543507.3583203.

120 Mei Yu, Jiujiang Guo, Jian Yu, Tianyi Xu, Mankun Zhao, Hongwei Liu, Xuewei Li, and Ruiguo Yu. TBDRI: block decomposition based on relational interaction for temporal knowledge graph completion. *Appl. Intell.*, 53(5):5072–5084, 2023. doi:10.1007/S10489-022-03601-5.

121 Mohamad Zamini, Hassan Reza, and Minou Rabiei. A review of knowledge graph completion. *Information*, 13(8):396, 2022. doi:10.3390/INFO13080396.

122 Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, and Kui Ren.

Data poisoning attack against knowledge graph embedding. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4853–4859. ijcai.org, 2019. doi:10.24963/ijcai.2019/674.

123 Kexin Zhang, Shuhan Liu, Song Wang, Weili Shi, Chen Chen, Pan Li, Sheng Li, Jundong Li, and Kaize Ding. A survey of deep graph learning under distribution shifts: from graph out-of-distribution generalization to adaptation. *CoRR*, abs/2410.19265, 2024. doi:10.48550/arXiv.2410.19265.

124 Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2731–2741, 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/d961e9f236177d65d21100592edb0769-Abstract.html.

125 Shuai Zhang, Yi Tay, Lina Yao, and Qi Liu. Quaternion knowledge graph embeddings. *Advances in neural information processing systems*, 32, 2019.

126 Yuxiao Zhang, Qingfeng Chen, Xinkun Hao, Haiming Pan, Qian Yu, and Kexin Huang. Defense against adversarial attack on knowledge graph embedding. *Emerging Trends in Cybersecurity Applications*, page 441, 2023.

127 Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. Dynamic graph neural networks under spatio-temporal distribution shift. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/2857242c9e97de339ce642e75b15ff24-Abstract-Conference.html.

128 Zhao Zhang, Fuzhen Zhuang, Hengshu Zhu, Chao Li, Hui Xiong, Qing He, and Yongjun Xu. Towards robust knowledge graph embedding via multi-task reinforcement learning. *IEEE Trans. Knowl. Data Eng.*, 35(4):4321–4334, 2023. doi:10.1109/TKDE.2021.3127951.

129 Tianzhe Zhao, Jiaoyan Chen, Yanchi Ru, Qika Lin, Yuxia Geng, and Jun Liu. Untargeted adversarial attack on knowledge graph embeddings. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1701–1711. ACM, 2024. doi:10.1145/3626772.3657702.

130 Yu Zhao, Huali Feng, and Patrick Gallinari. Embedding learning with triple trustiness on noisy knowledge graph. *Entropy*, 21(11):1083, 2019. doi:10.3390/E21111083.

131 Lei Zheng, Pei Quan, Yong Shi, and Lingfeng Niu. A brief survey of distribution robust graph neural networks. *Procedia Computer Science*, 242:1281–1286, 2024. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024). doi:10.1016/j.procs.2024.08.140.

132 Yongchun Zhu, Fuzhen Zhuang, Xiangliang Zhang, Zhiyuan Qi, Zhi-Ping Shi, Juan Cao, and Qing He. Combat data shift in few-shot learning with knowledge graph. *Frontiers Comput. Sci.*, 17(1):171305, 2023. doi:10.1007/S11704-022-1339-7.

133 Yuqicheng Zhu, Nico Potyka, Mojtaba Nayyeri, Bo Xiong, Yunjie He, Evgeny Kharlamov, and Steffen Staab. Predictive multiplicity of knowledge graph embeddings in link prediction. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 334–354. Association for Computational Linguistics, 2024. doi:10.18653/V1/2024.FINDINGS-EMNLP.19.

134 Yuqicheng Zhu, Nico Potyka, Jiarong Pan, Bo Xiong, Yunjie He, Evgeny Kharlamov, and Steffen Staab. Conformalized answer set prediction for knowledge graph embedding. *CoRR*, abs/2408.08248, 2024. doi:10.48550/arXiv.2408.08248.

Distances Between Formal Concept Analysis Structures

Alexandre Bazin
LIRMM, CNRS, Université de Montpellier, Montpellier, France

Giacomo Kahn
Université Lumière Lyon 2, INSA Lyon, Université Claude Bernard Lyon 1, France
Université Jean Monnet Saint-Etienne, DISP UR4570, Bron, France

Abstract

In this paper, we study the notion of distance between the most important structures of formal concept analysis: formal contexts, concept lattices, and implication bases. We first define three families of Minkowski-like distances between these three structures. We then present experiments showing that the correlations of these distances are low and depend on the distance between formal contexts.

2012 ACM Subject Classification Mathematics of computing → Discrete mathematics

Keywords and phrases Formal Concept Analysis, Implication Base, Concept Lattice, Pattern Mining, Ordinal Data Science

Digital Object Identifier 10.4230/TGDK.3.2.2

Category Research

Related Version *Previous Version*: `https://hal.science/hal-04475242`

Supplementary Material *Software (Source Code)*: `https://github.com/Authary/FCAD`
archived at `swh:1:dir:75951913d5a4222771718415a619dc6b1a97a6ed`
Software (Experiments Code): `https://github.com/Authary/experiments_distances_fca`
archived at `swh:1:dir:00363b3ec2ad2d63a8235a852392699f1ccf6688`

Funding This work was partially supported by the ANR SmartFCA project Grant ANR-21-CE23-0023 of the French National Research Agency.

Acknowledgements The authors thank the members of the SmartFCA project for their advice.

Received 2024-12-09 **Accepted** 2025-05-07 **Published** 2025-10-15

1 Introduction

Formal Concept Analysis (FCA [12]) is a mathematical framework that allows extracting patterns called concepts from data in the form of objects described by attributes, and organises them in an ordered structure called a concept lattice. Concept lattices are then used for exploratory search [15, 14], conceptual navigation [22, 1], and other applications – see [17] for a survey. The framework also handles implications between sets of attributes, that can be summarised by implication bases. In FCA, formal contexts, concept lattices and sets of implications are three representations of – or points of view on – the same entity and all three of them are well known, well studied, and well used in various fields of data mining [19, 18, 7].

We are interested in distances between these FCA structures. Given two data tables on the same objects and attributes, how far apart are the structures that are extracted from them? In this paper, we define three families of distances: one between formal contexts, one between implication bases, and one between concept lattices. For formal contexts, we consider the context as a set of pairs (the incidence relation) and use set-based analogues of Minkowski distances to define the *factual distance*. For concept lattices and implication bases, we consider the structures as representations of, respectively, the derivation operators and the closure operator of the

corresponding context and propose Minkowski-like distances: the *conceptual distance* and the *logical distance*. We show that these distances are metrics and we provide algorithms to compute them. We experimentally study the correlations between those distances on formal contexts that are *closer* or *farther apart* and observe that these correlations depend on the factual distance.

There are multiple expected applications for this work. The most direct one would be the comparison of concept lattices or implication bases, for instance to study the differences in the variability in different software product lines [3]. This contribution would then allow for the study of the trajectory of a given software collection between versions – how much a new version differs from older versions. The notion of trajectory can also be used for iterative processes such as Relational Concept Analysis [21, 2], to quantify how much variation there is between concept lattices in some given steps. In those two contexts, the objects and attributes are basically the same through time or during the process, and the labels of objects carry some significance, which is why our distances consider contexts with the same object and attributes sets. In distance-based machine learning, knowledge is often embedded in numerical vectors. This contribution allows for the direct computation of a distance between knowledge structures. More broadly, these distances could be used to define complexity indicators in triadic or polyadic datasets [24] as the relative distance of each *slice* of n-context to every other. Additionally, this is a contribution to Ordinal Data Science, as defined in its manifesto [23], and it seemed like an interesting question to be answered in itself.

The paper follows a classic structure: in Section 2 we define the necessary notions of FCA and distances, then we introduce our distances between FCA structures and the algorithms to compute them in Section 3. In Section 4 we experiment on the new distances: we study the correlations between them and compare them together and with Domenach's dissimilarity measure [9] on concept lattices.

2 Preliminaries

2.1 Formal Concept Analysis

In the following, we consider only finite sets.

Formal Concept Analysis (FCA) is a mathematical framework based on lattice theory that aims at structuring the information contained in the relation between *objects* and their *attributes* [12]. It is centered around the notion of *formal context*.

▶ **Definition 1** (Formal context). *A* formal context *is a triple* $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ *in which* $\mathcal{O}$ *is a set of* objects*,* $\mathcal{A}$ *is a set of* attributes *and* $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{A}$ *is a binary relation between objects and attributes. We say that* the object o is described by the attribute a *when* $(o, a) \in \mathcal{R}$*.*

Formal contexts can be represented as crosstables.

	a_1	a_2	a_3	a_4	a_5
o_1	×	×			
o_2		×	×	×	
o_3		×		×	×
o_4			×		×
o_5				×	×

Figure 1 A formal context with five objects and five attributes.

A formal context $\mathcal{C}$ gives rise to two *derivation operators*, both usually noted $\cdot'$, defined as:

$$\cdot' : \mathcal{P}(\mathcal{A}) \to \mathcal{P}(\mathcal{O})$$

$$A' = \{o \in \mathcal{O} \mid \forall a \in A, (o, a) \in \mathcal{R}\}$$

$$\cdot' : \mathcal{P}(\mathcal{O}) \to \mathcal{P}(\mathcal{A})$$

$$O' = \{a \in \mathcal{A} \mid \forall o \in O, (o, a) \in \mathcal{R}\}$$

where $\mathcal{P}(X)$ denotes the powerset of X.

For instance, in the Fig. 1 context, $\{a_2, a_4\}' = \{o_2, o_3\}$ and $\{a_1\}'' = \{a_1, a_2\}$. Both operators $\cdot'$ form a Galois connection and their compositions $\cdot''$ are closure operators. Throughout this paper, when in the presence of two different formal contexts $\mathcal{C}_1$ and $\mathcal{C}_2$, we shall use $\cdot'^i$ and $\cdot''^i$ to denote the derivation and closure operators of context $\mathcal{C}_i$.

▶ **Definition 2** (Formal concept). *In a formal context $(\mathcal{O}, \mathcal{A}, \mathcal{R})$, a* formal concept *is a pair (E, I) in which E is a set of objects, I is a set of attributes, and such that $E = I'$ and $I = E'$. As such, $I = I''$ and $E = E''$ are both closed sets. We call E the* extent *and I the* intent *of the concept.*

Visually, concepts correspond to maximal rectangles of crosses in the context's crosstable, up to permutation of rows and columns. In the Fig. 1 context, the pair $(\{o_2, o_3\}, \{a_2, a_4\})$ is a concept while the pair $(\{o_3, o_4\}, \{a_5\})$ is not as $\{a_5\}' = \{o_3, o_4, o_5\}$. Concepts can be ordered by the inclusion relation on their extents, i.e. $(E_1, I_1) \leq (E_2, I_2) \Leftrightarrow E_1 \subseteq E_2$. As per the basic theorem of formal concept analysis [12], the set of all concepts of a context $\mathcal{C}$ ordered in such a way forms a complete lattice called the *concept lattice of* $\mathcal{C}$. Additionally, all complete lattices are isomorphic to the concept lattice of some context.

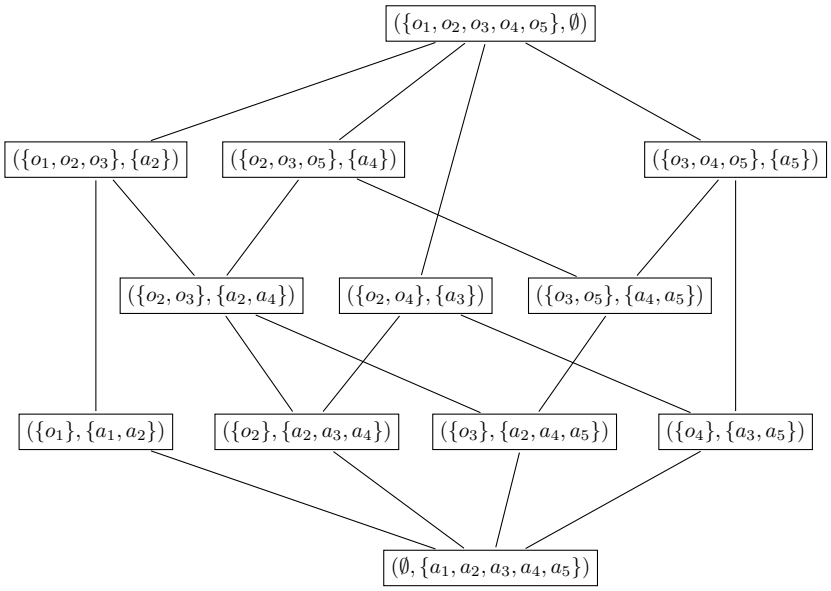


Figure 2 Concept lattice of the formal context depicted in Fig. 1.

▶ **Definition 3** (Implications). *In a formal context $(\mathcal{O}, \mathcal{A}, \mathcal{R})$, an* implication *is a pair of attribute sets (X, Y), usually noted $X \to Y$. An implication $X \to Y$* holds *in the context when $X' \subseteq Y'$ or, equivalently, $Y \subseteq X''$. In other words, the implication holds when all the objects described by X are also described by Y.*

In the Fig. 1 context, the implications $\{a_1\} \to \{a_1, a_2\}$ and $\{a_3, a_4\} \to \{a_2\}$ hold while the implication $\{a_3\} \to \{a_5\}$ does not. For simplicity's sake, we thereafter say "$X \to Y$" instead of "$X \to Y$ holds". Some implications can be inferred from others through Armstrong's axioms:

- if $Y \subseteq X$, then $X \to Y$ (*Reflexivity*)
- if $X \to Y$, then $X \cup Z \to Y \cup Z$ for all attribute sets Z (*Augmentation*)
- if $X \to Y$ and $Y \to Z$, then $X \to Z$. (*Transitivity*)

▶ **Definition 4** (Implication base). *An* implication base *of a formal context is an implication set $\mathcal{I}$ such that the set of implications that can be inferred from $\mathcal{I}$ through Armstrong's axioms is the set of all implications that hold in the context.*

Several implication bases with interesting properties exist in the literature [5, 4]. In this paper, we are interested in only one.

▶ **Definition 5** (Proper Premises). *Let $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ be a formal context and a an attribute. A* proper premise of a *is an inclusion-minimal, non-closed attribute set X such that $X \to \{a\}$, i.e. there is no $Y \subset X$ such that $Y \to \{a\}$.*

In the Fig. 1 example, the set $\{a_2, a_3\}$ is a proper premise of the attribute a_4 as no proper subset of $\{a_2, a_3\}$ implies $\{a_4\}$. The set of all implications $X \to \{a\}$ where a is an attribute and X is one of its proper premises forms an implication base.

▶ **Definition 6** (Logical closure). *Let $\mathcal{I}$ be an implication base. The* logical closure *of an attribute set X by $\mathcal{I}$, denoted $X^{\mathcal{I}}$, is defined as the largest $Y \supseteq X$ such that $X \to Y$ can be inferred from $\mathcal{I}$.*

For instance, the logical closure of the attribute set $\{a_1, a_3\}$ by the implication base $\mathcal{I} = \{\{a_1\} \to \{a_2\}, \{a_2, a_3\} \to \{a_4\}\}$ is $\{a_1, a_3\}^{\mathcal{I}} = \{a_1, a_2, a_3, a_4\}$. The logical closure, as its name indicates, is a closure operator. If $\mathcal{C}$ is a formal context and $\mathcal{I}$ an implication base of $\mathcal{C}$, then $.^{\mathcal{I}} = .''$.

2.2 Metrics

A *metric* on a set S is a function of distance between the elements of S satisfying the following axioms:

- $f(x, x) = 0$
- $f(x, y) > 0$ when $x \neq y$, (positivity)
- $f(x, y) = f(y, x)$, (symmetry)
- $f(x, z) \leq f(x, y) + f(y, z)$. (triangular inequality)

In this paper, we make use of two families of metrics between vectors and sets so as to build our own metrics between FCA structures. The first is the well-known family of Minkowski distances between vectors $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ defined as

$$D_p(X, Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}.$$

The second is the family of normalised set-based analogues of Minkowski distances [13] defined, for two sets X and Y, as

$$d_{2,q}(X, Y) = \frac{\sqrt[q]{(|X| - |X \cap Y|)^q + (|Y| - |X \cap Y|)^q}}{|X \cap Y| + \sqrt[q]{(|X| - |X \cap Y|)^q + (|Y| - |X \cap Y|)^q}}.$$

In this paper, we chose to use the Minkowski distance for sets. Other, more usual distances (e.g. Hamming distance, or another edit distance) might also be interesting, and in most cases they can be plugged into the calculations in Section 3 to describe new families of distances between FCA structures. In [8], some distances are described directly for binary relations and for (semi-)lattices. However, formal contexts are special cases of binary relation (bipartite graphs) and concept lattices are more than lattices as they take two dimensions into account – objects and attributes. Therefore, we did not make use of those distances and instead chose to propose new ones.

3 Distances Between FCA Structures

3.1 Aim

We aim at proposing distances between FCA structures. This is not a brand new endeavor. Distances between formal contexts can be obtained by considering contexts as being any more widely known structures, such as bipartite graphs or hypergraphs, and using existing distances for these structures. Similarity measures between concept lattices have already been studied [9]. However, these are not sufficient. What we want is a set of three distances that can be used to compare two entities in their three different forms (context, lattice and implication base) and the knowledge of how these three distances relate to each others. In this paper, we suppose that all pairs of structures we compare use the same objects and attributes.

In this section, we define families of distances for each of the usual structures of FCA, and show that they are metrics. The three families are based on the normalised set-based analogues of Minkowski distances $d_{2,p}$ [13]. In Section 4, we provide experimental results on the interaction of those distances.

3.2 Distance Between Contexts

As we only consider contexts on the same sets of objects and attributes, the distance between the contexts depends only on their incidence relations. Hence, we define our distances between formal contexts as a distance between binary relations seen as sets of pairs.

▶ **Definition 7.** *Let* $\mathcal{C}_1 = (\mathcal{O}, \mathcal{A}, \mathcal{R}_1)$ *and* $\mathcal{C}_2 = (\mathcal{O}, \mathcal{A}, \mathcal{R}_2)$ *be two formal contexts. The* factual distance (FD) *between* $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *is defined as*

$$FD_p(\mathcal{C}_1, \mathcal{C}_2) = d_{2,p}(\mathcal{R}_1, \mathcal{R}_2).$$

The two formal contexts depicted in Fig. 3 have a factual distance of ≈ 0.13.

	a_1	a_2	a_3	a_4
o_1	×	×	×	×
o_2	×	×	×	
o_3	×	×		
o_4	×			

	a_1	a_2	a_3	a_4
o_1	×	×	×	×
o_2	×	×		×
o_3	×	×		
o_4	×			

Figure 3 Two chain contexts. The two contexts have a factual distance of ≈ 0.13 with $p = 2$.

As $d_{2,p}$ is a metric, the factual distance is a metric.

3.3 Distance Between Concept Lattices

We consider concept lattices as pairs of functions that map sets of objects to the set of attributes they have in common and sets of attributes to the set of objects they all describe, i.e. we see concept lattices as representations of the derivation operators $\cdot'$. If (E, I) is a concept, then all subsets of E that are not subsets of lower neighbours in the lattice are mapped to I and reciprocally. This is notationally easier to express in terms of the derivation operators associated with the formal context of the lattice: object sets O are mapped to O'. As such, we define our distance between concept lattices as a distance between the derivation operators. For this reason, our distance makes use of the distances between the derivations of every element of the powerset of objects/attributes in both contexts. This has the added benefit of facilitating the comparison of concept lattices with different extents/intents.

▶ **Definition 8.** *Let $\mathcal{L}_1, \mathcal{L}_2$ be the two concept lattices of two contexts $\mathcal{C}_1$ and $\mathcal{C}_2$ with the same sets of objects $\mathcal{O}$ and attributes $\mathcal{A}$. We define the* lattice object distance *as*

$$LOD_{p,q}(\mathcal{L}_1, \mathcal{L}_2) = \frac{\sqrt[p]{\sum_{o \in \mathcal{O}} d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2))^p}}{\sqrt[p]{|\mathcal{O}|}}$$

and the lattice attribute distance *as*

$$LAD_{p,q}(\mathcal{L}_1, \mathcal{L}_2) = \frac{\sqrt[p]{\sum_{a \in \mathcal{A}} d_{2,q}(\mathcal{P}(\{a\}'^1), \mathcal{P}(\{a\}'^2))^p}}{\sqrt[p]{|\mathcal{A}|}}.$$

The conceptual distance *(CD) between $\mathcal{L}_1$ and $\mathcal{L}_2$ is then defined as*

$$CD_{p,q}(\mathcal{L}_1, \mathcal{L}_2) = min(LOD_{p,q}(\mathcal{L}_1, \mathcal{L}_2), LAD_{p,q}(\mathcal{L}_1, \mathcal{L}_2)).$$

In this definition, we chose to use the minimum between the lattice object distance and the lattice attribute distance. One could use the maximum between those two quantities to obtain a slightly different distance.

Figure 4 depicts the two chain concept lattices of the two contexts in Fig. 3. Even though they are isomorphic, their conceptual distance is ≈ 0.33 with $p = 2$ and $q = 1$.

In the following example, as well as in the experiments section (Section 4), we chose to fix $p = 2$ and $q = 1$ in our calculations. Fixing $q = 1$ creates an analogue to the Manhattan distance. Then, a generalised mean is computed over the set of attributes (resp. objects). With $p = 2$, we are using the *root mean square* deviation.

The conceptual distance takes its values in $[0, 1]$ and is a metric, satisfying the following axioms:

1. $CD(x, x) = 0$
2. $CD(x, y) > 0$ when $x \neq y$, (positivity)
3. $CD(x, y) = CD(y, x)$, (symmetry)
4. $CD(x, z) \leq CD(x, y) + CD(y, z)$. (triangular inequality)

These follow directly from the fact that $d_{2,p}$ is a metric:

1. because $\mathcal{L}_1 = \mathcal{L}_2 \Rightarrow \mathcal{P}(\{o\}'^1) = \mathcal{P}(\{o\}'^2)$ and $d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^1)) = 0$
2. because $d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2)) > 0$
3. because $d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2)) = d_{2,q}(\mathcal{P}(\{o\}'^2), \mathcal{P}(\{o\}'^1))$
4. because

$$d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^3)) \leq d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2)) + d_{2,q}(\mathcal{P}(\{o\}'^2), \mathcal{P}(\{o\}'^3))$$

Computing the conceptual distance is quite easy: for each object o, find the concepts with the smallest extents that contain o in both lattices. Their intents are $\{o\}'^1$ and $\{o\}'^2$ respectively. Then, computing $d_{2,q}(\mathcal{P}(\{o\}'^1), \mathcal{P}(\{o\}'^2))$ is straightforward. Algorithm 1 follows this principle. Finding the concept with the smallest extent that contains an object in the lattice $\mathcal{L}_i$ can be done in $O(|\mathcal{L}_i|)$ so the time complexity of Algorithm 1 is in $O((|\mathcal{O}| + |\mathcal{A}|) \times max(|\mathcal{L}_1|, |\mathcal{L}_2|))$.

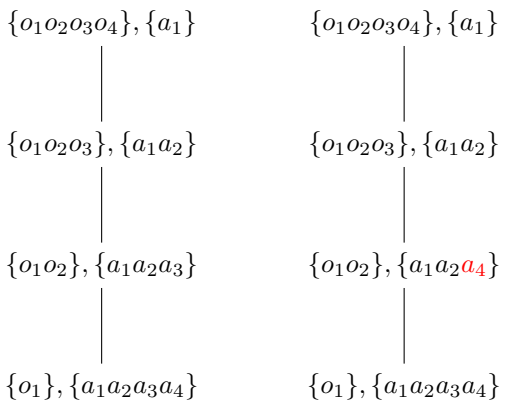


Figure 4 The two concept lattices of the Fig. 3 contexts. These have a conceptual distance of $CD_{2,1} \approx 0.33$ with $p = 2$ and $q = 1$. A small difference in the intents leads to a non-zero distance, even on isomorphic lattices with the same extents.

Algorithm 1 $CD_{p,q}$.

Input: Two concept lattices $\mathcal{L}_1$ and $\mathcal{L}_2$ with the same sets of objects $\mathcal{O}$ and attributes $\mathcal{A}$, p and q
Output: $CD_{p,q}(\mathcal{L}_1, \mathcal{L}_2)$

$LOD = 0$
foreach *object* $o \in \mathcal{O}$ **do**
  $LOD = LOD + (\sqrt[q]{(2^{|\{o\}'^1|} - 2^{|\{o\}'^1 \cap \{o\}'^2|})^q + (2^{|\{o\}'^2|} - 2^{|\{o\}'^1 \cap \{o\}'^2|})^q})^p$
$LOD = \sqrt[p]{LOD} / \sqrt[p]{|\mathcal{O}|}$
$LAD = 0$
foreach *attribute* $a \in \mathcal{A}$ **do**
  $LAD = LAD + (\sqrt[q]{(2^{|\{a\}'^1|} - 2^{|\{a\}'^1 \cap \{a\}'^2|})^q + (2^{|\{a\}'^2|} - 2^{|\{a\}'^1 \cap \{a\}'^2|})^q})^p$
$LAD = \sqrt[p]{LAD} / \sqrt[p]{|\mathcal{A}|}$
return $min(LOD, LAD)$

3.4 Distance between Implication Bases

For our distance between implication bases, we consider implication bases as functions mapping attribute sets X to attribute sets $Y = \{y \mid X \to \{y\}\}$, i.e. we see implication bases as representations of the closure operator $\cdot''$ on attributes. Note that, from Armstrong's axioms, we can infer that

$$X \to Y \Leftrightarrow \forall y \in Y, X \to \{y\}.$$

▶ **Definition 9.** *Let $\mathcal{I}_1, \mathcal{I}_2$ be two implication bases on the same attribute set $\mathcal{A}$. For an attribute $a \in \mathcal{A}$ and an implication base $\mathcal{I}$, we denote by $\mathcal{I}^a = \{X \mid a \in X^{\mathcal{I}}\}$ the set of attributes sets that imply a. The* logical distance *(LD) between $\mathcal{I}_1$ and $\mathcal{I}_2$ is then defined as*

$$LD_{p,q}(\mathcal{I}_1, \mathcal{I}_2) = \frac{\sqrt[p]{\sum_{a \in \mathcal{A}} d_{2,q}(\mathcal{I}_1^a, \mathcal{I}_2^a)^p}}{\sqrt[p]{|\mathcal{A}|}}.$$

Fig. 5 depicts the two proper premises implication bases of the contexts in Fig 3. These two implication bases have a logical distance of ≈ 0.23. Indeed, the attribute a_3 is implied by all supersets of $\{a_4\}$ only in the first context and the attribute a_4 is implied by all supersets of $\{a_3\}$ only in the second context.

$\{a_4\} \to \{a_2, a_3\}$ $\quad$ $\{a_4\} \to \{a_2\}$
$\{a_3\} \to \{a_2\}$ $\quad$ $\{a_3\} \to \{a_2, a_4\}$
$\emptyset \to \{a_1\}$ $\quad$ $\emptyset \to \{a_1\}$

Figure 5 The two proper premises bases of the Fig. 3 contexts. The logical distance, with parameters $p = 2$ and $q = 1$, between these two bases is ≈ 0.23.

The logical distance takes its values in $[0, 1]$ and is a metric, satisfying the following axioms:

1. $LD(x, x) = 0$
2. $LD(x, y) > 0$ when $x \neq y$, (positivity)
3. $LD(x, y) = LD(y, x)$, (symmetry)
4. $LD(x, z) \leq LD(x, y) + LD(y, z)$. (triangular inequality)

Just as those for the conceptual distance, these axioms follow from the fact that $d_{2,q}$ is a metric.

To compute the logical distance, one requires the knowledge of all the attribute sets X that imply a given attribute a. This is not explicitly contained in implication bases and retrieving it is the computationally most expensive part of computing the distance. We propose Algorithm 3 to compute the logical distance. We assume that the implication bases are proper premises bases. If this is not the case, other bases can be converted to proper premises bases in output-polynomial time [16].

The algorithm treats each attribute a separately. The first step is to compute the cardinalities of $\mathcal{I}_1^a \cap \mathcal{I}_2^a$, $\mathcal{I}_1^a$ and $\mathcal{I}_2^a$. To do so, we start with computing the attribute sets P that are minimal such that $P \to \{a\}$ in both implication bases ($commonPremises$). The cardinality of $\mathcal{I}_1^a \cap \mathcal{I}_2^a$ is then the number of attributes sets that contain one of the elements of $commonPremises$. To obtain it, we compute the union closure U_c of the set $commonPremises$, i.e. the minimal sets of attributes sets such that $X, Y \in U_c \Rightarrow X \cup Y \in U_c$. The set U_c ordered by set-inclusion forms a lattice. We use Algorithm 2 to associate to each element x of U_c the number of attribute sets that contain x but not its supersets in U_c, *i.e.* the size of the equivalence classes in the union-closed lattice. Algorithm 3 then sums those numbers (sum_c) to obtain the numbers of attribute sets containing one of the corresponding premises. The same approach is applied to compute the cardinalities of $\mathcal{I}_1^a$ and $\mathcal{I}_2^a$. As the size of the union closure is bounded by $2^{|\mathcal{A}|}$ (when all singletons are premises), the worst case complexity of Algorithm 3 is in $O(|\mathcal{A}| \times 2^{|\mathcal{A}|})$.

4 Experiments

In all these experiments, we used parameters $p = 2$ and $q = 1$ for all distances. A Python module[1] containing the three distances, as well as the script for the experiments themselves[2], are publicly available.

4.1 Correlation Between distances

The first question that may come to mind is "how do these distances compare to each others?". Let us consider the Fig. 6 example representing three contexts $\mathcal{C}_{B_3}$, $\mathcal{C}_{M3}$ and $\mathcal{C}_{N5}$ corresponding respectively to Boolean ($\mathcal{B}_3$), M_3 and N_5 concept lattices, and their associated proper premises implication bases $\mathcal{I}_{B_3}$, $\mathcal{I}_{M_3}$ and $\mathcal{I}_{N_5}$. We compute the factual, conceptual and logical distances between $\mathcal{C}_{\mathcal{B}_3}$ and the other two and obtain the following results:

[1] `https://github.com/Authary/FCAD`
[2] `https://github.com/Authary/experiments_distances_fca`

Algorithm 2 *sizeEQ.*

Input: A set U of premises
Output: $sizeEQ(U)$
Build a dictionary D mapping each premise P in U to the set of premises $P_2 \supset P$
$sum \leftarrow 0$
$over \leftarrow false$
while $over = false$ **do**
  $over \leftarrow true$
  foreach *premise* P *in* U **do**
    if *all* $P_2 \in D(P)$ *have been tagged* **then**
      $|P^{\equiv}| \leftarrow 2^{|\mathcal{A}|-|P|-1} - \sum_{P_2 \in D(P)} |P_2^{\equiv}|$
      Tag P
      $sum \leftarrow sum + |P^{\equiv}|$
      $over \leftarrow false$
return sum

Algorithm 3 *LD.*

Input: Two implication bases $\mathcal{I}_1$ and $\mathcal{I}_2$, p, q
Output: $LD_{p,q}(\mathcal{I}_1, \mathcal{I}_2)$
$Result \leftarrow 0$
foreach *attribute* a **do**
  $U_1 = unionClosure(\{P \mid P \to \{a\} \in \mathcal{I}_1\})$
  $U_2 = unionClosure(\{P \mid P \to \{a\} \in \mathcal{I}_2\})$
  $commonPremises = min(\{P_1 \cup P_2 \mid P_1 \to \{a\} \in \mathcal{I}_1, P_2 \to \{a\} \in \mathcal{I}_2\})$
  $U_c = unionClosure(commonPremises)$
  $sum_c = sizeEQ(U_c)$
  $sum_1 = sizeEQ(U_1)$
  $sum_2 = sizeEQ(U_2)$
  $Result = Result + (\sqrt[q]{(sum_1 - sum_c)^q + (sum_2 - sum_c)^q})^p$
return $\sqrt[p]{Result} / \sqrt[p]{|\mathcal{A}|}$

$$FD_{2,1}(\mathcal{C}_{\mathcal{B}_3}, \mathcal{C}_{M_3}) = 1 > FD_{2,1}(\mathcal{C}_{\mathcal{B}_3}, \mathcal{C}_{N_5}) = 0.51$$

$$CD_{2,1}(\mathcal{B}_3, M_2) = 0.75 > CD_{2,1}(\mathcal{B}_3, N_5) = 0.54$$

$$LD_{2,1}(\mathcal{I}_{\mathcal{B}_3}, \mathcal{I}_{M_3}) = 0.20 < LD_{2,1}(\mathcal{I}_{\mathcal{B}_3}, \mathcal{I}_{N_5}) = 0.22$$

We observe that $\mathcal{C}_{N_5}$ is factually and conceptually closer to $\mathcal{C}_{\mathcal{B}_3}$ while $\mathcal{C}_{M_3}$ is logically closer. The three distances therefore do not always agree. Are they at least correlated? In order to answer this question, we generated structures in different ways. As contexts are the easiest structure to manipulate, we explored the following approaches:

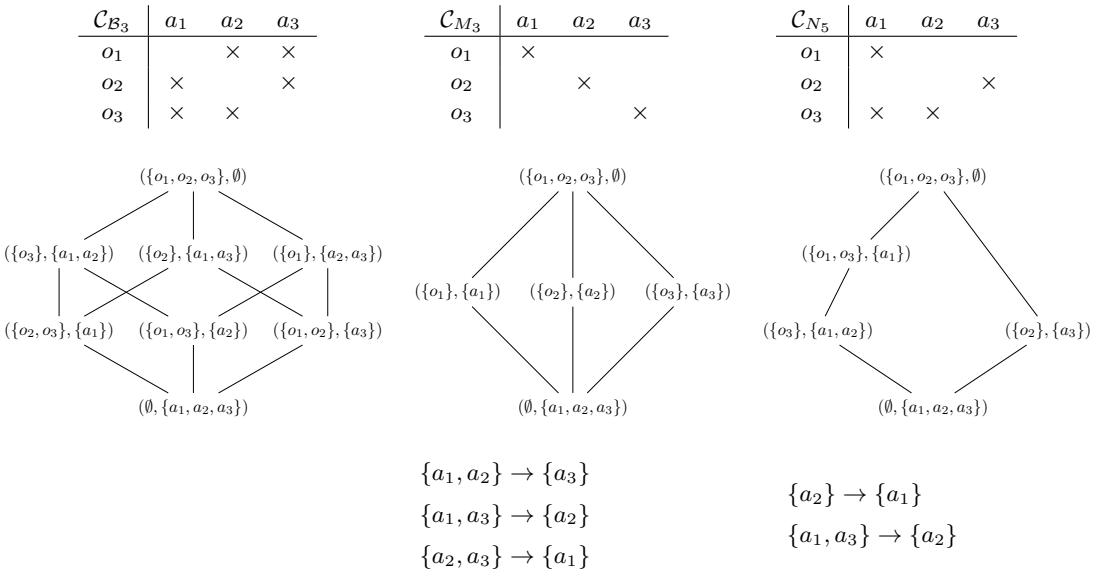


Figure 6 Three formal contexts $\mathcal{C}_{\mathcal{B}_3}$, $\mathcal{C}_{M_3}$ and $\mathcal{C}_{N_5}$ corresponding respectively to the Boolean, $N5$ and $M3$ concept lattices, and their associated implication bases (proper premises).

- starting from a full context and iteratively removing crosses either row by row or randomly chosen
- randomly generating contexts
  - by having each cross with a probability p
  - by randomly flipping each cross of a reference context with a probability p
- generating *pseudo-real* contexts by sampling real data.

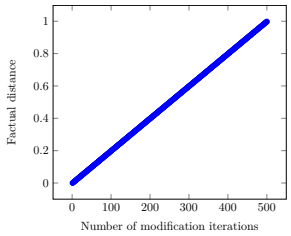
In each case, we computed three correlation coefficients, Pearson, Spearman and Kendall's τ.

4.2 Iterative Emptying of a Full Context

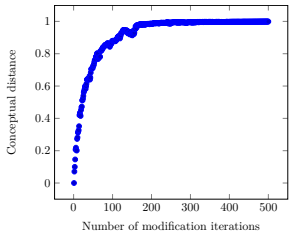
In a first series of experiment, we explored whether we could control the variation of the distances by purposefully modifying a context.

4.2.1 Random Removal of Crosses

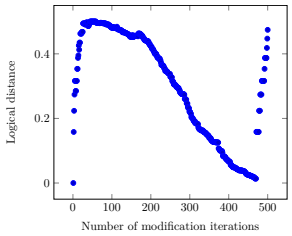
In a first experiment, we started with a full formal context $(\mathcal{O}, \mathcal{A}, \mathcal{R})$ with $|\mathcal{O}| = 50$, $|\mathcal{A}| = 10$ and $\mathcal{R} = \mathcal{O} \times \mathcal{A}$. We iteratively removed random crosses one by one and, at each step, computed the three distances between the current context and the initial one. Fig. 7 presents the progression of the three distances. We observe that while the factual distance increases linearly (which was expected), the other two distances behave very differently. In particular, while the factual and conceptual distances are increasing, the logical distance varies cyclically. This is because in both the full and the empty contexts each attribute implies all the others. Fig. 7 also depicts three diagrams illustrating respectively the relation between the factual (x-axis) and conceptual (y-axis) distances, the relation between the factual (x-axis) and the logical (y-axis) distances and the relation between the conceptual (x-axis) and the logical (y-axis) relations. Additionally, the figure also presents the values of the three correlation coefficients, Pearson, Spearman and Kendall's τ. In this experiment, the values of the three distances are fairly correlated.



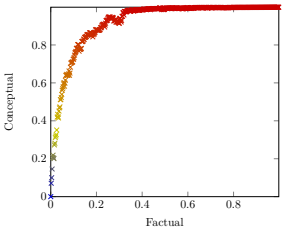
(a) Progression of the factual distance.



(b) Progression of the conceptual distance.

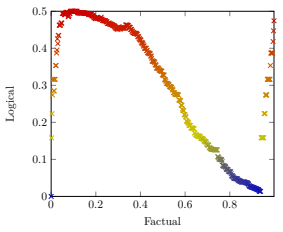


(c) Progression of the logical distance.



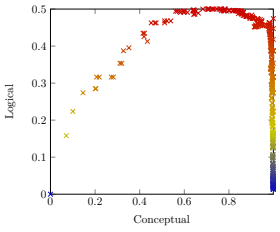
Pearson	Spearman	Kendall
0.70	0.97	0.92

(d) Factual (x-axis) and conceptual (y-axis) distances to the full context of contexts obtained by iteratively removing random crosses.



Pearson	Spearman	Kendall
−0.88	−0.88	−0.81

(e) Factual (x-axis) and logical (y-axis) distances to the full context of contexts obtained by iteratively removing random crosses.



Pearson	Spearman	Kendall
−0.43	−0.88	−0.75

(f) Conceptual (x-axis) and logical (y-axis) distances to the full context of contexts obtained by iteratively removing random crosses.

Figure 7 Iterative removal of crosses from a full context: random crosses.

4.2.2 Iterative Removal of Specific Crosses

In the second set of experiments, we tried to control the variation of the conceptual and logical distances by removing specific crosses in the starting context instead of random ones. We removed crosses row by row so that, *i.e.*, at every step, at most one row is neither full nor empty. Fig. 8 depicts the progression of the three distances and the relations between pairs of distances. By modifying the contexts in this way, the factual distance to the original full context still increases linearly but the conceptual distance increases more slowly. More interestingly, the values of the logical distance repeat every 10 iterations, *i.e.* the number of attributes:

0.1581 0.1118 0.0684 0.0395 0.0220 0.0121 0.0065 0.0034 0.0018 0.0009

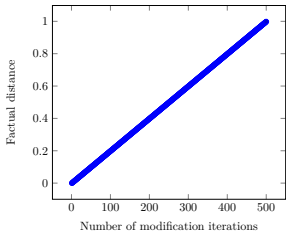
This phenomenon is due to the fact that empty rows do not impact the implications, so every 10 iterations results in a context equivalent to the full context with a single row missing some crosses. In this experiment, only the factual and conceptual distances are correlated.

4.3 Random Contexts

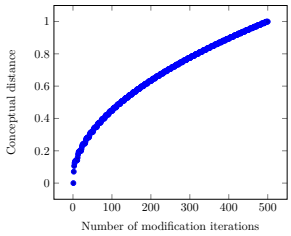
In a second series of experiments, we explored how random generations and modifications of contexts impact the distances, with a focus on how correlated the distances are.

4.3.1 Randomly Generated Contexts

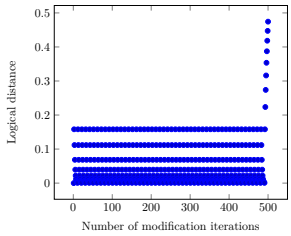
We randomly generated 1500 pairs (A, B) of formal contexts with 50 objects and 10 attributes, with a pair (o, a) having a probability 0.3 of being in the incidence relation. We then computed the distances between the contexts (resp. their associated lattices and implication bases) in each



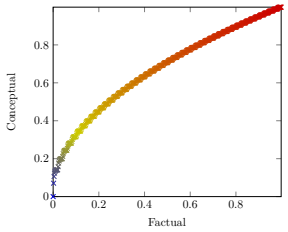
(a) Progression of the factual distance.



(b) Progression of the conceptual distance.

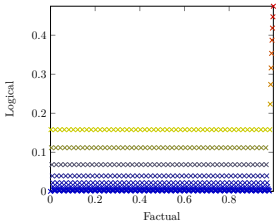


(c) Progression of the logical distance.



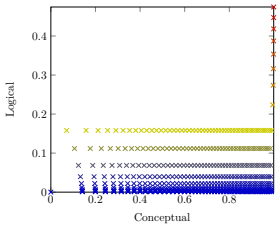
Pearson	Spearman	Kendall
0.98	1	0.99

(d) Factual (x-axis) and conceptual (y-axis) distances to the full context of contexts obtained by iteratively removing minimally aligned crosses.



Pearson	Spearman	Kendall
0.12	0.04	0.02

(e) Factual (x-axis) and logical (y-axis) distances to the full context of contexts obtained by iteratively removing minimally aligned crosses.



Pearson	Spearman	Kendall
0.10	0.04	0.02

(f) Conceptual (x-axis) and logical (y-axis) distances to the full context of contexts obtained by iteratively removing minimally aligned crosses.

Figure 8 Iterative removal of crosses from a full context: minimally aligned crosses.

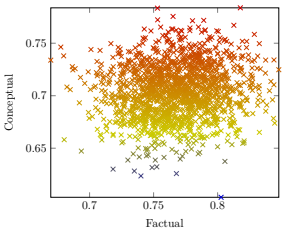
pair. Fig. 9 depicts three diagrams illustrating the relation between the factual (x-axis) and logical (y-axis) distances, the relation between the factual (x-axis) and the conceptual (y-axis) distances and the relation between the conceptual (x-axis) and the logical (y-axis) relations.

We observe that the three distances appear to be pairwise independent when the contexts are randomly generated in such a way. Fig. 9 also depicts the values of the three correlation coefficients, Pearson, Spearman and Kendall's τ. Their values confirm the independence, with the factual and conceptual distances being very slightly less independent. Note that Pearson measures linear correlation, Spearman assesses monotonic relationships and Kendall's τ measures rank correlation.

Interestingly, all factual distances are between 0.67 and 0.85, suggesting that random generation produces contexts that are far apart.

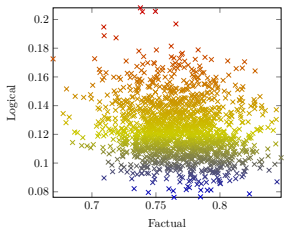
4.3.2 Randomly Modified Contexts

In a second batch of experiments, we generated 1500 other pairs of contexts such that A is a randomly generated context and B is obtained by randomly modifying A. All contexts contain 50 objects and 10 attributes. The contexts A were generated with a probability 0.3 for each cross. The modified contexts were obtained through the following algorithm: for each $(object, attribute)$ pair, with a probability 0.05, remove the pair from the incidence relation if it belongs to it or add it if it does not. We then computed the distances between the contexts (resp. their associate lattices and implication bases) in each pair. Fig. 10 depicts three diagrams illustrating the relation between the factual (x-axis) and logical (y-axis) distances, the relation between the factual (x-axis) and the conceptual (y-axis) distances and the relation between the conceptual (x-axis) and the logical (y-axis) relations. Fig. 10 also depicts the values of the three correlation coefficients, Pearson, Spearman and Kendall's τ.



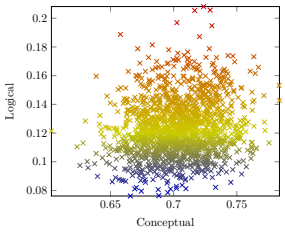
Pearson	Spearman	Kendall
0.09	0.10	0.06

(a) Factual (x-axis) and conceptual (y-axis) distances of random contexts.



Pearson	Spearman	Kendall
−0.16	−0.15	−0.10

(b) Factual (x-axis) and logical (y-axis) distances of random contexts.

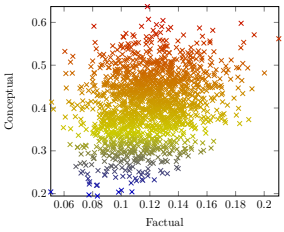


Pearson	Spearman	Kendall
0.14	0.13	0.09

(c) Conceptual (x-axis) and logical (y-axis) distances of random contexts.

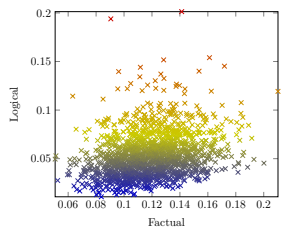
Figure 9 Randomly generated contexts : correlation between the distance measures.

Visually, we observe some slight positive correlation between the factual and conceptual distances and between the factual and logical distances. This is in opposition to the previous experiment with randomly generated contexts. All factual distances are below 0.2, suggesting that our modification algorithms successfully produces contexts that are close together. This result, together with the previous one on randomly generated contexts, hints at the correlations between the factual distance and the others being stronger for very close contexts.



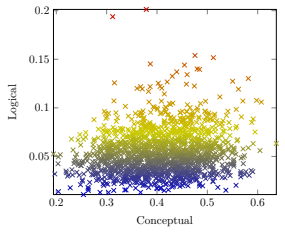
Pearson	Spearman	Kendall
0.25	0.24	0.16

(a) Factual (x-axis) and conceptual (y-axis) distances to a reference context of contexts obtained by randomly modifying the reference.



Pearson	Spearman	Kendall
0.26	0.28	0.19

(b) Factual (x-axis) and logical (y-axis) distances to a reference context of contexts obtained by randomly modifying the reference.



Pearson	Spearman	Kendall
0.14	0.13	0.09

(c) Conceptual (x-axis) and logical (y-axis) distances to a reference context of contexts obtained by randomly modifying the reference.

Figure 10 Randomly modified contexts: correlation between the distance measures (factual,logical), (factual,conceptual) and (conceptual,logical).

4.3.3 Variation of Correlation Relative to the Factual Distance

In the previous experiments, the distances seemed to be more correlated for low factual distances. This hinted at differences in correlations depending on factual distances. Let us check whether this is really the case. For this experiment, we generated pairs (A, B) of contexts such that A is a 50×10 randomly generated context and B is obtained by randomly flipping the truth value of each $(object, attribute)$ pair in A with a probability p. We made p vary from 0.025 to 0.5 with 0.025 increments. For each value of p, we generated 1000 pairs of contexts and computed the three distances between A and B. We then computed the three correlation coefficients for each pair of distances. Fig. 11 presents these correlation values (y-axes) for the different values of p (x-axes).

As a higher p results in factually more distant contexts, we observe that the correlations between the factual distance and the other two decrease when p increases. This confirms our previous observation that these distances are more correlated for low factual distances.

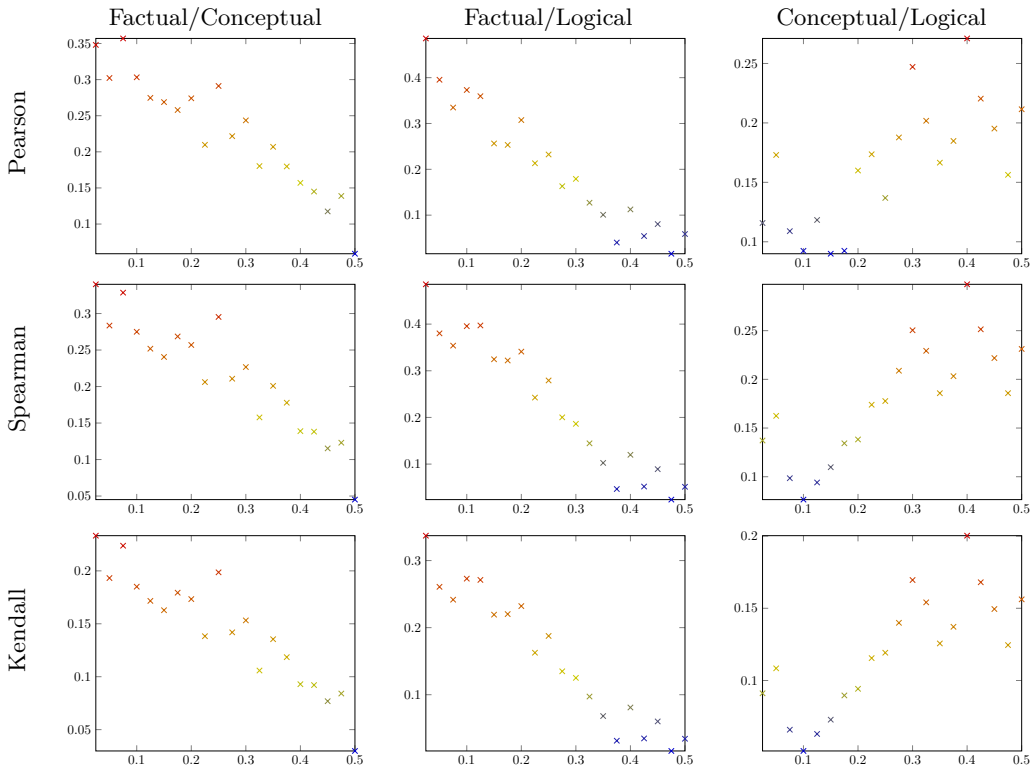


Figure 11 Correlation of the three distance measures for different values of probabilities used in the modification of contexts. Higher probabilities means higher factual distances.

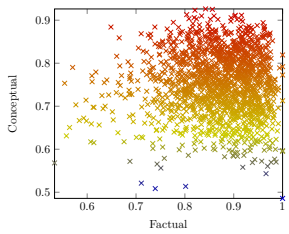
4.4 Pseudo-real data

In this experiment, we sampled real datasets to create pseudo-real contexts. We used the Mushroom dataset [20] and the data from *A statistical analysis of the work of Bob Ross*[3]. We generate pseudo-real data to minimise the risk of interference from the generation method, since generating random formal context is not a trivial task, as shown in [6]. Our sample strategy is the following: we uniformly sample a number n_o of objects along the dimension of objects and a number n_a of attributes along the dimension of attributes. Then, we keep the portion of the incidence relation that correspond to those objects and attributes. Finally, we rename the objects and attributes from 1 to, respectively, n_o and n_a so that two sampled contexts may have the same object and attribute sets. For this set of experiments, our goal is, again, to study the correlations between pairs of distances.

We sampled the Mushroom dataset (8124 objects and 119 attributes) into 1500 smaller contexts (56 objects and 11 attributes) and computed the three distances between pairs of such sampled contexts. The average density of the sampled contexts is 0.19.

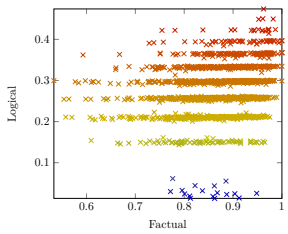
[3] `https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/`

The results are shown in Fig. 12. They show no strong correlation between the distances, except between the factual distance and the logical distance. In this experiment, the logical distance only takes values around a few quantities, unlike in the case of the removal of crosses where the logical distance had exactly the same 10 values.



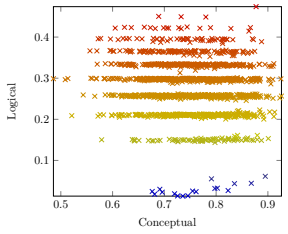
Pearson	Spearman	Kendall
0.016	-0.02	-0.01

(a) Factual (x-axis) and conceptual (y-axis) distances of random contexts.



Pearson	Spearman	Kendall
0.26	0.34	0.23

(b) Factual (x-axis) and logical (y-axis) distances of random contexts.



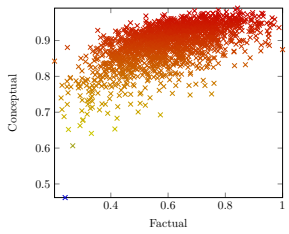
Pearson	Spearman	Kendall
-0.10	-0.12	-0.08

(c) Conceptual (x-axis) and logical (y-axis) distances of random contexts.

Figure 12 Randomly sampled contexts from the Mushroom dataset : correlation between the distance measures.

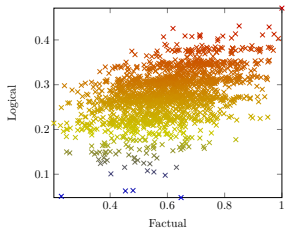
We sampled the Bob Ross dataset as well. It is a 433×133 context based on the apparition, or not, of some features in Bob Ross' paintings, for each episode of The Joy Of Painting. We sampled it into 1500 smaller, 44×10 contexts. In this case, the average density is 0.50.

The results are shown in Fig. 13. In this case, each pair of distances shows a stronger correlation, especially between the factual and conceptual distances. The difference between the two experiments might be related to the density of the contexts, or to intrinsic differences between the data sets, as Bob Ross' paintings usually had the same few elements appearing together in most episodes.



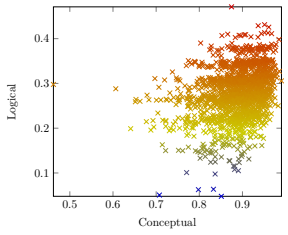
Pearson	Spearman	Kendall
0.60	0.60	0.42

(a) Factual (x-axis) and conceptual (y-axis) distances of random contexts.



Pearson	Spearman	Kendall
0.46	0.45	0.31

(b) Factual (x-axis) and logical (y-axis) distances of random contexts.



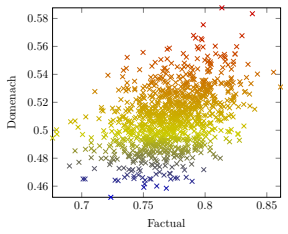
Pearson	Spearman	Kendall
0.30	0.31	0.21

(c) Conceptual (x-axis) and logical (y-axis) distances of random contexts.

Figure 13 Randomly sampled contexts from the Bob Ross dataset : correlation between the distance measures.

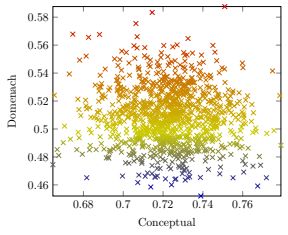
4.5 Comparison with Domenach's Dissimilarity Measure

Domenach's dissimilary measure is based on the *overhanging* relation [10] between sets of objects. Two sets are overhanged if one is a subset of the other and their closures are different. To compute a distance between concept lattices, Domenach defines two matrices, M_1 and M_2, based on the overhanging relation of pairs of objects in each concept lattice. The distance is then based on the L_1 norms of those matrices: $\frac{||M_1 - M_2||}{||M_1|| + ||M_2||}$.



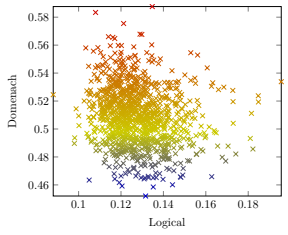
Pearson	Spearman	Kendall
0.41	0.40	0.27

(a) Factual distance (x-axis) and Domenach's dissimilarity (y-axis) of random contexts.



Pearson	Spearman	Kendall
−0.05	−0.03	−0.02

(b) Conceptual distance (x-axis) and Domenach's dissimilarity (y-axis) of random contexts.



Pearson	Spearman	Kendall
−0.17	−0.19	−0.13

(c) Logical distance (x-axis) and Domenach's dissimilarity (y-axis) of random contexts.

Figure 14 Randomly generated contexts : correlation between our distance measures and Domenach's dissimilarity.

We compared our distances with Domenach's dissimilarity measure on 1000 pairs of randomly generated contexts. Fig. 14 depicts the results. We observe that Domenach's dissimilarity measure is independent of our conceptual distance and slightly correlated with our factual distance.

5 Conclusion and Perspectives

We presented three distance families between the most important structures in formal concept analysis, *i.e.* formal contexts, concept lattices and implication bases. These structures represent three complementary points of view on the information contained in formal context: the factual, conceptual and logical points of views. We see the distances we studied in this paper as a first step towards the simultaneous exploitation of the three points of view in the analysis of data.

The applications could be distance-based machine learning, both supervised and unsupervised, or the measurement of the complexity of multidimensional data.

From our point of view, the applications for the work can be multiple. For example, one can study the trajectory and dynamics of a process such as Relational Concept Analysis [21] or Attribute Exploration [11], where the concept lattice and the implication base are built iteratively. Note that since our distances consider contexts with identical dimensions, studying the trajectory of the process would have to be *a posteriori*, and not in an online manner.

Our experiments indicate that, of our distances, only the factual distance is (barely) correlated with the other two and that their correlations depend on the factual distance. In particular, there is no correlation when generating contexts by coin-flipping. As this method has been shown not to produce truly random lattices [6], and the correlation in pseudo-real data is slightly higher, it remains to be seen whether our results are a product of the generation method. In any case, we believe that this is interesting because the concept lattice is supposed to be a lossless representation of the information contained in the formal context but small variations in the context can produce large variations in the lattice. The variation of the correlation w.r.t. other distances should also be studied once we better understand how to control the conceptual and logical distances in the generation of data. Our experiments also highlight the need to study the metric spaces induced by the distances, and their relations, as experimental results are insufficient.

Future work includes the extension of these distances to contexts defined on different sets of objects and attributes, and to the polyadic concept analysis framework.

Resource Availability Statement

The source code for the computation of the distances is hosted at `https://github.com/Authary/FCAD`. The source code for the experiments if hosted at `https://github.com/Authary/experiments_distances_fca`.

References

1 Alexandre Bazin, Jessie Carbonnel, and Giacomo Kahn. On-demand generation of aoc-posets: Reducing the complexity of conceptual navigation. In *Foundations of Intelligent Systems: 23rd International Symposium, ISMIS 2017, Warsaw, Poland, June 26-29, 2017, Proceedings 23*, pages 611–621. Springer, 2017. doi:10.1007/978-3-319-60438-1_60.

2 Alexandre Bazin, Jessie Galasso, and Giacomo Kahn. Polyadic relational concept analysis. *International Journal of Approximate Reasoning*, 164:109067, 2024. doi:10.1016/j.ijar.2023.109067.

3 Alexandre Bazin, Marianne Huchard, and Pierre Martin. Towards analyzing variability in space and time of products from a product line using triadic concept analysis. In *Proceedings of the 27th ACM International Systems and Software Product Line Conference-Volume B*, pages 85–89, 2023. doi:10.1145/3579028.3609019.

4 Karell Bertet, Christophe Demko, Jean-François Viaud, and Clément Guérin. Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. *Theoretical Computer Science*, 743:93–109, 2018. doi:10.1016/j.tcs.2016.11.021.

5 Karell Bertet and Bernard Monjardet. The multiple facets of the canonical direct unit implicational basis. *Theoretical Computer Science*, 411(22-24):2155–2166, 2010. doi:10.1016/j.tcs.2009.12.021.

6 Daniel Borchmann and Tom Hanika. Some experimental results on randomly generating formal contexts. In *CLA*, volume 1624, pages 57–69, 2016. URL: https://ceur-ws.org/Vol-1624/paper5.pdf.

7 Victor Codocedo and Amedeo Napoli. Formal concept analysis and information retrieval–a survey. In *International Conference on Formal Concept Analysis*, pages 61–77. Springer, 2015. doi:10.1007/978-3-319-19545-2_4.

8 Elena Deza, Michel Marie Deza, Michel Marie Deza, and Elena Deza. *Encyclopedia of distances*. Springer, 2009.

9 Florent Domenach. Similarity measures of concept lattices. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*, pages 89–99. Springer, 2015. doi:10.1007/978-3-662-44983-7_8.

10 Florent Domenach and Bruno Leclerc. Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Mathematical Social Sciences*, 47(3):349–366, 2004. doi:10.1016/j.mathsocsci.2003.09.008.

11 Bernhard Ganter, Sergei Obiedkov, Bernhard Ganter, and Sergei Obiedkov. Attribute exploration. *Conceptual exploration*, pages 125–185, 2016. doi:10.1007/978-3-662-49291-8_4.

12 Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer Science & Business Media, 1999.

13 Kathy J Horadam and Michael A Nyblom. Distances between sets based on set commonality. *Discrete Applied Mathematics*, 167:310–314, 2014. doi:10.1016/j.dam.2013.10.037.

14 Marianne Huchard, Pierre Martin, Emile Muller, Pascal Poncelet, Vincent Raveneau, and Arnaud Sallaberry. Rcaviz: Exploratory search in multi-relational datasets represented using relational concept analysis. *International Journal of Approximate Reasoning*, page 109123, 2024. doi:10.1016/j.ijar.2024.109123.

15 Priscilla Keip, Alain Gutierrez, Marianne Huchard, Florence Le Ber, Samira Sarter, Pierre Silvie, and Pierre Martin. Effects of input data formalisation in relational concept analysis for a data model with a ternary relation. In *International Conference on Formal Concept Analysis*, pages 191–207. Springer, 2019. doi:10.1007/978-3-030-21462-3_13.

16 Claudio L Lucchesi and Sylvia L Osborn. Candidate keys for relations. *Journal of Computer and System Sciences*, 17(2):270–279, 1978. doi:10.1016/0022-0000(78)90009-0.

17 Jonas Poelmans, Dmitry I Ignatov, Sergei O Kuznetsov, and Guido Dedene. Formal concept analysis in knowledge processing: A survey on applications. *Expert systems with applications*, 40(16):6538–6560, 2013. doi:10.1016/j.eswa.2013.05.009.

18 Jonas Poelmans, Dmitry I Ignatov, Sergei O Kuznetsov, and Guido Dedene. Fuzzy and rough formal concept analysis: a survey. *International Journal of General Systems*, 43(2):105–134, 2014. doi:10.1080/03081079.2013.862377.

19 Jonas Poelmans, Sergei O Kuznetsov, Dmitry I Ignatov, and Guido Dedene. Formal concept analysis in knowledge processing: A survey on models and techniques. *Expert systems with applications*, 40(16):6601–6623, 2013. doi:10.1016/j.eswa.2013.05.007.

20 UCI Machine Learning Repository. Mushroom. UCI Machine Learning Repository, 1981. DOI: https://doi.org/10.24432/C5959T.

21 Mohamed Rouane-Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. Relational concept analysis: mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, 67:81–108, 2013. doi:10.1007/s10472-012-9329-3.

22 Sebastian Rudolph, Christian Săcărea, and Diana Troancă. Conceptual navigation for polyadic formal concept analysis. In *IFIP International Workshop on Artificial Intelligence for Knowledge Management*, pages 50–70. Springer, 2016. doi:10.1007/978-3-319-92928-6_4.

23 Gerd Stumme, Dominik Dürrschnabel, and Tom Hanika. Towards ordinal data science. *Transactions on Graph Data and Knowledge (TGDK)*, 2023. doi:10.4230/TGDK.1.1.6.

24 George Voutsadakis. Polyadic concept analysis. *Order*, 19:295–304, 2002. doi:10.1023/A:1021252203599.

GraphRAG on Technical Documents – Impact of Knowledge Graph Schema

Henri Scaffidi
School of Computer Science and Software Engineering, The University of Western Australia, Australia

Melinda Hodkiewicz
School of Mechanical Engineering, The University of Western Australia, Australia

Caitlin Woods
School of Computer Science and Software Engineering, The University of Western Australia, Australia

Nicole Roocke
Minerals Research Institute of Western Australia, Australia

Abstract

Retrieval Augmented Generation (RAG) is seeing rapid adoption in industry to enable employees to query information captured in proprietary data for their organisation. In this work, we test the impact of domain-relevant knowledge graph schemas on the results of Microsoft's GraphRAG pipeline. Our approach aims to address the poor quality of GraphRAG responses on technical reports rich in domain-specific terms. The use case involves technical reports about geology, chemistry and mineral processing published by the Minerals Research Institute of Western Australia (MRIWA). Four schemas are considered: a simple five-class minerals domain expert-developed schema, an expanded minerals domain schema, the Microsoft GraphRAG auto-generated schema, and a schema-less GraphRAG. These are compared to a conventional baseline RAG. Performance is evaluated using a scoring approach that accounts for the mix of correct, incorrect, additional, and missing content in RAG responses. The results show that the simple five-class minerals domain schema extracts approximately 10% more entities from the MRIWA reports than the other schema options. Additionally, both the five-class and the expanded eight-class minerals domain schemas produce the most factually correct answers and the fewest hallucinations. We attribute this to the minerals-specific schemas extracting more relevant, domain-specific information during the Indexing stage. As a result, the Query stage's context window includes more high-value content. This contributes to the observed improvement in answer quality compared to the other pipelines. In contrast, pipelines with fewer domain-related entities in the KG retrieve less valuable information, leaving more room for irrelevant content in the context window. Baseline RAG responses were typically shorter, less complete, and contained more hallucinations compared to our GraphRAG pipelines. We provide a complete set of resources at `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main`. These resources include links to the MRIWA reports, a set of questions (from simple to challenging) along with domain-expert curated answers, schemas, and evaluations of the pipelines.

2012 ACM Subject Classification Information systems → Specialized information retrieval

Keywords and phrases RAG, minerals, local search, global search, entity extraction, competency questions

Digital Object Identifier 10.4230/TGDK.3.2.3

Category Research

Supplementary Material *InteractiveResource*: `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain` [30]

Funding *Nicole Roocke*: MRIWA Grant No. M10530. DOI: `https://doi.org/10.71342/574784944778`

Received 2025-03-14 **Accepted** 2025-08-13 **Published** 2025-10-15

1 Introduction

Recent advancements in Large Language Models (LLMs) have revolutionised AI's capability to process and understand natural language, demonstrating strong performance in complex practical tasks. However, LLMs are often limited to their pre-trained knowledge and therefore falter when additional technical domain-specific information is required. Retrieval Augmented Generation (RAG) systems address this issue by enabling the LLM to access and incorporate information from a retrieval source – structured or unstructured knowledge such as documents, tables, or graphs. This process grounds the LLM's responses in additional, domain-specific information that complements its pre-trained dataset. RAG operates through two steps: retrieval and generation. Retrieval involves searching through the retrieval source to rank the information based on its relevance to the query. The most relevant portions are then passed to the LLM for the generation step, where they are combined with the original query to produce more targeted responses. Knowledge graphs (KGs) store explicit knowledge as structured data and are used to enhance the retrieval step [11, 25].

Industry's challenge is that unstructured technical reports contain unstandardised formatting, a mix of text, graphics, and tables, and language requiring domain expertise for interpretation. A specific example of this is MRIWA (Minerals Research Institute of Western Australia). This organisation has a 40-year archive of such reports holding valuable data and insights for the state's minerals industry and research community. Work to understand the impact of different KG schemas on the performance of KG-RAG on complex technical texts is in its infancy. Complex (non-academic) technical texts written by and for industry users are challenging to source as organisations are often reluctant to release them. To use these reports for reproducible KG-RAG research, the documents must be accompanied by KG schemas, question and answer sets, and pipeline answer evaluations by domain experts. This project seeks to address these challenges by evaluating the impact of different KG schemas on the quality of generated responses to a set of curated queries for MRIWA's minerals research technical reports. We release all the artifacts at `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main` to support FAIR[1] research.

The paper is organised as follows. Section 2 reviews the literature on RAG, identifies gaps, and defines the research question. Section 3 describes the structure and content of MRIWA's reports. Section 4 outlines the process of data selection, pipeline functionality, and the experimental setup. Section 5 analyses the results of the experiment.

2 Literature Review

2.1 Overview of Retrieval Augmented Generation

The RAG paradigm is still evolving and can be broadly classified into three processes: baseline RAG, advanced RAG and modular RAG [25]. Baseline RAG has two steps: retrieval and generation. In the retrieval step, text from the document is segmented into chunks, then encoded using an embedding model and stored in a vector database. The RAG system uses the same encoding model for the user query and computes a similarity score between the vectors for the query and the indexed chunk. The top-k chunks with the highest similarity score are used to generate a response to the prompt. Baseline RAG, or "Naïve" RAG, often struggles to identify appropriate chunks, especially when the query requires large amounts of information (exceeding the top-k restriction) to be retrieved and then aggregated across documents, and when the query requires the system to synthesise insights not explicitly stated in reports [11].

[1] FAIR stands for Findable, Accessible, Interoperable, and Reusable.

Advanced RAG uses pre-retrieval and post-retrieval strategies to improve indexing and streamline the retrieval process. For example, there is growing interest in utilising additional content, specifically structured data such as knowledge graphs, to enhance the retrieval stage in RAG. This is called GraphRAG [26, 27]. Specifically, altering the retrieval source structure, from unstructured text to graph-based knowledge, for example, can improve the quality of retrieved content, which helps to reduce LLM hallucinations [16] and improve summarisation performance [7]. A modular RAG framework introduces new modules into the pipeline, allowing for substitution or reconfiguration of specific steps, making the process more adaptable to specific use-cases beyond traditional sequential retrieval and generation [11].

2.2 Domain-Specific Knowledge Graph Schemas and Construction

A KG is a collection of triples, where each triple represents a relation (edge) between two entities (nodes): *relation(Entity1, Entity2)*. The KG schema defines the types of entities and relations permissible in the KG. GraphRAG relies on a KG to retrieve relevant information. However, manually curating a KG is labor-intensive and time-consuming, often relying on crowd-sourced efforts like Freebase [3] and Wikidata [35]. As a result, automated KG construction methods are crucial for efficiently extracting structured knowledge from unstructured or semi-structured data, significantly reducing the manual burden [42]. The challenge lies in designing a KG schema that balances domain specificity with adaptability, ensuring the KG can incorporate new concepts, evolving terminology, and varied technical document structures.

Broadly, there are two approaches to KG schema development: top-down and bottom-up. The top-down approach develops a schema based on domain knowledge and the use case, sometimes informed by existing ontologies if available [24]. Bottom-up approaches identify concepts of interest using entity discovery processes to discover strings referring to semantic entities and then classify them into types, which collectively define the schema [36, 41, 42]. In practice, a combination of both approaches is usually deployed.

Within the minerals domain are several technical disciplines, each with its own processes, norms, and data models. These disciplines contribute at different stages along the mineral processing value chain – find and define the deposit (geology, geochemistry, geostatistics), consider options for processing (geometallurgy), identify risks and quantify the value (mining finance), mine planning and execution (mining engineering), mineral processing (metallurgists) to produce the saleable product (marketing), environmental management and closure (environmental engineers). These disciplines draw on the sciences – including geology, chemistry, physics, and mathematics – as well as all branches of engineering and business. While there has been work on developing schemas and ontologies in subdomains such as geology [6, 12], there is no publicly available KG schema or ontology that spans the entire minerals domain value chain. An ontology is a formal and semantically rich specification used when reasoning and interoperability are required.

2.3 KG-RAG Retrieval Processes

As the literature on RAG has grown, more studies have sought to integrate KGs into the pipeline, particularly as the field acknowledges the limitations of vector search over unstructured text [11]. Broadly, these efforts can be categorised into two approaches: using LLM-generated KG queries [10, 17, 18, 37, 39], and machine learning-based KG querying [15, 16, 23].

2.4 Microsoft's GraphRAG

Microsoft's GraphRAG [7], introduced in 2024, uses LLMs to generate the KG from text. Graph machine learning is utilised to build a bottom-up hierarchical cluster of the KG using an LLM to pre-summarise each cluster and produce a set of community reports. The KG and community reports are retrieved via two search options: **local search** and **global search**. Local search identifies semantically related entities within the KG, based on the prompt, and retrieves closely connected information. This search method is suited to answer targeted queries about specific entities [19]. Global search retrieves all community reports at a specific level of the hierarchical cluster and conducts Map-Reduce to synthesise a response. This search method is designed to answer high-level aggregation queries about the dataset.

Given the novelty of GraphRAG (released July 2, 2024), there is no research testing it in the minerals domain. As shown in Table 1, the published literature on Microsoft GraphRAG is limited. In our analysis, surveys were excluded, and only studies that evaluated Microsoft GraphRAG's question–answering performance were included. Across the 13 identified studies, most used broad, general-domain public datasets such as UltraDomain or HotpotQA, while only a small proportion applied GraphRAG to narrow industrial and technical domain corpora. LLM-based judging was the predominant evaluation approach, with few examples of human assessment or automated retrieval metrics. Citation counts are generally low, and impact factors, where applicable, are modest. These trends highlight a gap in the literature, which our work addresses by evaluating GraphRAG's performance in a real-world context within the minerals domain using human subject-matter expert evaluation.

Table 1 Summary statistics from published studies evaluating Microsoft GraphRAG ($n = 13$).

Metric	Value
Average citations (Google Scholar, Aug 2025)	1.15
Average impact factor (where reported, $n = 2$)	6.15
Papers using general-domain public datasets rather than narrow industrial and technical domain corpora	9 out of 13
Papers using LLMs as evaluators of pipeline performance	9 out of 13

Many organisations are seeking to use RAG to query their internal documents, and given their existing commitment to the Microsoft ecosystem, Microsoft's GraphRAG has become a relevant product worth investigating.

2.5 Research Question

Based on the gaps in the literature discussed in this section we propose the following research question: Given a set of minerals research reports that are accessed in a GraphRAG pipeline, how do different KG schemas of the report's subject matter impact the quality of responses to local and global search queries?

3 Data

The Minerals Research Institute of Western Australia (MRIWA) is a statutory government body whose purpose is to foster and promote minerals research for the benefit of the State of Western Australia. The organisation provides financial assistance to industry and research entities, with project outcomes typically presented in technical reports. MRIWA has accumulated a portfolio of over 300 such reports, each detailing project undertakings and findings [22].

Currently, querying the reports is limited to keyword search on project abstracts on the MRIWA website, enabling users to identify relevant reports which are then available for download. This approach is often ineffective, as information in the full reports may not be adequately captured in their abstracts. Once reports are downloaded, the process of locating specific information in the full reports can be challenging due to their technical content, length (the mean number of pages is 163, with approximately 7% of reports more than 400 pages long) and inconsistent formats. These issues create a challenge for MRIWA, their stakeholders, and research community members to extract data and insights that are relevant to the sector.

To illustrate the complexity of the report content, 50 text samples are analysed. Each sample is randomly selected from a page in the MRIWA report repository. The samples are a minimum of one sentence in length and consist solely of body text. Based on the samples, the academic content in the reports has an average Flesch Reading Ease of 45 out of 100, where plain English typically scores between 60 and 70 [32]. Specialised domain expertise is often required to interpret this technical content. Below are examples of typical text found in MRIWA's reports:

> *"In the footwall Li zone, the absence of petalite and fluid inclusion work indicate formation of (feldspar-free) quartz-spodumene assemblage. . . "* [38].
>
> *"The interaction with gangue may be advantageous (for example, pyrite may enhance copper leaching by increasing the fluid progress into the core, as was presented by Dutrizac et al. (1971)), or be detrimental (such as reagent consumption in carbonate dissolution (Sinclair and Thompson, 2015)), depending on the gangue type."* [13]
>
> *"The low shaft power needed for this type of vertical flow liquid-liquid contacting ensures a uniform drop size in the dispersion, resulting in very low entrainment (Ju et al. 1991)."* [28]

MRIWA's reports exist as either scanned PDFs or converted Word documents and their content includes tables, figures, and text. With no standardised format or template, significant formatting inconsistencies exist across all reports. Figure 1 highlights these inconsistencies, including variations in figure captions, column layouts, headers and footers, and section headings.

MRIWA seeks to support their stakeholders to query these reports using natural language, unlocking information in these historical documents to foster new insights for the WA minerals industry.

In addition to the MRIWA reports, the data set used for this project includes a set of competency questions and subject matter expert ideal answers.

3.1 Competency Questions and Ideal Answers

The validation input for our analysis consists of a set of competency questions to assess the effectiveness of a minerals domain RAG pipeline. The questions were developed by MRIWA's CEO and several members of the MRIWA staff. As representatives of the typical users of the RAG pipeline, they provided questions that reflect real-world information needs in the minerals research sector. The quantity of information required to answer each question and the complexity of each question varies. We have categorised these questions into five groups based on the CRAG [40] classification scheme: simple, simple with condition, set, aggregation, and post-processing-heavy (see Table 2 and Table 3). The CRAG paper defines eight question types, five of which apply to our question set. The three CRAG question types not covered by the CEO's questions are comparison, multi-hop, and false-premise.

For each competency question the MRIWA team developed an ideal answer[2]. These ideal answers are designed specifically for use with our evaluation approach outlined in Section 4.5.

[2] https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/data/mriwa_cqa

Table 2 Number of MRIWA competency questions by CRAG [40] classification.

Question Type	Description	Count
Simple	Questions asking for simple facts that are unlikely to change overtime, such as the birth date of a person and the authors of a book.	4
Simple with Condition	Questions asking for simple facts with some given conditions, such as stock prices on a certain date and a director's recent movies in a certain genre.	2
Set	Questions that expect a set of entities or objects as the answer (e.g., "what are the continents in the southern hemisphere?").	4
Aggregation	Questions that require aggregation of retrieval results to answer (e.g., "how many Oscar awards did Meryl Streep win?").	2
Post-Processing-Heavy	Questions that need reasoning or processing of the retrieved information to obtain the answer (e.g., "how many days did Thurgood Marshall serve as a Supreme Court justice?").	3

Table 3 MRIWA competency questions grouped by query type.

Simple
Identify which MRIWA reports reference MERIWA or MRIWA.
Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with in any capacity (including being listed in references)?
Which MRIWA report is related to the East Kimberley region?
Which MRIWA reports relate to leaching?
Simple with Condition
Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a researcher?
Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a sponsor?
Set
Extract all references to MERIWA and MRIWA from the MRIWA reports.
Identify any references to nickel or Ni in the MRIWA reports.
Which elements are considered in the MRIWA reports?
Which regions of Western Australia are referenced in the MRIWA reports?
Aggregation
Which MRIWA report author has been involved in more than one report/project?
What is the average number of references in each MRIWA report?
Post-Processing-Heavy
Which MRIWA reports relate to exploration?
Which MRIWA reports relate to mining extraction?
Which MRIWA reports relate to mineral processing?

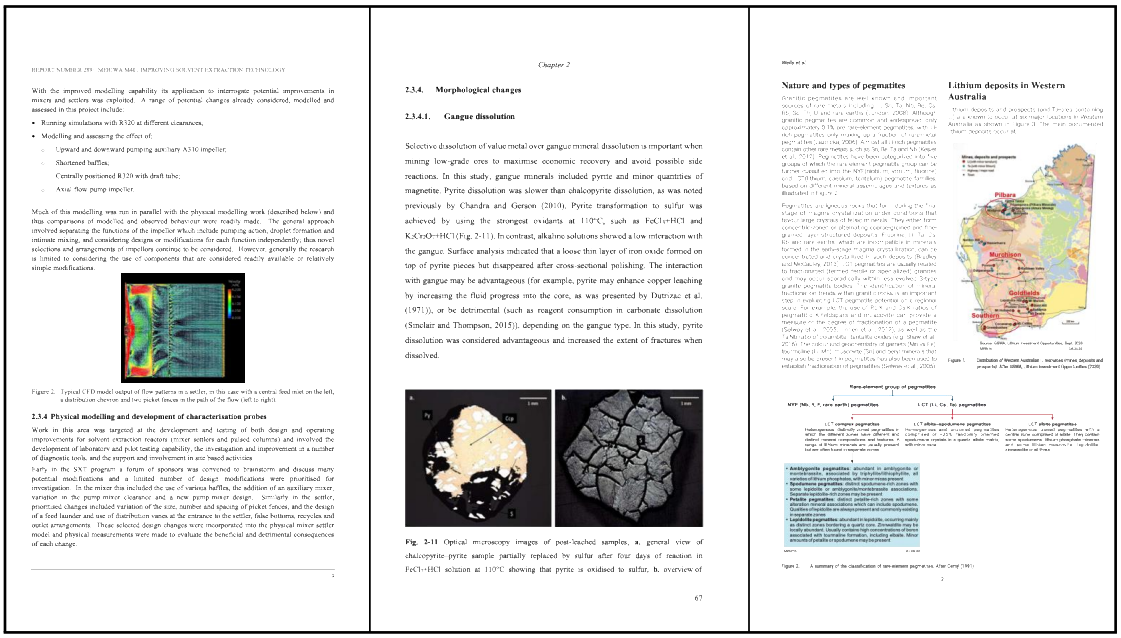


Figure 1 Examples of formatting inconsistencies across typical pages in MRIWA reports M0401 [28] (left), M0488 [13] (middle), and M0532 [38] (right). There is inconsistent formatting of figure captions, column layouts, headers and footers, and section headings, for example.

Some of the queries test GraphRAG's ability to find and return concepts not explicitly referenced in the query. For example, the question *"Which MRIWA report is related to the East Kimberley region"* requires the model to identify towns, mine sites, and geological locations in the East Kimberley region, not rely on the regular expression *East Kimberley*. Similarly, for *"reports relating to leaching"* we hope to find reports relating to leaching from a minerals processing perspective (the context of the question) and which mention any chemical process that extracts valuable minerals from ore by dissolving with a solvent.

4 Process

This section describes sample selection, KG schema development, and the experimental setup, which is based on four KG schema options, each examined using the local and global search strategies in Microsoft GraphRAG.

4.1 MRIWA Report Sample Selection

A subset of MRIWA's reports is sampled based on domain-expert opinion informed by natural language processing to count unique tokens, for instance. The selection process aims to ensure the data set covers multiple minerals domains and represents the complexity of the report set. A subset of 15 reports was selected, totalling approximately 2.7 million tokens – comparable in scale to the datasets used in the original Microsoft GraphRAG paper [7]. We believe this scale increases both the reliability of our results and the usefulness of the resources. The selection process is described in detail in the GitHub repository[3]. The selected PDF reports were converted to individual ".txt" files using the PyPDF2 [9] Python package for use in subsequent steps.

[3] `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/blob/main/supplementary_materials/mriwa_report_sample_selection.md`

4.2 Knowledge Graph Schema Development

The development of a minerals-domain KG schema began with a review of the concepts in MRIWA report executive summaries. These executive summaries (300-1000 words) provide an overview of topics and domain concepts covered in the full report content. Initially, we explored annotation classes (`Activity`, `Physical Object`, `Process`, `Property`, and `State`) used in our previous industrial annotation work[4]. This resulted in almost every item in the text having an annotation, and much of what was being annotated was irrelevant to addressing the questions being asked. We then explored discipline-specific (e.g., geology and mining) entities, but these were difficult to annotate consistently due to the spans. Following this experience, and with the assistance of MRIWA subject matter experts, we landed on more abstract concepts like `Naturally Occurring Object` to describe geological formations (present in many of the reports) and separated these from `Processed Object`, which are produced in some engineered process. We developed separate classes for objects, processes, location-specific concepts and for identifying organisations. We called this the Minerals Domain Schema (MDS). Early tests showed GraphRAG performed well with these aggregated concepts.

We then extended this schema to include some additional subclasses, splitting `Processed Object` into `Processed Material` and `Manufactured Product` (mirroring the mineral processing and manufactured output in mining), and sub-classing `Process` into `Natural Process`, `Lab testing Process` and `Industrial Process`. This was called the Expanded Minerals Domain Schema (EMDS). For a summary of these two schemas, see Table 4. We used our QuickGraph annotation tool[5] [2] throughout this iterative process. Figure 2 provides a detailed example of entity typing under the EMDS from our manual annotation sessions. If the MDS were applied instead, the same entities would be labelled as follows:

- Entities labelled as `Natural_Process`, `Lab_Process`, or `Industrial_Process` would all be labelled as `Process`.
- Entities labelled as `Processed_Material` or `Manufactured_Product` would both be labelled as `Processed_Object`.

In considering the entity types in the MDS and EMDS, we also considered the possibility that future work might wish to align the schema classes to the Basic Formal Ontology (BFO) [14] or another top level ontology. Thus, the entity types `Naturally_Occurring_Object`, `Processed_Object` and `Manufactured_Product` can be mapped to the BFO class `bfo:MaterialEntity`, `Process` to `bfo:Process`, `Organisation` to `bfo:ObjectAggregate`, and `Site_Location_Boundary` to `bfo:ImmaterialEntity`.

Whilst defining relationships between entities would render the schemas more expressive, we do not, as GraphRAG [7] supports only the pre-definition of entity types. Relations are generated dynamically by the LLM during *Indexing* (see Section 4.3.1). Results of GraphRAG entity extraction using the schemas, including quantities of entities extracted, are discussed in Section 5.1.1.

4.3 GraphRAG

Microsoft's GraphRAG approach [7] is designed to answer complex semantic queries and aggregation queries. Figure 3 shows the two GraphRAG phases: *Indexing* and *Query*.

[4] `https://github.com/nlp-tlp/maintie`
[5] `https://quickgraph.tech/`

Table 4 Minerals Domain Knowledge Graph Schema and Expanded Minerals Domain Knowledge Graph Schema.

Minerals Domain Knowledge Graph Schema (MDS)		
Entity Type	**Description**	**Superclass**
Naturally Occurring Object	Objects formed naturally.	Object
Processed Object	Objects formed through human intervention.	Object
Process	Activities in which one or more objects participate.	N/A
Site Location Boundary	Immaterial spatial regions.	N/A
Organisation	Person/s organised together for a purpose.	N/A
Expanded Minerals Domain Knowledge Graph Schema (EMDS)		
Entity Type	**Description**	**Superclass**
Naturally Occurring Material	Objects formed naturally.	Object
Processed Material	Objects formed and materials used in an industrial process.	Object
Manufactured Product	Objects that are the product of a manufacturing process.	Object
Natural Process	Naturally occurring processes.	Process
Lab testing Process	Processes involving chemical and other analyses.	Process
Industrial Process	Processes that produce a product.	Process
Site Location Boundary	Immaterial spatial regions.	N/A
Organisation	Person/s organised together for a purpose.	N/A

4.3.1 GraphRAG Indexing

The *Indexing* stage takes our multi-domain subset of MRIWA reports and splits them into 300-token chunks with 100-token overlap. Utilising the specified schema, the LLM extracts both entities and relations from each chunk. Entities and relations are in the form `{name, type, description}` and `{source entity, target entity, description}` respectively. Entities with the same `name` and `type` are merged by the LLM to summarise all the descriptions into one summary per entity. Relations with the same `source entity` and `target entity` are merged in the same manner. This process yields a KG of MRIWA reports. An example of an entity and relation relating the concept *Rutile* is displayed in Table 5, depicting the detail that is captured through *Indexing*.

Community detection is conducted on the KG using the Leiden algorithm [34]. The algorithm is recursively applied to generate a hierarchical cluster of the graph. The LLM summarises each community into a community report. Each community report provides a summary of key points and outlines the important entities and relations present within the community. These community reports vary in their level of abstraction, depending on the level of the community within the hierarchy. The rationale, here, is to enable the LLM to answer high-level aggregation queries, hence addressing a baseline RAG weakness. The text chunks, entities, relations, and community reports are all vectorised and stored for retrieval during the *Query* stage.

4.3.2 GraphRAG Query: Local and Global Search

The *Query* stage of GraphRAG comprises two search options: **local search** and **global search**. Local search vectorises the query and identifies a set of semantically related entities within the KG. Based on connectedness to these entities, additional candidate entities, relations, community

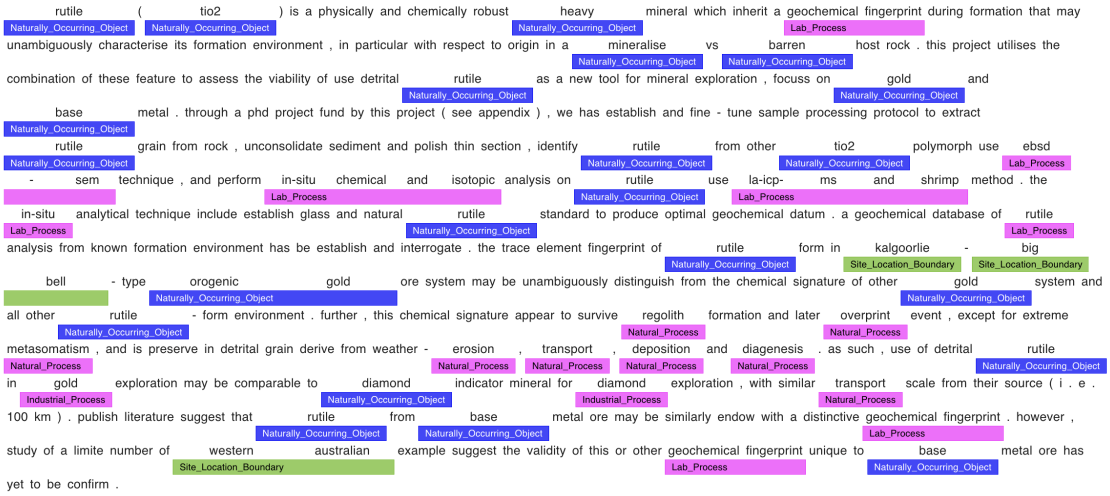


Figure 2 Example of a manually annotated pre-processed executive summary from MRIWA Report 448 using the Expanded Minerals Domain Schema.

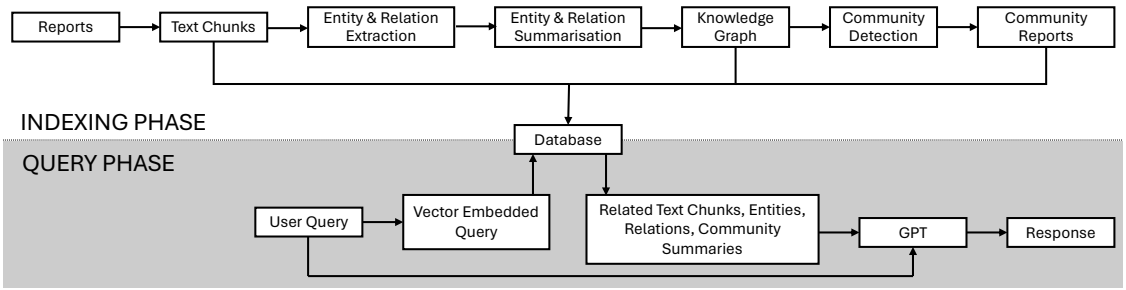


Figure 3 Two phases of the Microsoft GraphRAG pipeline (derived from [19]).

reports, and text chunks are identified. These information sources are ranked and filtered, then passed to the context window of the LLM to produce a response. This search method is suited to answer targeted queries about specific entities [19].

In contrast, global search is designed to answer high-level aggregation queries about the dataset. First, all community reports at a specific level of the hierarchy are retrieved and chunked. In parallel, each chunk is used by the LLM to produce an intermediate response to the query. Each response is ranked by importance, where the most important are passed to the context window for the LLM to produce a final response.

In addition to local and global search, the GraphRAG framework includes other query methods such as **basic search** and **DRIFT search**. The basic search option is an implementation of baseline vector RAG and was used in our experiments for comparison. The DRIFT search method was released after our experiments were conducted and is therefore not evaluated.

4.4 Experimental Setup

The experiment utilises GraphRAG Version 0.1.1 [20] in its default configuration with GPT-4o-Mini (temperature: 0.0). We run GraphRAG *Indexing* in four separate pipelines, each using one of the following KG schemas as the `entity_types` parameter for entity and relation extraction.

Table 5 Example GraphRAG entity and relation of rutile and gold exploration.

Entity Name	Entity Type	Entity Description
Rutile	Naturally Occurring Object	Rutile is a **naturally occurring mineral** primarily composed of titanium dioxide (TiO2) and is significant in various geological contexts. It is commonly found in metamorphic and igneous rocks, including sedimentary deposits, and is characterized by its high refractive index, strong dispersion, and distinct chemical properties. Rutile often occurs...
Relation Source Entity	**Relation Target Entity**	**Relation Description**
Rutile	Gold Exploration	Rutile is a mineral that plays a significant role in gold exploration. It is studied for its trace element signatures, which provide insights into the conditions under which it formed. As an indicator mineral, **rutile assists in identifying potential gold deposits**, making it a valuable **tool in the gold exploration process**. Its properties allow it to serve as a geochemical indicator, helping geologists locate areas that may contain gold. Overall, rutile's...

1. **Minerals Domain Schema (MDS):**
   See Table 4. The LLM is not told the descriptions of the schema's entity types, only the names of the entity types. Pipeline configuration[6].
2. **Expanded Minerals Domain Schema (EMDS):**
   See Table 4. The LLM is not told the descriptions of the schema's entity types, only the names of the entity types. Pipeline configuration[7].
3. **Auto-Generated Schema (AGS):**
   This schema is generated by an LLM through the GraphRAG Prompt Tuner. We utilise the default LLM prompts, not the auto-tuned LLM prompts, to ensure a fair comparison between the other schema pipelines. We only utilise the entity type names generated by the Prompt Tuner, as follows.
   `{chemical process, mineral, geological survey, geochemistry, exploration technique, sample, formation, project, research study}`
   Pipeline configuration[8].
4. **Schema-less (SL):**
   This pipeline leaves the KG schema unspecified, and instead prompts the LLM to "identify all entities needed from the text in order to capture the information and ideas in the text". Pipeline configuration[9].

[6] https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/minerals_domain_schema
[7] https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/expanded_minerals_domain_schema
[8] https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/auto_generated_schema
[9] https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src/schema_less

Additionally, we included **Baseline RAG** for comparison purposes. For this, we utilise a basic search pipeline from GraphRAG Version 2.3.0 [21] in its default configuration with GPT-4o-Mini (temperature: 0.0, chunk size: 300 tokens, chunk overlap: 100 tokens). The pipeline configuration is available[10].

Our code is available at `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/src`.

4.5 Competency Question Evaluation

Each of the four pipelines, through both local and global search, as well as the basic search pipeline, are asked 15 competency questions defined by MRIWA (see Table 3). Results can be accessed at `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/results`.

Evaluation of RAG systems focuses on two key aspects: retrieval quality, which refers to the usefulness of retrieved information for answering a query, and generation quality, which measures the ability to generate accurate and relevant responses from the retrieved information [11]. GraphRAG uses our minerals reports and summarises them into a KG. It is not plausible to define a correct set of nodes/summaries to retrieve for a certain query, given that the number of nodes in the generated KG exceeds 195,000 for each of our trials. This precludes the use of many traditional metrics of retrieval quality, including Hit Rate and Mean Reciprocal Rank, which focus on the frequency at which the system retrieves the "correct" piece of information. Studies that have existing SQL or Graph databases are able to define concrete queries or answers for which the LLM's output can be compared, enabling the use of metrics including Average Overall Execution Accuracy [31]. This technique is also not possible with the size of the GraphRAG-generated database of our technical reports. Our evaluation of generation quality is constrained to the use of our MRIWA-defined ideal answers, which contain key information expected in a completely correct response, as a basis of comparison. Given the technical nature of MRIWA's reports, however, this process of comparison requires subject matter expertise which further precludes using automated LLM-based evaluation approaches, like Retrieval Augmented Generation Assessment (RAGAS) [8] and Automated RAG Evaluation System (ARES) [29], which use non-expert general-purpose LLMs to conduct the evaluation.

For example, the answer to our simplest question *"identify which MRIWA reports reference MERIWA or MRIWA"* should be a simple list of report identifiers (MRIWA reports 432, 448, etc.). Instead, when using GraphRAG with the auto-generated schema, the response (shown below) is very verbose, containing a subset of the correct report numbers as well as information about the reports that was not requested and in some cases not completely correct.

> *"The following MRIWA reports reference either MERIWA or MRIWA:*
> *1. **MRIWA REPORT M532**: This report focuses on the geology, mineralogy, and geometallurgy of electric vehicle (EV) materials deposits in Western Australia. It highlights the significance of lithium production and provides insights into local resources associated with these deposits. The report is a comprehensive study that contributes to understanding the characteristics and potential of EV materials in the region [Data: Reports (12493); Entities (290, 10929, 1445, 23142, +more)].*
> *2. **MRIWA REPORT NO. 488**: This report details research findings on fluid-rock interactions and lixiviant/oxidant behavior for metal recovery from deep ore bodies. It is part of a broader initiative to understand the processes involved in mineral extraction and recovery [Data: Sources (9809, 9748, 9861); Entities (290, 10929, 1445, 23142, +more)].*

[10] `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/supplementary_materials/basic_search`

*3. **FINAL REPORT TO MRIWA**: This document encapsulates the findings and analyses from the By-products (LEAF) Project M451, detailing the project's outcomes, methodologies, and implications regarding soil amendments and their effects on leachate. It serves as a vital resource for understanding the impacts of soil amendments on leachate and the broader environmental context [Data: Reports (11245); Entities (290, 10929, 1445, 23142, +more)].*
These reports illustrate the collaborative efforts and research initiatives overseen by MRIWA, contributing to advancements in mineral research and environmental management in Western Australia. The references to MRIWA in these reports underscore its role in funding and overseeing significant research projects in the region."

Evaluation of long-form answers, like the above, in technical domains is in its infancy, and traditional benchmarking methods that rely on standard train/test datasets have shown limitations in assessing LLM-generated responses [1, 4]. The process still depends, in large part, on evaluation against domain-expert provided model answers.

To evaluate these RAG pipelines on the competency questions we used a human-in-the-loop approach, incorporating subject matter evaluation and reference to the ideal answers.

Our classification approach is informed by evaluation measures for 1) Quality (Context Relevance, Answer Faithfulness, Answer Relevance) and 2) Required Abilities (Noise Robustness, Negative Rejection, Information Integration, and Counterfactual Robustness) [11]. Rather than provide an assessment against each of these seven dimensions, which was very time-consuming and difficult to agree on, we created a five-level classification approach shown in Table 6. We found that the wording enabled evaluators to provide a yes/no answer for each level. While based on the aforementioned dimensions, our classification scheme is coherent with the metrics proposed in RAGEval [43]: Completeness (B1 and B2), Irrelevance (B3 and B4), and Hallucination (B5).

Table 6 Evaluation of GraphRAG pipeline performance.

Code	Name	Criteria
B1	High Answer Relevance	All *ideal answer* information is present in the response.
B2	Medium Answer Relevance	Some *ideal answer* information is present in the response.
B3	Medium Noise Robustness	Some factually correct information that is closely related to the question, but that does not directly answer the question, is present in the response.
B4	Low Noise Robustness	Some factually correct information that is irrelevant to the question is present in the response.
B5	Hallucination	A hallucination is present in the response.

Two independent raters classified all 135 responses to the competency questions using the classification scheme. Each question's ideal answer outlined the specific information required for a response to be considered fully correct. Prior to rating, the raters reviewed each ideal answer and agreed on how the classification scheme would apply to that particular question, ensuring consistent interpretation. For example, the question "Identify which MRIWA reports reference MERIWA or MRIWA" has the following ideal answer and agreed upon evaluation scheme:

`Final-Report_448_MRIWA_M0448.txt`: MRIWA
`Final-Report_278_MERIWA_M0384.txt`: MERIWA

`Final-Report_289_MERIWA_M0401.txt`: MERIWA
`Final-Report_532-MRIWA_M0532.txt`: MRIWA
`Final-Report_479_MRIWA_M0479.txt`: MRIWA
`Final-Report_459-489_MRIWA_M0459-M0484.txt`: MRIWA
`Final-Report_458_MRIWA_M0458.txt`: MRIWA
`Final-Report_476_MRIWA_M0476.txt`: MRIWA
`Final-Report_432_MRIWA_M0432.txt`: MRIWA
`Final-Report_470_MRIWA_M0470.txt`: MRIWA
`Final-Report_295_MERIWA_M0416.txt`: MERIWA
`Final-Report_494_MRIWA_M0494.txt`: MRIWA, MERIWA
`Final-Report_488_MRIWA_M0488.txt`: MRIWA
`Final-Report_451_MRIWA_M0451.txt`: MRIWA
`Final-Report_505_MRIWA_M0505.txt`: MRIWA

B1 – All above reports noted as referencing MRIWA/MERIWA.
B2 – Some of above reports noted as referencing MRIWA/MERIWA.
B3 – Discussions about MRIWA/MERIWA in MRIWA's reports generally, but not specifying or referring to the specific MRIWA report/s that mention MRIWA/MERIWA.
B4 – Discussions not about or unrelated to MRIWA/MERIWA in MRIWA's reports (e.g., MRIWA appearing in non-MRIWA reports).
B5 – False statement.

Multiple performance classifications (B1 – B5) may be attributed to a single response. For example, a response may contain "all ideal answer information" (B1) from our MRIWA-defined ideal answer but also include "some factually correct information that is irrelevant to the question" (B4).

5 Results and Discussion

5.1 KG Schema and Local-Global Search Performance

The evaluation results (averaged across the raters) using the five-level classification system for each GraphRAG pipeline (with four KG schema options, and local/global search) and the Baseline RAG pipeline are shown in Table 7. All pipeline responses are available[11]. As described in Section 4.5, the GraphRAG responses contain multiple paragraphs, relevant and irrelevant answers and facts, and in some cases reference the retrieved information. The average length of responses is 302.40 words (396.97 tokens).

We make the following observations about the schema pipeline data in Table 7 before presenting in-depth discussions on the performance of specific KG schemas in the following subsections. In discussing the performance classifications, we consider a good performance as having a high total for B1 and B2 (relevance) and a low total for B5 (hallucinations). B3 and B4 (robustness) are informative in that a high total indicates that additional peripheral information is being returned that may or may not be relevant to the question.

Relevance (B1 and B2):

- All schemas exhibit similar levels of relevance (B1 and B2).
- 57% of responses exhibit B2 (contain some ideal answer information) indicating majority of responses are partially correct.

[11] `https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain/tree/main/supplementary_materials/performance_analysis`

Table 7 Average total number of responses per pipeline that exhibit each performance classification (B1–B5).

	Local Search					Global Search				
	B1	**B2**	**B3**	**B4**	**B5**	**B1**	**B2**	**B3**	**B4**	**B5**
Minerals Domain Schema	2.0	9.0	11.0	4.0	2.0	0.0	8.5	9.0	8.5	1.5
Expanded Minerals Domain Schema	0.5	9.5	8.0	3.0	2.5	1.0	8.0	7.0	9.5	1.5
Auto-Generated Schema	2.0	9.0	6.5	1.0	3.0	0.5	8.0	5.5	9.5	4.0
Schema-less	1.0	8.0	8.0	3.5	3.5	0.5	8.0	9.5	9.5	2.0
	Basic Search									
	B1	**B2**	**B3**	**B4**	**B5**	–	–	–	–	–
Baseline RAG	1.0	8.5	6.0	4.0	3.5	–	–	–	–	–

- Only approximately 7 of the 120 responses received a B1 score (all ideal answer information present in the response).
- Local search slightly outperforms global search on relevance (B1 and B2).

Robustness (B3 and B4):

- Global search pipelines produced more irrelevant information (B4) than local search pipelines.
- The MDS pipeline with local search produced fewer examples of peripheral information (B3 and B4) than MDS using global search.
- The MDS pipeline produced the second highest amount of peripheral information (B3 and B4) and produced the most relevant peripheral information (B3).

Hallucination (B5):

- The MDS pipeline produced hallucinations (B5) at a lower rate (though comparable with EMDS) than the other pipelines, only hallucinating on the following four competency questions: Two *Simple* queries "Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with in any capacity (including being listed in references)?", "Which MRIWA report is related to the East Kimberley region?", one *Simple with Condition* query "Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a sponsor?", and one *Aggregation* query "What is the average number of references in each MRIWA report?"
- All local search variants hallucinated on the *Aggregation* query "What is the average number of references in each MRIWA report?", while none of the global search pipelines hallucinated.
- The highest number of hallucinations is from the AGS pipeline using global search, hallucinating on four out of the 15 competency questions.

General Observations:

- The MDS and EMDS pipelines using local search had the most desirable set of scores.
- Global search results of all pipelines are not clear to rank in terms of performance.

5.1.1 Comparison of KG schema selection

These results suggest that GraphRAG on MRIWA's reports performs the best when the MDS or EMDS is used with local search. To elicit possible reasons for this we examined the number of entities in the KG generated by each pipeline as shown in Table 8. We make the following observations.

The MDS pipeline extracts approximately 10% more entities from MRIWA's reports than the other three pipelines. We hypothesise that with an additional 20,000 entities, the MDS pipeline extracts and retrieves more relevant information from MRIWA's reports. As a result, its context

Table 8 Statistics of each pipeline's generated knowledge graph.

	Minerals Domain Schema	Expanded Minerals Domain Schema	Auto-Generated Schema	Schema-less
# Entities	**218274**	202878	198090	195930
# Relations	59087	57580	60602	67791
# Communities	6056	5867	5807	6421

window includes more high-value content, which contributes to the observed improvement in answer quality compared to the AGS and SL pipelines. This also explains the additional relevant peripheral information present in MDS responses compared to EMDS.

In contrast, pipelines with fewer entities in the KG retrieve less valuable information, leaving more room for irrelevant content in the context window. This increases the risk of distracting the LLM, a phenomenon known as the Distraction Problem [5], and would explain the worse response quality to competency questions.

The five entities in the MDS represent the abstract concepts in the MRIWA texts. The MDS pipeline extracted the highest number of entities from MRIWA's reports compared to the other pipelines. This observation raises an important question: why does specifying the MDS lead to more entities – but not more relations or communities – being extracted into the KG compared to other schema options? In the MDS pipeline, we only provided GraphRAG with five abstract entity types: `Naturally_Occurring_Object`, `Processed_Object`, `Process`, `Site_Location_Boundary`, and `Organisation`, without any examples to define these terms. We hypothesise that the MDS may be easier for a general-purpose LLM like GPT-4o-Mini to apply because 1) limiting the number of classes reduces complexity for the LLM, and 2) the classes are broad and require less domain-specific expertise to interpret. This may explain why the EMDS schema – which adds domain-specific subclasses to these same five types – resulted in fewer entities being extracted. However, we found no prior literature that explains this phenomenon.

These results suggest there may be an optimal level of abstraction when defining entity types for GraphRAG pipelines. A schema that is too specific may constrain the LLM and reduce recall, while one that is too generic may overlook important distinctions. Striking the right balance appears to be a critical consideration in KG schema design for GraphRAG, and we identify this as an area for future research.

Figures 4 and 5 show that object-related entity types dominate the KGs in both the MDS and EMDS pipelines, which is expected given the nature of MRIWA's reports. However, we did not observe a clear difference in how well the pipelines answered object-related versus process-related questions, so the impact of this imbalance is unclear.

Another finding is that the LLM occasionally labelled entities with types that were not part of the specified schema. For example, in the MDS pipeline, approximately 3% of the 218,000 extracted entities were assigned types including, but not limited to, `Person`, `Technology`, and `Drillhole`, despite not being among the five MDS-defined entity types. This demonstrates the impact of the non-deterministic nature of LLM outputs, even with temperature set to 0.0, and highlights a limitation of using prompts alone to constrain entity extraction. This limitation may be mitigated by incorporating a post-processing validation step to enforce schema compliance and by tailoring the GraphRAG few-shot examples in the entity extraction prompt to the domain of interest.

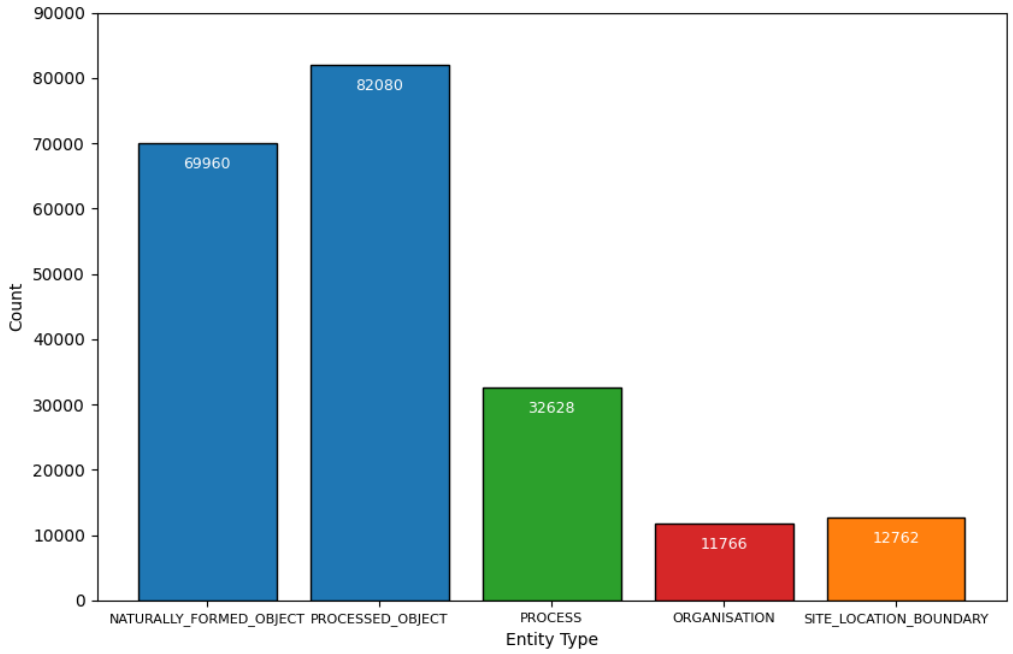


Figure 4 Distribution of MDS entity types extracted by the MDS GraphRAG pipeline.

5.1.2 Comparison of Local and Global Search

The results in Table 7 suggest global search responses tend to contain more peripheral information (higher B3 and B4 counts) than local search responses.

This result is expected as global search is designed to answer queries at a higher level of abstraction. In doing so, the approach includes a larger breadth of information during the generation procedure (approximately 15 to 20 times the token usage of local search), which may explain the inclusion of additional peripheral information in the final response.

This finding implies that global search is better suited for high-level queries requiring broad information coverage. We observe this to be true for our dataset and competency questions. For example, global search responses to "Which elements are considered in the MRIWA reports?" and "Which regions of Western Australia are referenced in the MRIWA reports?" typically list more elements and regions than local search, and offer more detailed answers. On average, global search responses are 15% longer. However, neither method fully answered these queries, each capturing only a subset of the *ideal answer*.

5.1.3 Comparison of GraphRAG and Baseline RAG

Compared to the GraphRAG pipelines, baseline RAG shows comparable B1 and B2 scores, indicating similar answer relevance. However, the baseline RAG answers are significantly shorter and narrower in scope. On average, GraphRAG responses are 2,273 characters long, while baseline RAG responses average 1,628 characters – a 39.64% difference. This brevity often corresponds to a reduced subset of the ideal answer information. For example, in response to the question "Which regions of Western Australia are referenced in the MRIWA reports?", baseline RAG returned just two regions, whereas the MDS global pipeline listed eight. Baseline RAG also includes slightly less relevant peripheral information (B3) and slightly more irrelevant information (B4) than the MDS and EMDS local pipelines. Baseline RAG's B5 score is roughly double that of the MDS and EMDS local pipelines, indicating a higher rate of hallucinations.

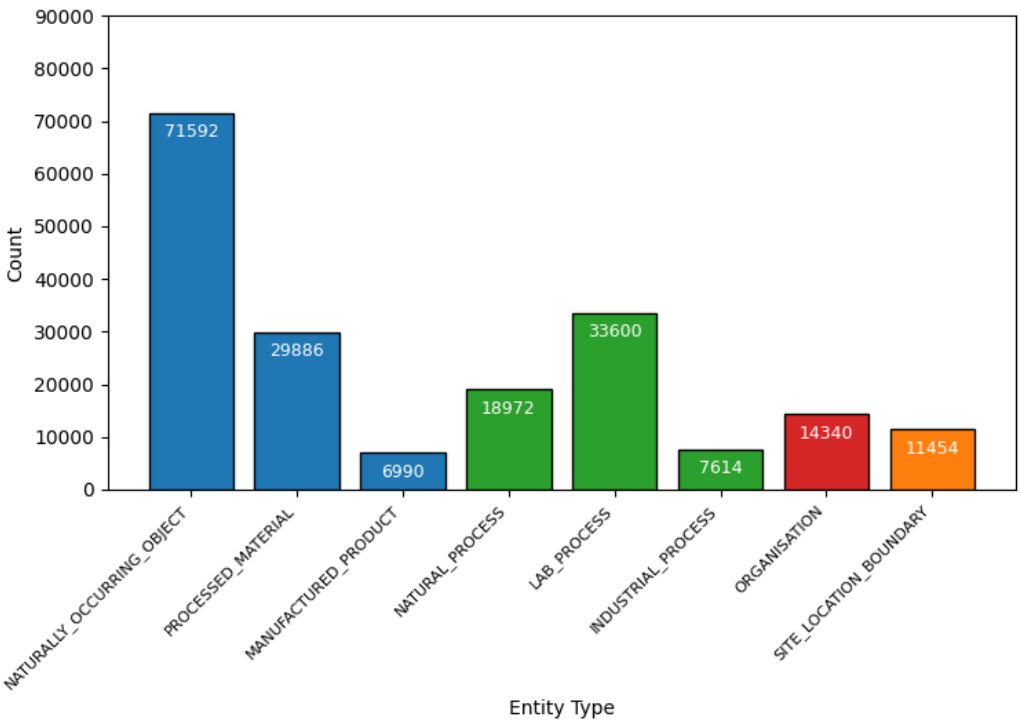


Figure 5 Distribution of EMDS entity types extracted by the EMDS GraphRAG pipeline.

5.1.4 Reflection on Performance Classification Scheme

The scores discussed in Table 7 were assigned by two people using the ideal answers provided by the MRIWA experts. The inter-rater agreement was high, with a Cohen's Kappa of 0.804 (out of 1) and percentage agreement of 91.56%. These values were calculated across all 675 binary ratings, comparing agreement on the presence or absence of each performance criterion. The agreement scores indicate strong consistency in how the classification scheme was applied. The main area of difference was a debate around the distinction between B3 (Medium Noise Robustness) and B4 (Low Noise Robustness). The second marker suggests that B3 (Medium Noise Robustness) and B4 (Low Noise Robustness) could be combined, as having information that does not directly answer the question, whether on-topic or not, is equally unhelpful to the user. However, the MRIWA team argued that providing a wide range of information could help users make new connections between different findings. Ultimately, we keep B3 (Medium Noise Robustness) and B4 (Low Noise Robustness) separate, but we note the value of each performance classification may depend on the user's goal and the specific question being asked.

We recognise a limitation of the B2 classification: it applies equally to answers containing nearly all ideal facts (e.g., 14/15) and those containing very few (e.g., 1/15). This lack of granularity means that poor responses can still receive a medium relevance score. The impact of this depends on the scale of the dataset and the length of the ideal answer. When ideal answers are short (e.g., 5 key facts), B2 may offer sufficient resolution. However, for more complex questions requiring many facts, the classification becomes less representative of actual answer quality, and an additional rating criterion may be needed to capture this distinction.

5.2 Performance of the MDS pipeline based on type of competency question

In this section we explore the impact of the type of question on the performance of the MDS GraphRAG pipeline by looking at some specific examples in more detail. We chose the MDS pipeline as it has strong performance compared to the other pipelines, as discussed in the preceding sections.

5.2.1 Simple Query: Identification of MRIWA Organisation in Reports

The question *"Identify which MRIWA reports reference MERIWA or MRIWA."* requires keyword search for the organisation "MRIWA/MERIWA". A correct response would list all 15 MRIWA report identifiers because each report contains "MRIWA" and/or "MERIWA".

We compare the MDS local search response with the MDS global search response for this question. With local search, the response correctly identifies three of the 15 reports available. The response also identifies an additional MRIWA report not provided in our sample but that is mentioned in one of our 15 reports. In contrast, global search identifies eight MRIWA reports correctly.

Generally, the ability to partially answer this question with no hallucinations is likely due to "MRIWA" and "MERIWA" being tagged with our MDS entity type `Organisation` during *Indexing*, allowing the related information to be retrieved during the *Query* stage. The global search identifies more correct reports because it incorporates a broader range of information during generation, using approximately 15 to 20 times more tokens than local search.

5.2.2 Set Query: Identification of Chemical Elements

The question, *"Which elements are considered in the MRIWA reports?"* requires the system to collate the set of all chemical elements across documents. A good response might identify, for each report, the elements of focus and the context in which they are mentioned. Elements frequently mentioned in our report subset include, but are not limited to, Gold, Carbon, Copper, Sulfur, Iron, Potassium, Phosphorus and Lithium.

We compare the MDS local search response with the MDS global search response for this question. With local search, the response identifies five chemical elements, which are a small subset of all the elements mentioned in the reports. Two include Rubidium and Manganese, neither of which rank in the top 10 most frequently mentioned elements in the documents. With only four MRIWA reports being explicitly mentioned in the response, the result demonstrates an inability of local search to effectively aggregate information about chemical elements across reports.

With global search, the response identifies seven elements, a small improvement over local search. However, no specific MRIWA reports are mentioned in the response. This limitation indicates a disadvantage of the global search technique, as it only utilises community reports when generating a response. By utilising these high-level summaries of the document repository, the names of the underlying information sources may not be available to the pipeline during final response generation. Additionally, the global search response notes several minerals in MRIWA's reports, which is irrelevant information.

5.2.3 Aggregation Query: Mathematical Aggregation of Report Citations

The question, *"What is the average number of references in each MRIWA report?"*, requires the system to both retrieve all references and calculate the average quantity of references in each report. None of our pipelines answered this question successfully.

Local search responses often contain an abundance of hallucinations regarding how many references are in each document. This search strategy first identifies relevant entities in the KG – for a question of this nature, it is not obvious what entities would be relevant given the schema. Nonetheless, if relevant entities are identified, then all 300-token chunks of reference lists need to be retrieved and pieced together – this does not occur and explains the failure to answer the question.

Global search responses typically contain statements to the effect of *"I am sorry but I am unable to answer this question given the provided data."* Given that global search only retrieves community reports of summarised information, full reference lists are not included in the final context window, resulting in the response failure.

5.2.4 Post-Processing-Heavy Query: Identification of Mining Extraction in Reports

The question *"Which MRIWA reports relate to mining extraction?"*, focuses on a specific process, rather than an object or organisation. A good response might identify a set of reports related to mining extraction and describe each report's contribution to the topic.

We compare the MDS local search response with the MDS global search response for this question. With local search, the response successfully identifies MRIWA Report 488, which is the only report in our subset that relates to mining extraction. However, the response contains four more MRIWA reports, which are not related to mining extraction.

With global search, the response does not successfully identify Report 488, instead identifying four unrelated MRIWA reports. The response also identifies an additional report "Laboratory Flotation Test", which is a community report, not a MRIWA report. This is a common issue across all global search responses. Since we use GraphRAG's default LLM prompts, the LLM is not aware of the surrounding context of the system and, therefore, fails to understand the difference between a MRIWA report and a generated community report.

These complex semantic queries, generally, are not answered well. The LLM, being a non-expert in minerals research, often fails to correctly distinguish between concepts such as "mineral processing", "mining extraction", and "exploration". This issue may extend to other technical domains, too. A solution may be to fine-tune the LLM on the text being retrieved to improve response quality.

5.2.5 Identification of Organisational Involvement in Projects

The questions *"Which MRIWA reports has Commonwealth Scientific Industrial Research Organisation been involved with as a researcher?"*, and *"Which MRIWA report author has been involved in more than one report/project?"* both require the system to identify involvement of entities in projects. The key difference is the former question names a specific organisation to be identified, whilst the latter question specifies "authors" to be identified.

We compare the MDS local search response to each question. The response to the CSIRO question successfully identifies three of the six MRIWA reports in which CSIRO is involved in as a researcher. In contrast, the response to the authors question is incorrect. Specifically, the five authors identified are valid authors, but none have been involved in multiple projects. This question is not answered successfully by any of the pipelines tested.

The explanation for this result may lie in the schema, as we specify `Organisation` as an entity type but not `Person` or `Author`. Further, the authors of reports are typically contained within document metadata which GraphRAG does not capture during the *Indexing* stage. This information may assist the LLM in answering metadata-related queries.

5.3 Phrase-by-Phrase Analysis

To investigate the depth of answers, we do a phrase-by-phrase analysis of the responses to the question, "Which MRIWA report is related to the East Kimberley region?". We classify phrases into five categories, as is displayed in Figure 6. These figures aim to give the reader a visual sense of the complexity of evaluating RAG responses. We want to see answers with cool colours (green and blue) and without hot colours (red and pink).

Figure 6 Minerals Domain Schema pipeline responses (left) and Auto-Generated Schema pipeline responses (right) to East Kimberley question. Each schema has two responses, Local Search (upper) and Global Search (lower).

In general, pipelines using a MRIWA-specific schema produce responses with a larger quantity of correct and relevant statements, compared to the auto-generated schema and schema-less pipelines. This finding is consistent with the results of our performance analysis.

The Local Search responses tend to be more concise and contain more correct and relevant statements (blue and green) than global search responses. This result is consistent with the fact that low-level targeted queries, such as *"Which MRIWA report is related to the East Kimberley region?"* are answered better with Local Search.

Default GraphRAG usually provides an unasked-for final paragraph starting with the words "overall", "in conclusion", "in summary", or "this report highlights". These paragraphs usually contain no new information. This behaviour may be altered by changing the default GraphRAG LLM prompts and parameters. For example, the `response_type` parameter of the `LocalSearch` and `GlobalSearch` classes allow a description of the response format to be specified – the parameter defaults to "Multiple Paragraphs".

None of the pipelines identify MRIWA Report 448 as being related to the East Kimberley region. While the region is not a primary focus of the report, mentions of the Kimberley appear in several data tables and references. However, these mentions do not occur near the term "Report 448" (or equivalent). Since the report is initially split into 300-token chunks, "Report 448" and "Kimberley" do not appear in the same chunk. As a result, the LLM fails to associate Report 448 with the Kimberley, leading to its omission in responses. Capturing document structure and metadata may help trace the source report of specific entities.

Among the pipelines that correctly identify MRIWA Report 459-484 as primarily focused on the East Kimberley, all fail to capture key details from the report's abstract when generating a summary. This result suggests GraphRAG's summarisation process during *Indexing* may lose important depth typically found in a technical report abstract.

5.4 Limitations

GPT-4o-Mini is a non-deterministic LLM. Therefore, repetitions of the experiment are likely to yield different results [33]. The impact of non-determinism may be prevalent during entity and relation extraction. Given the same KG schema and set of text chunks, the entities and relations identified may not be identical between trials, resulting in variation of the final KG, impacting GraphRAG responses. As Edge et al. [7] note, this issue may be solved by conducting multiple gleanings, where the process of entity extraction is repeated to detect additional entities missed by the LLM.

A risk of using OpenAI's GPT is its closed-source nature, which can lead to vendor lock-in and raise privacy and confidentiality concerns. Additionally, changes to the underlying model may impact repeatability. While using an open-source LLM like Llama can provide more control, it does not eliminate the possibility of updates affecting results and may introduce performance differences.

Finally, our experiments utilise a subset of 15 MRIWA reports and do not include tables or figures from the reports. These may provide valuable information for the construction of a more comprehensive KG. However, all the MRIWA reports are publicly available, and for the 15 reports we have used, we provide the questions, our ideal answers, schemas, and details on how we performed each evaluation. This will allow others to replicate and extend our work.

6 Conclusions

This paper explores the application of Microsoft's GraphRAG pipeline to technical reports about geology, chemistry and mineral processing published by MRIWA. The project assesses how 1) domain-specific knowledge graph schema, and 2) the selection of local or global GraphRAG search options, impact the quality of GraphRAG responses to MRIWA-defined queries. Pipeline performance evaluation is based on domain-expert curated answers, and original scoring systems considering relevance, robustness and hallucinations.

GraphRAG with a simple five-class minerals-domain schema (MDS) extracts approximately 10% more entities from MRIWA's reports compared to pipelines using a complex eight-class minerals-domain schema (EMDS), an auto-generated schema (AGS), and a schema-less (SL)

GraphRAG. While MDS responses contain more relevant peripheral information, both MDS and EMDS produce the most factually correct answers with the fewest hallucinations. These results suggest that a domain-specific schema can assist GraphRAG with entity extraction on technical texts, leading to better response quality.

References

1 Bradley P Allen, Lise Stork, and Paul Groth. Knowledge engineering using large language models. *arXiv preprint arXiv:2310.00637*, 2023. doi:10.48550/arXiv.2310.00637.

2 Tyler Bikaun, Michael Stewart, and Wei Liu. Quickgraph: A rapid annotation tool for knowledge graph extraction from technical text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2022. doi:10.18653/v1/2022.acl-demo.27.

3 Kurt Bollacker, Robert Cook, and Patrick Tufts. Freebase: A shared database of structured general human knowledge. In *AAAI*, volume 7, pages 1962–1963, 2007. URL: http://www.aaai.org/Library/AAAI/2007/aaai07-355.php.

4 Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024. doi:10.1145/3641289.

5 Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729, 2024. doi:10.1145/3626772.3657834.

6 Armita Davarpanah, Hassan A Babaie, and W Crawford Elliott. Knowledge-based query system for the critical minerals. *Applied Computing and Geosciences*, 22:100167, 2024. doi:10.1016/j.acags.2024.100167.

7 Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A Graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi:10.48550/arXiv.2404.16130.

8 Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of Retrieval Augmented Generation. *arXiv preprint arXiv:2309.15217*, 2023. doi:10.48550/arXiv.2309.15217.

9 Mathieu Fenniak and PyPDF2 Contributors. PyPDF2: A pure-Python PDF library, version 3.x. GitHub, 2022. Accessed: 17-Jul-2025. URL: https://github.com/py-pdf/pypdf.

10 E. Friis. Langchain. GitHub, 2024. Accessed: 09-May-2024. URL: https://github.com/langchain-ai.

11 Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-Augmented Generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. doi:10.48550/arXiv.2312.10997.

12 Luan Fonseca Garcia, Mara Abel, Michel Perrin, and Renata dos Santos Alvarenga. The geocore ontology: a core ontology for general use in geology. *Computers & Geosciences*, 135:104387, 2020. doi:10.1016/j.cageo.2019.104387.

13 Tania Hidalgo, Laura Kuhar, Andreas Beinlich, and Andrew Putnis. Understanding fluid–rock interactions and lixiviant/oxidant behaviour for the in-situ recovery of metals from deep ore bodies. MRIWA Report M0488, Minerals Research Institute of Western Australia, 2020. Report No. 488.

14 International Organization for Standardization. ISO/IEC 21838-1:2021 Information technology — Top-level ontologies (TLO) - Part 1: Requirements. International standard, International Organization for Standardization, 2021.

15 Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, et al. Think and retrieval: A hypothesis knowledge graph enhanced medical large language models. *arXiv preprint arXiv:2312.15883*, 2023. doi:10.48550/arXiv.2312.15883.

16 Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*, 2023. doi:10.48550/arXiv.2305.18846.

17 Jerry Liu. Llamaindex. GitHub, 2022. Accessed: 09-May-2024. URL: https://github.com/jerryjliu/llama_index.

18 Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*, 2023. doi:10.48550/arXiv.2310.01061.

19 Microsoft. GraphRAG. https://microsoft.github.io/graphrag/. Accessed: 2024-09-30.

20 Microsoft. GraphRAG. https://github.com/microsoft/graphrag/tree/v0.1.1, 2024. Accessed: 2024-09-30.

21 Microsoft. GraphRAG. https://github.com/microsoft/graphrag/tree/v2.3.0, 2025. Accessed: 2025-07-25.

22 Mineral Research Institute of Western Australia. Project Portfolio. https://www.mriwa.wa.gov.au/research-projects/project-portfolio/, 2024. [Online; accessed 9-May-2024].

23 Pin Ni, Ramin Okhrati, Steven Guan, and Victor Chang. Knowledge graph and deep learning-based text-to-graphql model for intelligent medical consultation chatbot. *Information Systems*

Frontiers, 26(1):137–156, 2024. doi:10.1007/s10796-022-10295-0.

24 Daniela Oliveira, Ratnesh Sahay, and Mathieu d'Aquin. Leveraging ontologies for knowledge graph schemas. In *KGB@ ESWC*, pages 24–36, 2019. URL: https://ceur-ws.org/Vol-2489/paper3.pdf.

25 Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36:3580–3599, 2024. doi:10.1109/tkde.2024.3352100.

26 Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph Retrieval-Augmented Generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024. doi:10.48550/arXiv.2408.08921.

27 Tyler Thomas Procko and Omar Ochoa. Graph Retrieval-Augmented Generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169. IEEE, 2024.

28 David Robinson. Improving solvent extraction technology. MRIWA Report M0401, Minerals Research Institute of Western Australia, 2010. Report No. 289.

29 Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for Retrieval-Augmented Generation systems. *arXiv preprint arXiv:2311.09476*, 2023. doi:10.48550/arXiv.2311.09476.

30 Henri Scaffidi, Melinda Hodkiewicz, Caitlin Woods, and Nicole Roocke. nlp-tlp/GraphRAG-on-Minerals-Domain. InteractiveResource (visited on 2025-10-07). URL: https://github.com/nlp-tlp/GraphRAG-on-Minerals-Domain, doi:10.4230/artifacts.24886.

31 Juan Sequeda, Dean Allemang, and Bryon Jacob. A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise SQL databases. In *Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, pages 1–12, 2024. doi:10.1145/3661304.3661901.

32 Tanya Serry, Tonya Stebbins, Andrew Martchenko, Natalie Araujo, and Brigid McCarthy. Improving access to COVID-19 information by ensuring the readability of government websites. *Health Promotion Journal of Australia*, 34(2):595–602, 2023. doi:10.1002/hpja.610.

33 Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024. doi:10.48550/arXiv.2407.10457.

34 Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019. doi:10.1038/s41598-019-41695-z.

35 Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on World Wide Web*, pages 1063–1064, 2012. doi:10.1145/2187980.2188242.

36 Chengbin Wang, Xiaogang Ma, Jianguo Chen, and Jingwen Chen. Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences*, 112:112–120, 2018. doi:10.1016/j.cageo.2017.12.007.

37 Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*, 2023. doi:10.48550/arXiv.2308.11761.

38 M. Wells, M. Aylmore, and B. McInnes. The geology, mineralogy and geometallurgy of EV materials deposits in Western Australia. MRIWA Report M532, Geological Survey of Western Australia, 2022. Report No. 228.

39 Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-Augmented Generation with knowledge graphs for customer service question answering. *arXiv preprint arXiv:2404.17723*, 2024. doi:10.1145/3626772.3661370.

40 Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. CRAG-comprehensive RAG benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490, 2024. doi:10.48550/arXiv.2406.04744.

41 Huaxuan Zhao, Yueling Pan, and Feng Yang. Research on information extraction of technical documents and construction of domain knowledge graph. *IEEE Access*, 8:168087–168098, 2020. doi:10.1109/access.2020.3024070.

42 Lingfeng Zhong, Jia Wu, Qian Li, Hao Peng, and Xindong Wu. A comprehensive survey on automatic knowledge graph construction. *ACM Computing Surveys*, 56(4):1–62, 2023. doi:10.1145/3618295.

43 Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, et al. Rageval: Scenario specific RAG evaluation dataset generation framework. *arXiv preprint arXiv:2408.01262*, 2024. doi:10.48550/arXiv.2408.01262.