Journal of Data Intelligence, Vol. 2, No. 3 (2021) 326–347

DISCOVER RELATIONS IN THE INDUSTRY 4.0 STANDARDS VIA UNSUPERVISED LEARNING ON KNOWLEDGE GRAPH EMBEDDINGS

ARIAM RIVAS
Research Center L3S, Leibniz Univ. of Hannover, Appelstraße 9a
Hannover, 30167, Germany
ariam.rivas@tib.eu

IRLÁN GRANGEL-GONZÁLEZ
Robert Bosch Corporate Research GmbH, Robert-Bosch-Campus 1
Stuttgart, 71272, Germany
irlan.grangelgonzalez@de.bosch.com

DIEGO COLLARANA
University of Bonn & Fraunhofer IAIS, Zwickauerstraße 46
Dresden, 01069, Germany
diego.collarana.vargas@iais.fraunhofer.de

JENS LEHMANN
University of Bonn & Fraunhofer IAIS, Zwickauerstraße 46
Dresden, 01069, Germany
jens.lehmann@iais.fraunhofer.de

MARIA-ESTHER VIDAL
TIB Leibniz Information Centre for Science and Technology, Welfengarten 1B
Hannover, 30167, Germany
maria.vidal@tib.eu

Industry 4.0 (I4.0) standards and standardization frameworks *provide a unified way* to describe smart factories. Standards specify the main components, systems, and processes inside a smart factory and the interaction among all of them. Furthermore, standardization frameworks classify standards according to their functions into layers and dimensions. Albeit informative, frameworks can categorize similar standards differently. As a result, interoperability conflicts are generated whenever smart factories are described with miss-classified standards. Approaches like ontologies and knowledge graphs enable the integration of standards and frameworks in a structured way. They also encode the meaning of the standards, known relations among them, as well as their classification according to existing frameworks. This structured modeling of the I4.0 landscape using a graph data model provides the basis for graph-based analytical methods to uncover alignments among standards. This paper contributes to analyzing the relatedness among standards and frameworks; it presents an unsupervised approach for discovering links among standards. The proposed method resorts to knowledge graph embeddings to determine relatedness among standards-based on similarity metrics. The proposed method is agnostic to the technique followed to create the embeddings and to the similarity measure. Building on the similarity values, community detection algorithms can automatically create communities of highly similar standards. Our approach follows the *homophily principle*, and assumes that related standards are together in a community. Thus, alignments across standards are predicted and interoperability issues across them are solved. We empirically evaluate our approach on a knowledge graph of 249 I4.0 standards using the Trans* family of embedding models for knowledge graph entities. Our results are promising and suggest that relations among standards can be detected accurately.

Keywords: Industry 4.0, Standard, Knowledge Graph, Embedding, Unsupervised Learning

1. Introduction

The international community recognizes Industry 4.0 (I4.0) as the fourth industrial revolution. The main objective of I4.0 is the creation of *smart factories* by combining the Internet of Things (IoT), Internet of Services (IoS), and Cyber-Physical Systems (CPS). In smart factories, humans, machines, materials, and CPS cooperate intelligently to produce individualized products. This cooperation requires effective communication and the resolution of interoperability issues generated whenever the same products are described with different standards. Different industrial communities have defined standardization frameworks aligning standards according to their features and expressiveness. Relevant examples are the Reference Architecture for Industry 4.0 (RAMI4.0) [1] or the Industrial Internet Connectivity Framework (IICF) in the US [2]. Despite the capacity to categorize existing standards, standardization frameworks may present divergent interpretations of the same standard. Mismatches among standard classifications generate semantic interoperability conflicts that negatively impact communication effectiveness in smart factories.

Database and Semantic web communities have extensively studied the problem of data integration [3, 4, 5], and various approaches have been proposed to support data-driven pipelines to transform industrial data into actionable knowledge in smart factories [6, 7, 8]. Ontology-based approaches have also contributed to create a shared understanding of the domain [9], and specifically Kovalenko and Euzenat [4] have equipped data integration with diverse methods for ontology alignment. Furthermore, Lin *et al.* [10] identify interoperability conflicts across domain specific standards (e.g., RAMI4.0 model and the IICF architecture), while works by Grangel-Gonzalez *et al.* [11, 12, 13] show the relevant role that Descriptive Logic, Datalog, and Probabilistic Soft Logic play in liaising I4.0 standards. Certainly, the extensive literature in data integration provides the foundations for enabling the semantic description and alignment of "similar" things in a smart factory. Nevertheless, finding alignments across I4.0 requires the encoding of domain specific knowledge represented in standards of diverse nature and standardization frameworks defined with different industrial goals. We rely on state-of-the-art knowledge representation and discovery approaches to embed meaningful

associations and features of the I4.0 landscape, to enable interoperability.

Problem Statement and Objectives. We address the problem of determining relatedness across I4.0 standards described in terms of their main features and standardization frameworks. Our goal is uncovering alignments among related standards, i.e., standards that define the same type of components of a smart factory. Moreover, we aim at providing precise classification of the standards and contributing to a more precise categorizations in the standardization frameworks.

Proposed Solution. We propose a knowledge-driven approach able to integrate standards and standardization frameworks into a knowledge graph. Knowledge graphs are data structures that represent data, knowledge, and actionable insights using a graph data model. They naturally model entities and their relationships, mono- and multi-valued attributes, and the neighborhoods of an entity. The features of the standards represented in a knowledge graph are exploited to build latent representations in a low-dimensional space, i.e., embeddings. Values of similarity metrics between embeddings are used in conjunction with state-of-the-art community detection algorithms to identify patterns among standards. The *homophily* prediction principle is performed in each community to discover new links between standards and frameworks. Our approach is general and agnostic with the technique to create the embeddings, the similarity measure, and the community detection methods. However, as a proof of concept, the Trans* family of embedding models is utilized in our evaluations. Moreover, different similarity measures are evaluated to determining relatedness among standards based on the embeddings. State-of-the-art community detection methods (e.g., SemEP [14], Metis[15] and KMeans [16]) are applied to group together similar standards. We assess the performance of the proposed methods in a knowledge graph composed of 249 I4.0 standards connected by 736 relations. These relationships have been extracted from the literature. The experiments are executed following various configurations and baselines. The observed results are promising and demonstrate the benefits of exploiting knowledge graphs for the computation of alignments across standards. These outcomes provide evidence of the accuracy of the uncovered patterns and the discovered relations.

Contributions. This paper is an extension of our previous work [17] where we determined relatedness among standards and analyzed their properties to detect unknown relations. This work is built on the results reported on our previous paper and presents the following contributions:

1. A more detailed analysis of the state of the art.
2. A formalization of problem of finding relations among I4.0 standards. It presents *I4.0*$\mathcal{RD}$, a knowledge-driven approach to unveil these relations. *I4.0*$\mathcal{RD}$ exploits the semantic description encoded in a knowledge graph via the creation of embeddings, to identify then communities of standards that should be related.
3. An extensive evaluation of *I4.0*$\mathcal{RD}$ in different embeddings learning models, similarity measures, and community detection algorithms. The evaluation material is available at `https://github.com/i40-Tools/I40KG-Embeddings`.

The rest of the paper is organized as follows: Section 2 summarizes the the state of the art. Section 3 illustrates the interoperability problem presented in this paper. Section 4 presents the proposed approach, while the architecture of the proposed solution is explained in Section 5. Results of the empirical evaluation of our methods are reported in Section 6. Finally, we close with the conclusion and future work in Section 7.

2. Related Work

There have been a great deal of research in recent years investigating key aspects of discovering communities of standards. Furthermore, many approaches are proposed to corroborate and extend the knowledge of the standardization frameworks and resolving semantic interoperability issues.

2.1. *Solving Interoperability in I4.0*

Zeid *et al.* [18] study different approaches to achieve interoperability of different standardization frameworks. In this work, the current landscape for smart manufacturing is described by highlighting the existing standardization frameworks in different regions of the globe. Lin *et al.* [10] present similarities and differences between the RAMI4.0 model and the IIRA architecture. Based on the study of these similarities and differences, the authors proposed a functional alignment among layers in RAMI4.0 with the functional domains and crosscutting functions in IIRA. Monteiro *et al.* [19] further report on the comparison of the RAMI4.0 and IIRA frameworks. In this work, a cooperation model is presented to align both standardization frameworks. Furthermore, mappings between RAMI4.0 IT Layers and the IIRA functional domain are established. Moreover, the IIRA and RAMI4.0 frameworks are compared based on different features, e.g., country of origin, source organization, basic characteristics, application scope, and structure. It further details where correspondences exist between the IIRA viewpoints and RAMI4.0 layers. In [20], Darmois *et al.* present the main contributions to the analysis of IoT standardization. This work has defined knowledge areas used for the classification of standards and identifies the standardization gaps. The purpose is to support the interoperability in complex IoT systems and provide guidelines that contribute to the semantic interoperability approaches. Aligning standardization frameworks is useful to solve the interoperability problems but not all standards are classified by layers in the standardization frameworks. However, these approaches aim to solve interoperability problems by mapping the different frameworks without creating a common vocabulary that semantically represents the standards. In this article, we propose an approach to solve interoperability problems among I4.0 standards by discovering unknown relationships.

2.2. *Ontology-based Approaches in I4.0*

Ontology-based approaches have contributed to creating a shared understanding of the I4.0 domain. Lelli *et al.* [9] propose the reuse of existing ontologies as one of the main principles in ontology design. For this purpose, they make use of Linked Open Vocabulary (LOV) and collect 22 ontologies related to IoT. They state that project developers in the IoT community do not reuse existing works, damaging the attempt to define a shared understanding of smart interoperability. Kovalenko and Euzenat [4] have equipped data integration with diverse methods for ontology alignment. They examine the problems of ontological correspondence in the context of engineering knowledge integration. Kovalenko and Euzenat present technologies for defining mappings between ontologies to support data integration. Finally, they illustrate how mappings can be generated from definitions in the Expressive and Declarative Ontology Alignment Language (EDOAL). These approaches are limited to representing the existing characteristics of the knowledge domain in ontologies, which is useful because it enables data integration in Industry 4.0. However, there are standards that are not classified in any standardization framework and this limits the solution of the interoperability problem. In this work, we employ the Standard Ontology, (*STO*) for representing the main properties of standards and standardization frameworks, as well as relationships among them [11].

2.3. *Knowledge Graphs and Semantic Data Integration*

Sebastian *et al.* [6] propose a semantically annotated knowledge graph for Industry 4.0 related standards, norms, and frameworks. The I4.0 knowledge graph helps to overcome Industry 4.0 challenges that require a comprehensive knowledge of the different standards. Furthermore, the I4.0 knowledge graph considers the semantics and relations between standards and the standardization framework. Garofalo *et al.* [21] outline Knowledge Graph Embeddings for I4.0 use cases. Existing techniques for generating embeddings on top of KG are examined. Further, the analysis of how these techniques can be applied to the I4.0 domain is described; specifically, it identifies the predictive maintenance, quality control, and context-aware robots as the most promising areas to apply the combination of KGs with embeddings. These approaches mentioned above support data-driven pipelines to transform industrial data into actionable knowledge in smart factories. Galinski [22] examines the problem of semantic data integration and interoperability between standards. This work emphasizes the need for metadata, data models, and metamodels for standards. It also presents an interesting description of which data to consider when describing a standard. Hodges *et al.* [7] propose an approach for semantic integration of standards to achieve interoperability between them by means of ontologies; relevant standards and well-known ontologies to represent standards are also identified. Albeit representing domain-specific knowledge, the approaches mentioned above cannot solve interoperability issues across I4.0 standards. We overcome this limitation by exploiting embeddings over a knowledge graph of I4.0 standards to predict relatedness among standards.

Summary of the Related Work. The approaches presented in this section describe and characterize existing knowledge in the I4.0 domain. However, in our view, two directions need to be considered to enhance the knowledge in the domain; 1) the use of a KG based approach to encoding the semantics; and 2) the use of machine learning techniques to discover and predict new communities of standards based on their relations. Our goal is uncovering alignments among related standards. Nevertheless, finding alignments across I4.0 requires the encoding of domain-specific knowledge represented in standards of diverse nature.

3. Motivating Example

Existing efforts to achieve interoperability in I4.0, mainly focus on the definition of standardization frameworks. A standardization framework defines different layers to group related I4.0 standards based on their functions and main characteristics. Typically, classifying existing standards in a certain layer is not a trivial task and it is biased by the point of view of the community that developed the framework. RAMI4.0 and IICF are exemplar frameworks, the former is developed in Germany while the latter in the US; they meet specific I4.0 requirements of certain locations around the globe. RAMI4.0 classifies the standards OPC UA and MQTT into the Communication layer, stating that both standards are similar. Contrary, IICF presents OPC UA and MQTT at distinct layers, i.e., the framework and the transport layers, respectively. Furthermore, independently of the classification of the standards made by standardization frameworks, standards have relations based on their functions. Therefore, IEC 61580 is an international standard defining communication protocols for intelligent electronic devices and ISO 15531 is an standard for industrial automation systems. Both standards are not not classified at all. Figure 1 depicts these relations across the frameworks RAMI4.0 and IICF, and the standards; it illustrates interoperability issues in the I4.0 landscape.

Existing data integration approaches rely on the description of the characteristics of entities to

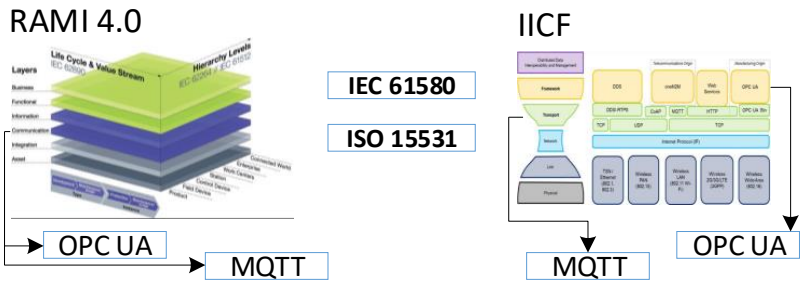


Fig. 1. **Motivating Example**. The RAMI4.0 and IICF standardization frameworks are developed for diverse industrial goals; they classify standards in layers according to their functions, e.g., OPC UA and MQTT under the communication layer in RAMI4.0, and OPC UA and MQTT in the framework and transport layers in IICF, respectively. Further, some standards, e.g., IEC 61580 and ISO 15531, are not classified yet.

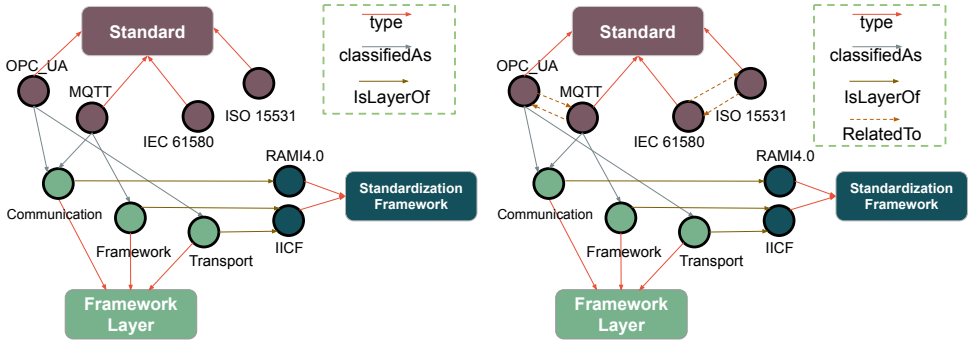
solve interoperability by discovering alignments among them. Specifically, in the context of I4.0, semantic-based approaches have been proposed to represent standards, known relations among them, as well as their classification according to existing frameworks [10, 23, 24, 25]. Despite informative, the structured modeling of the I4.0 landscape only provides the foundations for detecting interoperability issues. We propose *I4.0*$\mathcal{RD}$, an approach capable of discovering relation over I4.0 knowledge graphs to identify unknown relations among standards. Our proposed methods exploit relations represented in an I4.0 knowledge graph to compute the similarity of the modeled standards. Then, an unsupervised graph partitioning method determines the communities of standards that are similar. *I4.0*$\mathcal{RD}$ explores communities to identify possible relations of standards, enhancing, thus, interoperability.

4. Problem Definition and Proposed Solution

We tackle the problem of unveiling relations between I4.0 standards. Relations among standards and standardization frameworks (e.g., in Figure 2 (a)) are represented in a knowledge graph named I4.0KG. Nodes in a I4.0KG correspond to standards and frameworks; edges represent relations among standards, as well as the standards grouped in a framework layer. An I4.0KG is defined as follows:

Given sets V_e and V_t of entities and types, respectively, a set E of labelled edges representing relations, and a set L of labels. An I.40KG is defined as $\mathcal{G} = (V_e \cup V_t, E, L)$:

- The types Standard, Frameworks, and Framework Layer belong to V_t.
- I4.0 standards, frameworks, and layers are represented as instances of V_e.
- The types of the entities in V_e are represented as edges in E that belong to $V_e \times V_t$.
- Edges in E that belong to $V_e \times V_e$ represent relations between standards and their classifications into layers according to a framework.
- Properties *relatedTo*, *Type*, *classifiedAs*, *isLayerOf* correspond to labels in L that represent the relations between standards, their type, their classification into layers, and the layers of a framework, respectively.



(a) Actual I4.0 KG (b) Ideal I4.0 KG

Fig. 2. **Example of I4.0KGs**. (a) shows known relationships among standards to Framework Layer and Standardization Framework. (b) depicts all the ideal relationships between the standards expressed with the property *relatedTo*. Standards OPC UA and MQTT are related, as well as the standards IEC 61968 and IEC 61400. Our aim is discovering relations *relatedTo* in (b).

4.1. *Problem Statement*

Let $\mathcal{G}' = (V_e \cup V_t, E', L)$ and $\mathcal{G} = (V_e \cup V_t, E, L)$ be two I4.0 knowledge graphs. $\mathcal{G}'$ is an *ideal* knowledge graph that contains all the *existing relations* between standard entities and frameworks in V_e, i.e., an oracle that knows whether two standard entities are related or not, and to which layer they should belong; Figure 2 (b) illustrates a portion of an ideal I4.0KG, where the relations between standards are explicitly represented. $\mathcal{G} = (V_e \cup V_t, E, L)$ is an *actual* I4.0KG, which only contains a portion of the relations represented in $\mathcal{G}'$, i.e., $E \subseteq E'$; it represents those relations that are known and is not necessarily complete. Let $\Delta(E', E) = E' - E$ be the set of relations existing in the ideal knowledge graph $\mathcal{G}'$ that are not represented in $\mathcal{G}$. Let $\mathcal{G}_{\text{comp}}$=$(V_e \cup V_t, E_{\text{comp}}, L)$ be a *complete* knowledge graph, which includes a relation for each possible combination of elements in V_e and labels in L, i.e., $E \subseteq E' \subseteq E_{\text{comp}}$. Given a relation $e \in \Delta(E_{\text{comp}}, E)$, the problem of discovering relations consists of determining whether $e \in E'$, i.e., if a relation represented by an edge r=$(e_i\ l\ e_j)$ corresponds to an existing relation in the ideal knowledge graph $\mathcal{G}'$. Specifically, we focus on the problem of discovering *relations* between standards in $\mathcal{G} = (V_e \cup V_t, E, L)$. We are interested in finding the maximal set of relationships or edges E_a that belong to the ideal I4.0KG, i.e., find a set E_a that corresponds to a solution of the following optimization problem:

$$\underset{E_a \subseteq E_{comp}}{\operatorname{argmax}} |E_a \cap E'|$$

Considering the knowledge graphs depicted in Figures 2 (a) and (b), the problem addressed in this work corresponds to the identification of edges in the ideal knowledge graph that correspond to unknown relations between standards.

4.2. *Proposed Solution*

We propose a relation discovery method over I4.0KGs to identify unknown relations among standards. Our proposed method exploits relations represented in an I4.0KG to compute similarity values between the modeled standards. Further, an unsupervised graph partitioning method determine the parts of the I4.0KG or communities of standards that are similar. Then, the *homophily* prediction

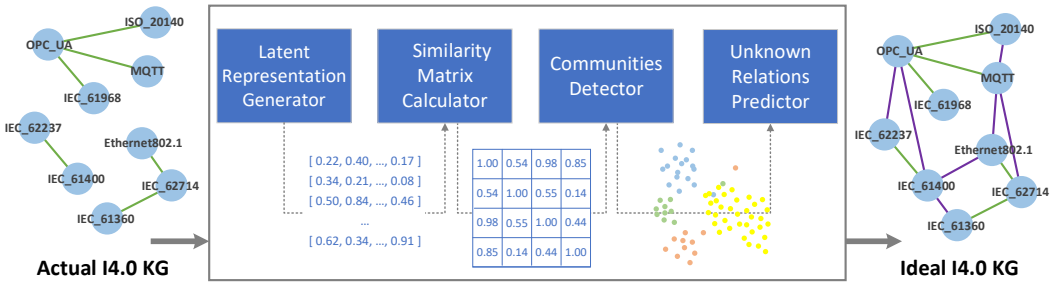


Fig. 3. **The *I4.0RD* Architecture.** *I4.0RD* receives the actual I4.0 KG and outputs an extended version of the I4.0KG including novel relations. Embeddings for each standard are created using the Trans* family of models, and similarity values between embeddings are computed; these values are used to partition standards into communities. Finally, the homophily prediction principle is applied to each community to discover unknown relations. A KG closer to the ideal I4.0 KG is generated.

principle is applied in a way that similar standards in a community are considered to be related.

5. The *I4.0RD* Architecture

Figure 3 presents *I4.0RD*, a pipeline that implements the proposed approach. *I4.0RD* receives an I4.0KG $\mathcal{G}$, and returns an I4.0KG $\mathcal{G}'$ that corresponds to a solution of the problem of discovering relations between standards. First, in order to compute the values of similarity between the entities an I4.0KG, *I4.0RD* learns a latent representation of the standards in a high-dimensional space. Our approach resorts to the Trans* family of models to compute the embeddings of the standards. Then, a distance metric for vector spaces is applied to compute the values of similarity between standards. Next, community detection algorithms are applied to identify communities of related standards. METIS [15], KMeans [16], and SemEP [14] are methods included in the pipeline to produce different communities of standards. Finally, *I4.0RD* applies the *homophily* principle to each community to predict relations or alignments among standards.

5.1. *Properties of the relation relatedTo between Standards*

The relation *relatedTo* was extracted from the literature and represents a relation that connects two standards. Beside *relatedTo* is an equivalent relation that satisfies three properties, i.e., the relation is reflexive, symmetric, and transitive. They are defined as follow:

- Reflexive: $\forall e_i \in V_e (e_i, \textit{relatedTo}, e_i)$
- Symmetric: $\forall e_i, e_j \in V_e ((e_i, \textit{relatedTo}, e_j) \Leftrightarrow (e_j, \textit{relatedTo}, e_i))$
- Transitive:

$$\forall e_i, e_j, e_k \in V_e : ((e_i, \textit{relatedTo}, e_j) \wedge (e_j, \textit{relatedTo}, e_k)) \Rightarrow (e_i, \textit{relatedTo}, e_k)$$

An example of the transitivity property of *relatedTo* is presented with the following three standards: *IEC 61310 P3 E2; IEC 61310 P1 E2; IEC 61310 P2 E2*. From the literature the next relations are known: *(IEC 61310 P3 E2, relatedTo, IEC 61310 P1 E2)* $\wedge$ *(IE 61310 P1 E2, relatedTo, IEC 61310 P2 E2)* and that implies: *(IEC 61310 P3 E2, relatedTo, IEC 61310 P2 E2)*. Since the property

relatedTo between standards is an equivalent relation, the transitive closure of the relations is materialized in I4.0KG. Thus, we can capture implicit relations between I4.0 standards. Figure 4 shows the relation *relatedTo* before it materialized in I4.0KG (cf. Figure 4a) and after being materialized Figure 4b. Figure 4b illustrates how after the transitive closure of the relations, the I4.0 standards knowledge graph is more connected. The graphs were plotted using Cytoscape[a].

Learning Latent Representations of Standards. *I4.0RD* utilizes the Trans* family of models to compute latent representations, e.g., vectors, of entities and relations in an I4.0 knowledge graph. In particular, *I4.0RD* utilizes TransE, TransD, TransH, and TransR. These models differ on the representation of the embeddings for the entities and relations (Wang et al. [26]). Suppose e_i, e_j, and p, denote the vectorial representation of two entities related by the labeled edge p in an I4.0 knowledge graph. Furthermore, $\|x\|_2$ represents the Euclidean norm.

TransE, TransH, and TranR represent the entity embeddings as $(e_i, e_j \in \mathbb{R}^d)$,while TransD characterizes the entity embeddings as: $(e_i, w_{e_i} \in \mathbb{R}^d - e_i, w_{e_j} \in \mathbb{R}^d)$. As a consequence of different embedding representations, the scoring function also varies. For example, TransE is defined in terms of the score function $\|e_i + p - e_j\|_2^2$, while $\|M_p e_i + p - M_p e_j\|_2^2$ defines TransR[b] Furthermore, TransH score function corresponds to $\|e_{i\perp} + d_p - e_{j\perp}\|_2^2$, where the variables $e_{i\perp}$ and $e_{j\perp}$ denote a projection to the hyperplane w_p of the labeled relation p, and d_p is the vector of a relation-specific translation in the hyperplane w_p. To learn the embeddings, *I4.0RD* resorts to the PyKeen (Python KnowlEdge EmbeddiNgs) framework [27]. As hyperparameters for the models of the Trans* family, we use the ones specified in the original papers of the models. The hyperparameters include embedding dimension (set to 50), number of epochs (set to 500), batch size (set to 64), seed (set to 0), learning rate (set to 0.01), scoring function (set to 1 for TransE, and 2 for the rest), margin loss (set to 1 for TransE and 0.05 for the rest). All the configuration classes and hyperparameters are open in GitHub [c].

Computing Similarity Values Between Standards. Once the algorithm–Trans* family–that computes the embeddings reaches a termination condition, e.g., the maximum number of epochs, the I4.0KG embeddings are learned. As the next step, *I4.0RD* calculates a *similarity symmetric matrix* between the embeddings that represent the I4.0 standards. Any distance metric for vector spaces can be utilized to calculate this value. However, as a proof of concept, *I4.0RD* applies the Cosine Similarity and the Inverse Euclidean Distance. Let u be an embedding of the Standard-A and v an embedding of the Standard-B, the similarity score, between both standards, is defined by Cosine Similarity[d] as follows:

$$cosine(u, v) = \frac{u.v}{||u||_2||v||_2}$$

The Inverse Euclidean Distance[e] between the vectors u and v, is defined as follows:

$$d(u, v) = 1 - ||u - v||_2$$

After building the *similarity symmetric matrix*, *I4.0RD* applies a threshold to restrict the similarity values. *I4.0RD* relies on percentiles to calculate the value of such a threshold. Further, *I4.0RD*

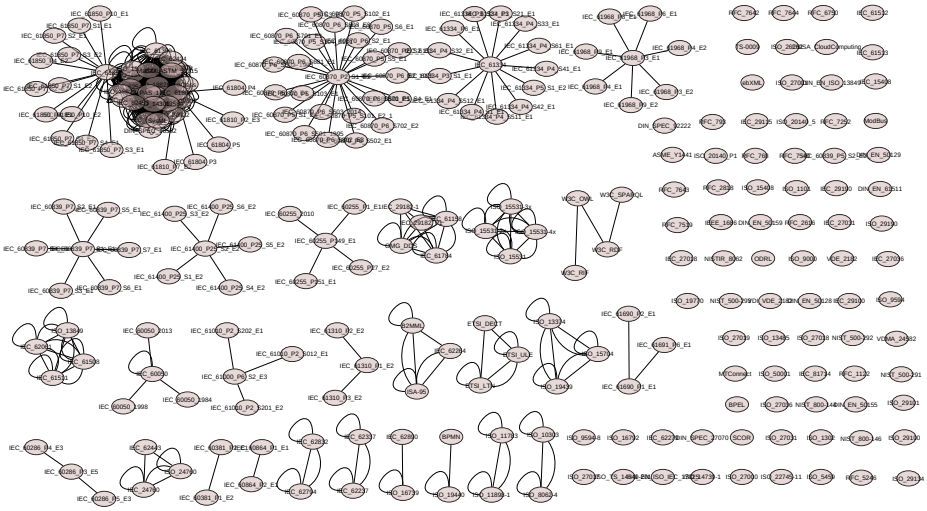
[a] `https://cytoscape.org/`

[b] M_p corresponds to a projection matrix $M_p \in \mathbb{R}^{dxk}$ that projects entities from the entity space to the relation space; further $p \in \mathbb{R}^k$.

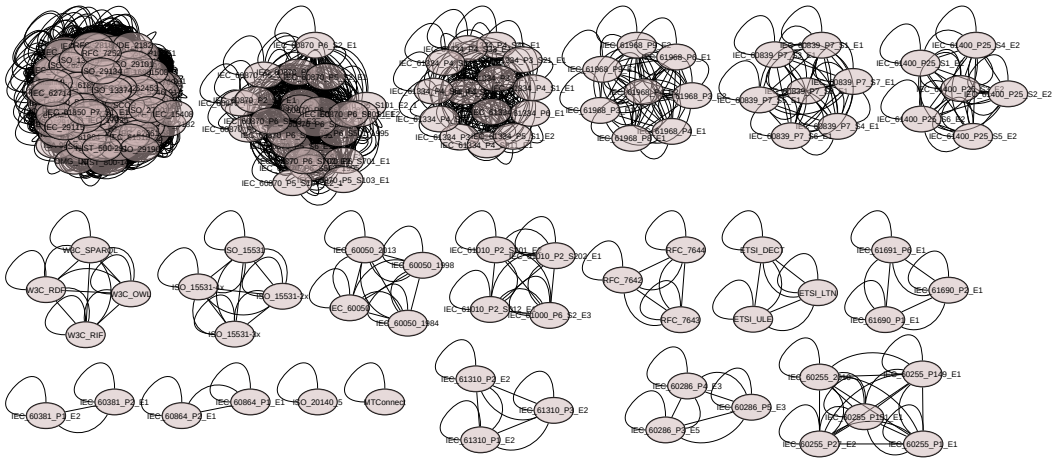
[c] `https://github.com/i40-Tools/I4.0KG-Embeddings`

[d] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cosine.html`

[e] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.euclidean.html`



(a) Explicit relations between I4.0 KG standards by the property *relatedTo*



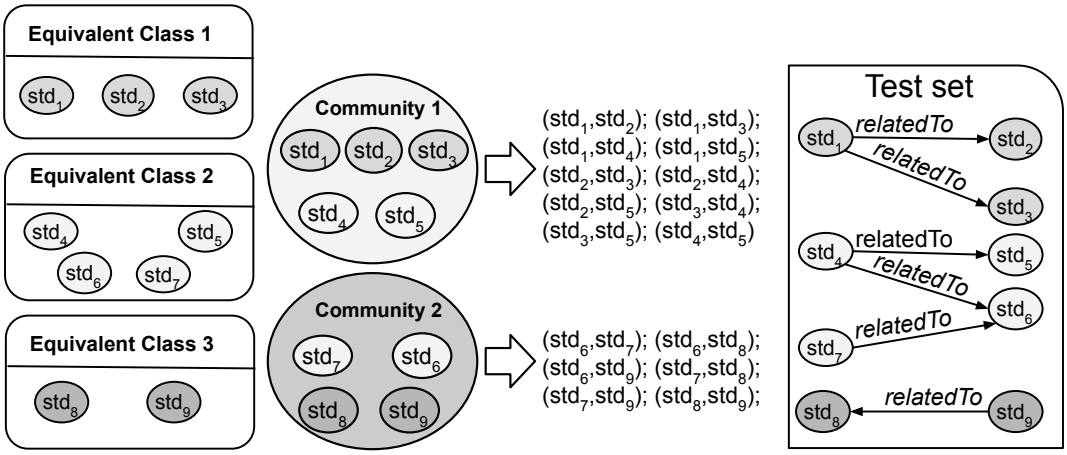
(b) Transitive closure of the property *relatedTo* between I4.0 KG standards

Fig. 4. **Relations between I4.0 KG standards**. (a) Using explicit relations between standards in I4.0 KG, 109 connected components are found. (b) Applying transitive closure of the property *relatedTo*, 20 connected components are found, eight less than in (a). Standards in I4.0 KG are more connected and new relations in the connected components correspond to meaningful relations.

utilizes the function Kernel Density Estimation (KDE) to compute the density of both similarity measures, Cosine Similarity and Inverse Euclidean Distance; it sets to zero the similarity values lower than the given threshold.

Detecting Communities of Standards. *I4.0*$\mathcal{RD}$ maps the problem of computing groups of potentially related standards to the problem of community detection. Once the embeddings are learned, the standards are represented in a vectorial way according to their functions preserving their semantic characteristics. Using the embeddings, *I4.0*$\mathcal{RD}$ computes the similarity between the standards in the I4.0 KG as mentioned in the previous section. The values of similarity between standards are utilized to partition the set of standards in a way that standards in a community are highly similar but dissimilar to the standards in other communities. As proof of concept, three state-of-the-art community detection algorithms have been used in *I4.0*$\mathcal{RD}$: SemEP, METIS, and KMeans. They implement diverse strategies for partitioning a set based on the values of similarity, and our goal is to evaluate which of the three is more suitable to identify meaningful connections between standards.

Discovering Relations Between Standards. New relations between standards are discovered in this step; the *homophily* prediction principle is applied over each of the communities and all the standards in a community are assumed to be related. Figure 5 depicts an example where new relations are computed from two communities; unknown relations correspond to connections between standards in a community that did not existing in the input I4.0KG. Figure 5a shows the equivalent classes of the I4.0KG example. The Community 1 has five standards where three of them belong to Equivalent Class 1 and the other two belong to Equivalent Class 2. Applying the homophily prediction principle to Community 1, six new relations are found between standards from Equivalent Class 1 and Equivalent Class 2, these are: $(std_1, std_4), (std_2, std_4), (std_3, std_4), (std_1, std_5), (std_2, std_5), (std_3, std_5)$. These new relations are evaluated by expert to proof that they correspond to meaningfully relations.



(a) Equivalent classes induced by the property *relatedTo*

(b) Application of the Homophily Prediction Principle

(c) Known Relations used to determine discovered relations between standards

Fig. 5. **Discovering relations between standards**. (a) The homophily prediction principle is applied on two communities, as a result, 16 relations between standards are found. (b) Five out of the 16 found relations correspond to meaningfully relations.

6. Empirical Evaluation

We use as baseline the equivalent classes induced by the property *relatedTo*. An equivalent class is induced by equivalent relations like *relatedTo* that satisfies three properties, i.e., the relation is reflexive, symmetric, and transitive. The equivalent classes are partitions of the set of standards induced by the relation *relatedTo*. Figure 6 shows the number of partitions and how many standards each partition of our baseline has. Equivalent Class 1 has the highest number of standards with 148. All the standards in each equivalent class are related to each other but isolated from the other equivalent classes. Assuming that the different combinations of similarity measures together with the community detection algorithms are effective predictors of the standards communities, then the distances between the equivalent classes and the communities discovered should be close. The Average Category-based Score measure assesses the distance between Communities and the baseline.

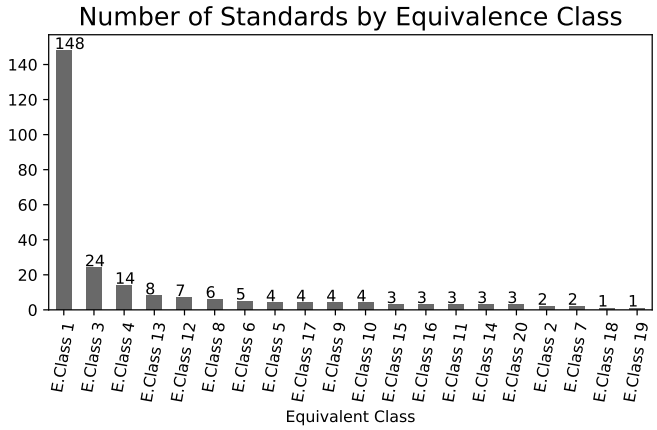


Fig. 6. **Baseline of Equivalent Classes.** I4.0KG has 20 Equivalent Classes and most of them have less than 10 standards except the Equivalent Classes 1, 2 and 5.

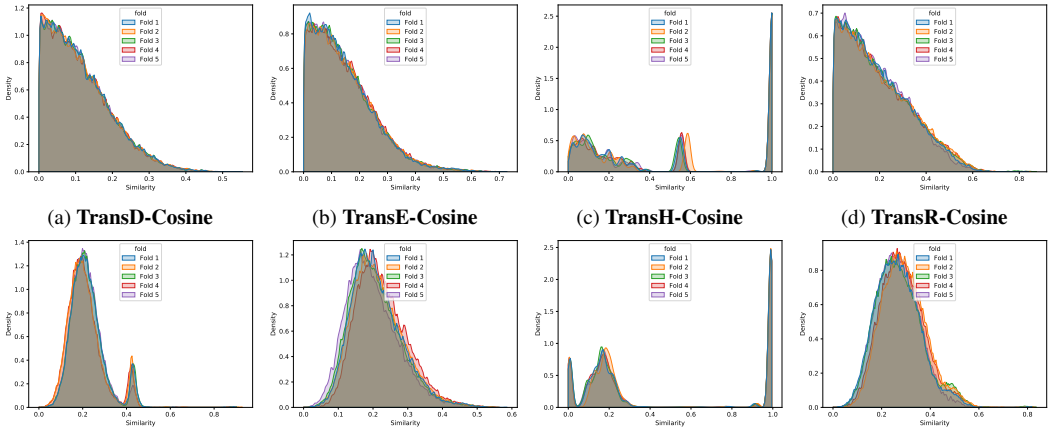
We report on the impact that the knowledge encoded in I4.0 knowledge graph has in the behavior of *I4.0*$\mathcal{RD}$. In particular, we assess the following research questions:

RQ1) How the function used to determine the relatedness between standards impact on the outcome of the problem of uncovering relations among standards?

RQ2) Does a semantic community based analysis on I4.0KG allow for improving the quality of predicting new relations on the I4.0 standards landscape?

RQ3) What is the effect of combining distinct similarity measures, embedding techniques, and community detection algorithms in the task of detecting the relatedness among standards?

Experiment Setup: Four embedding algorithms are considered to build the standards embedding. Each of these algorithms is evaluated independently. Next, a similarity matrix for the standards embedding is computed. Cosine Similarity and Inverse Euclidean Distance are considered as similarity measures. The similarity matrix is required for applying the community detection algorithms. In our experiments, three algorithms are used to compute the Communities. In total, we evaluate twenty-four combinations between embedding algorithms, similarity measure and community detection algorithms. To assure statistical robustness, we execute 5-folds cross-validation with one run. For



(a) **TransD-Cosine** (b) **TransE-Cosine** (c) **TransH-Cosine** (d) **TransR-Cosine**

(e) **TransD-Euclidean** (f) **TransE-Euclidean** (g) **TransH-Euclidean** (h) **TransR-Euclidean**

Fig. 7. **Similarity density by Cosine and Inverse Euclidean Distance of each fold per Trans* methods**. Results from the Inverse Euclidean Distance in all the Trans* methods have higher similarity values than Cosine similarity. Figures 7a, 7b, and 7d show that all folds have values close to zero, i.e., with embeddings created by TransD, TransE, and TransR the standards are very different from each other. However, TransH in both similarity measure (cf. Figure 7c and Figure 7g), exploits properties of the standards and generates embeddings with a different distribution of similarity, i.e., values between 0.0 and 0.4, as well as values close to 1.0. According to known characteristics of the I4.0 standards, the TransH distribution of similarity using both Cosine Similarity and Inverse Euclidean Distance better represents their relatedness.

the purposes of understanding how the Trans* methods, similarity measures, and community detection algorithms are performing, we evaluate the similarity density of the standards by Trans* methods, also the quality of the generated Communities, the accuracy of the Communities in discovering new relationships and the distance between the Communities and the baseline using Cosine and Inverse Euclidean Distance.

Implementation: Our proposed approach is implemented in Python 2.7 and integrated with the PyKeen (Python KnowlEdge EmbeddiNgs) framework [27], METIS 5.1 [f], SemEP [g], and Kmeans [h]. The experiments were executed on a GPU server with ten chips Intel(R) Xeon(R) CPU E5-2660, two chips GeForce GTX 108, and 100 GB RAM.

Thresholds for Computing Values of Similarity. Figure 7 depicts the density function of each fold for each embedding algorithm using the similarity metrics Cosine Similarity and Inverse Euclidean Distance. We notice that Inverse Euclidean Distance finds a higher density of similar standards than the Cosine Similarity metric in all Trans* methods. Figures 7a and 7b show the values of the folds of TransD and TransE, in Cosine Similarity, where all the similarity values are close to 0.0, i.e., all the standards are different. Figure 7d suggests that all the folds have similar behavior with values between 0.0 and 0.5 and a short group of standards with similarity values in 0.8. Figure 7c and Figure 7g shows a group of standards similar with values close to 1.0 and the rest of the standards between 0.0 and 0.4. The percentile of the similarity matrix is computed with a threshold of 0.85. That means all values of the similarity matrix which are less than the percentile computed, are filled

[f] `http://glaros.dtc.umn.edu/gkhome/metis/metis/download`
[g] `https://github.com/SDM-TIB/semEP`
[h] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

with 0.0 and then, these two standards are dissimilar. After analyzing the density of each fold (cf. Figure 7), the thresholds of TransH and TransR using Cosine Similarity are set to 0.50 and 0.75, respectively. The reason is because the two cases with a high threshold find all similar standards and it will not be possible to create more than one Community of standards. The thresholds of the similarity matrix using Inverse Euclidean Distance are also modified for the same reason.TransD, TransH and TransR are set to 0.95, 0.60 and 0.75, respectively. In the case of TransH, there is a high density of values close to 1.0; it indicates that for a threshold of 0.85, the percentile computed is almost 1.0. the values of the similarity matrix less than the threshold are filled with 0.0; values of 0.0 represent that the compared standards are not similar.

6.1. *Impact of Metrics for Determining Relatedness among Standards*

There are a variety of metrics to evaluate the quality of clusters. We used five recognized cluster metrics to estimate the quality of the communities from the I4.0KG embeddings. All the metrics are normalized in the range [0,1] where higher is better score.

a) **Conductance (InvC)**: measures relatedness of entities in a community, and how different they are to entities outside the community [28]. The inverse of Conductance is reported: $1 - Conductance(K)$, where $K = \{k_1, k_2, ..., k_n\}$ the set of standards communities obtained by the cluster algorithm, and k_i are the computed clusters.

b) **Performance (P)**: sums up the number of intra-community relationships, plus the number of non-existent relationships between communities [28]. Higher values indicate that a cluster is both internally dense and externally sparse.

c) **Total Cut (InvTC)**: sums up all similarities among entities in different communities [29]. The Total Cut values are normalized by dividing the sum of the similarities between the entities. The inverse of Total Cut is reported as follows: $1 - NormTotalCut(K)$

d) **Modularity (M)**: is the value of the intra-community similarities between the entities divided by the sum of all the similarities between the entities, minus the sum of the similarities among the entities in different communities, in case they are randomly distributed in the communities [30]. The value of the Modularity is in the range of $[-0.5, 1]$, which can be scaled to $[0, 1]$ by computing: $\frac{Modularity(K)+0.5}{1.5}$.

e) **Coverage (Co)**: compares the fraction of intra-community similarities between entities to the sum of all similarities between entities [28]. Higher coverage values mean that there are more edges within clusters than edges linking different clusters.

6.2. *Quality of the Predicted Relations among Standards*

The quality of the predicted relations among standards is evaluated by the accuracy. In order to measure the accuracy of the predicted relations in the communities, we are comparing them with the relations in the test set. The test set (TS) is used to validate the results and it is represented as $TS = \{\langle s, p, o\rangle | s, o \in V_e, p \in \textit{relatedTo}\}$ and V_e are standards (cf. Figure 5c). Considering we are applying the homophily prediction principle in the communities, all the standards in a community (c) are related to each other (cf. Figure 5b). Homophily prediction in a community is defined as

follow: $H(c) = \{\langle s,p,o\rangle | s,o \in c \wedge p \in \textit{relatedTo} \wedge s \neq o\}$. Then, we are selecting from TS the set of triples $\langle s,p,o\rangle$ where s or o are standards from cluster c; it is defined as follow: $S(c,TS) = \{\langle s,p,o\rangle | \langle s,p,o\rangle \in TS \wedge (s \in c \vee o \in c)\}$. Finally, is evaluated the percentage of predicted relations $acc(c)$ among standards in community c; $acc(c) = \frac{|S(c,TS) \cup H(c)|}{|S(c,TS)|}$, where the numerator corresponds to number of discovered relations from c. Since we are executing 5-folds cross-validation with one run, is reported the average of the accuracy.

6.3. *Impact of Community Detection Methods*

Average Category-based Score: We compared our baseline, Equivalent Classes, with the communities generated by the community detection algorithms. Given a Community **C** of standards, the average Category-based Score, $\mathcal{C}(C)$, corresponds to the average of the 'Category-based' measure for each pair of standards in the clusters of **C**. Values of $\mathcal{C}(C)$ are in the ranges between 0.0 and 1.0. A value equal to 0.0 indicates that there is no intersection between the classes of equivalence of the pairs of standards in the clusters of **C**, whereas a value closed to 1.0 represents that almost all the pairs of standards in each cluster of **C** share exactly the same classes of equivalence. Let EC be the Equivalent classes, EC_i be the set of standards in the Equivalent Class i, C_k be the set of standards in the Community k and $Comb(n)$ represents the number of pair of standards given a set of standards with cardinality n; it is computed by the number of two combinations of a set of n elements, $Comb(n, r = 2) = \frac{n!}{(n-2)!*2!} = \frac{n*(n-1)}{2}$. The Average Category-based Score is defined as follows:

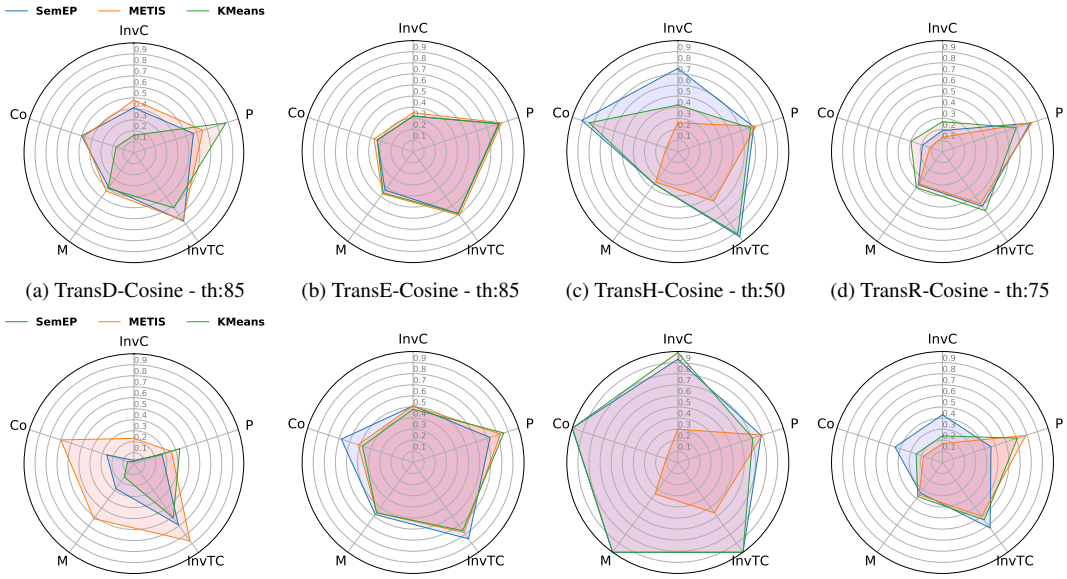
$$\mathcal{C}(C_k) = \frac{\sum_{i=1}^{|EC|} Comb(|C_k \cup EC_i|)}{Comb(|C_k|)}$$

$$avg(\mathcal{C}) = \frac{\sum_{k=1}^{|C|} \mathcal{C}(C_k)}{|C|}$$

Quality of the communities: We evaluated three community detection algorithms with two different similarity metrics and four Trans* methods. Considering the five metrics for assessing the communities, the best communities are obtained by Inverse Euclidean Distance, TransH, and with the SemEP and KMeans algorithms. Figure 8g shows how the InvTC, M, and Co have values close to one for SemEP and KMeans. The Performance (P) for SemEP and KMeans is 0.8 and 0.7 respectively, which means that communities built by KMeans have more external links to other communities than communities by SemEP. The inverse of Conductance (InvC) is high in both SemEP and KMeans with 0.93 and 0.99 respectively. This metric measures the relatedness of standards in a community, and how different they are from standards outside the community.

The *I4.0*$\mathcal{RD}$ accuracy: Figure 9b shows the best performance for TransH-KMeans achieving 100% of accuracy. However, KMeans is only able to discover three communities of standards while our baseline is already known to have twenty equivalence classes. This means that KMeans is clustering our 249 standards into just three clusters. K-Means finds the optimal number of clusters by computing the K-Elbow curve, but the results are not close to our baseline. Nevertheless, SemEP achieves an accuracy of over 90% in both similarity measures and furthermore, the number of communities discovered is very close to our baseline, reaching a mean of 16 communities. All the communities are assessed against the baseline to validate their closeness to the equivalence classes.

Baseline: TransH is selected as the best embedding according to the results achieved in the metrics for determining relatedness among Standards (cf. Figure 8) and quality of the predicted relations



(a) TransD-Cosine - th:85 (b) TransE-Cosine - th:85 (c) TransH-Cosine - th:50 (d) TransR-Cosine - th:75

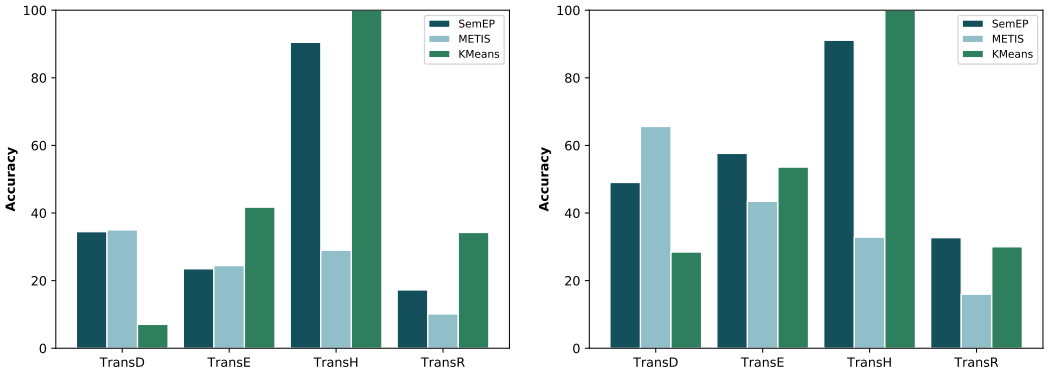
(e) TransD-Euclidean - th:95 (f) TransE-Euclidean - th:85 (g) TransH-Euclidean - th:60 (h) TransR-Euclidean - th:75

Fig. 8. **Quality of the generated communities**. Communities evaluated in terms of prediction metrics using the SemEP, METIS, and KMeans algorithms. Communities are derived for each combinations of Trans* method and similarity measure. In this case higher values are better. Our approach exhibits the best performance with TransH embeddings in both Cosine Similarity and Inverse Euclidean Distance, i.e., Figure 8c and Figure 8g. SemEP achieves the highest values in the five evaluated parameters using Inverse Euclidean Distance and in four of the five evaluated parameters with Cosine Similarity.

among standards (cf. Figure 9). Taking TransH as the best embedding the communities generated by the three community detection algorithms and the two similarity measures are evaluated. Figure 10 depicts the results of the measure Average Category-based Score for both similarity measures. The combination SemEP and TransH achived the best performance in both similarity measures, see Figure 10a and Figure 10b. Although KMeans has the highest accuracy, the performance in the measure Average Category-based Score where it is compared with the baseline is one of the lowest. In contrast, SemEP has the highest values for this measure and is also over 90% accuracy, which means that the communities discovered by SemEP are the closest to our baseline and with high accuracy.

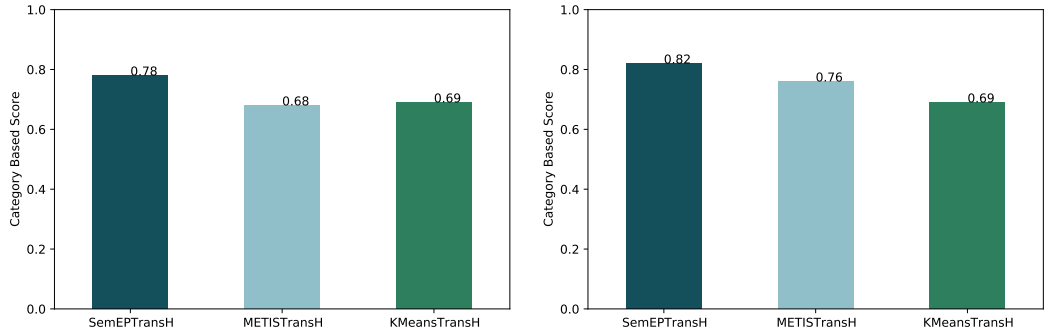
Network analysis: The I4.0KG is updated with the communities found by the combination of TransH, Inverse Euclidean Distance, and SemEP which is the best performer for the metrics evaluated. With the updated I4.0KG we are adding new links predicted by the communities. Table 1 shows the analysis of I4.0KG with new predicted links against our baseline. We improve the standards connectivity by predicting new links.

RQ1 - Corroborating the quality of communities in I4.0KG. To compute accuracy of *I4.0*$\mathcal{RD}$, we executed a five-folds cross-validation procedure. To that end, the data set is divided into five consecutive folds shuffling the data before splitting into folds. Each fold is used once as validation, i.e., test set while the remaining fourth folds form the training set. Figure 8 depicts the impact of metrics for evaluate communities. The best results are obtained with the combination of the Inverse Euclidean Distance and TransH with SemEP and KMeans algorithms, see Figure 8g. The values obtained for this



(a) Cosine Similarity Measure (b) Inverse Euclidean Distance Measure

Fig. 9. The ***I4.0RD*** **accuracy**. Percentage of the test set for the property *relatedTo* is achieved in each cluster. Figure 9a and Figure 9b shows the precision of the community detection algorithms by the measure Cosine Similarity and Inverse Euclidean Distance respectively. Our approach exhibits the best performance using TransH embedding and with the SemEP and KMeans algorithms in both similarity measures reaching an accuracy by up to 90%.



(a) Cosine Similarity Measure (b) Inverse Euclidean Distance Measure

Fig. 10. **Average Category Based Score respect to Equivalence Class**. Figure 10a and Figure 10b shows how similar our communities are to the baseline. Our approach exhibits the best performance with Inverse Euclidean Distance and SemEP achieving 82%.

combination for both SemEP and KMeans are high except for the metric **Performance** (P). SemEP and KMeans have values of 0.8 and 0.7 respectively, which means that communities built by KMeans have more external links to other communities than communities by SemEP.

RQ2 - Predicting new relations between standards. In order to assess the second research question, the data set is divided into five consecutive folds. Each fold comprises 20% of the relationships between standards. Next, the precision measurement is applied to evaluate the main objective: to unveil uncovered associations and, at the same time, corroborate knowledge patterns that are already known. As shown in Figure 9, the best performances for the property *relatedTo* are achieved by TransH embeddings in combination with the SemEP and KMeans algorithm in both similarity measures. KMeans reaches higher accuracy than SemEP, however, KMeans discover only three communities of standards while our baseline is already known to have twenty Equivalence Classes. On the other hand, the number of communities discovered by SemEP is very close to our baseline, reaching

Table 1. **Network connectivity analysis**. Table 1 shows the statistics for I4.0KG after transitive closure of the property *relatedTo* between standards and the statistics I4.0KG with the new links predicted by combining TransH, Inverse Euclidean Distance and SemEP. Results reveal a general improvement in connectivity when predicting new links. The Number of edges, Avg. number of neighbors and Network density increase predicting new links and this allows for fewer connected components and improves data integration. Measures that improve are highlighted in **bold**. The network analysis was performed by Cytoscape [31].

Statistic	Baseline	TransH-Inv.EuclideanDistance-SemEP
Number of nodes	249	249
Number of edges	22,969	**23,207**
Avg. number of neighbors	91.245	**92.201**
Network diameter	1	**3**
Network radius	1	1
Characteristic path length	1.000	1.001
Clustering coefficient	0.976	0.974
Network density	0.368	**0.372**
Connected components	20	**13**
Multi-edge node pairs	11,360	**11,479**
Number of self-loops	249	249

a mean of 16 communities. The communities of standards discovered using TransH embeddings, Inverse Euclidean Distance, and the SemEP algorithm contribute to the resolution of interoperability in I4.0 standards. To provide an example of this, we observe a resulting cluster with the standards *IEC 60255 P27 E2, IEC 60255 P151 E1, IEC 60255 2010, IEC 60255 P1 E1, IEC 60255 P149 E1* and *MTConnect*. The former provides an information model for describing manufacturing data. The latter offers a vocabulary for manufacturing equipment. It is important to note that the standard *MTConnect* are not related to the training set nor in I4.0KG. The membership of those standards in the cluster means that them should be classified together in the standardization frameworks. Besides, it also suggests to the creators of the standards that they might look after possible existing synergies between them. This example suggests that the techniques employed in this work are capable of discovering new communities of standards. These communities can be used to improve the classification that the standardization frameworks provide for the standards.

RQ3 - Comparison with the baseline of equivalent classes. From the combination of four Trans embeddings, two similarity measures, and three community detection algorithms we asses 24 results. In both the evaluation of the quality of the communities and the accuracy of new relations, the best results are reached with the TransH embedding, SemEP and KMeans as cluster algorithms, and both similarity metrics. Finally, in the evaluation with the baseline, the best similarity metric is Inverse Euclidean Distance and the best clustering algorithm is SemEP. Figure 10b shows Average Category Based Score achieved by SemEP respect to Equivalence Class. We reach quite high values which means that almost all the pairs of standards in each community share the same equivalence classes.

6.4. *Discussion*

The techniques proposed in this paper rely on known relations between I4.0 standards to discover novel patterns and new relations. During the experimental study, we observe that these techniques could group together not only standards that are known to be related, but also standards whose relat-

edness is implicitly represented in the I4.0KG. This feature facilitates the detection of high-quality communities as reported in Figure 8, as well as for an accurate discovery of relations between standards (cf. Figure 9) and for the evaluation with the baseline of equivalent classes, as shown in Figure 10. As observed, the accuracy of the approach can be benefited from the application of the Trans* family algorithms, e.g., TransH, and from similarity measures, e.g., Inverse Euclidean Distance. Additionally, SemEP groups in the same communities highly similar standards, and leads our approach into high-quality discoveries. Our results suggest that the techniques TransH, Inverse Euclidean Distance, and SemEP uncover meaningful communities with high quality because, the performance of the five metrics for evaluating communities are close to one, which means that standards in a community are different from standards outside the community, and there are more edges within communities than edges linking different communities. Also, the accuracy is up 90% which means that are discovered over 90% of the relationships and evaluating with the baseline achieving 82%, i.e., almost all the pairs of standards in each community share exactly the same equivalence classes. Moreover, the number of communities is close to the number of equivalent classes in the baseline.

To understand why the aforementioned combination of TransH, Inverse Euclidean Distance and SemEP produces the best results, we analyze in detail both techniques. TransH introduces the mechanism of projecting the relation to a specific hyperplane [32], enabling, thus, the representation of relations with cardinality many to many. Since the materialization of transitivity and symmetry of the property *relatedTo* corresponds to many to many relations, the instances of this materialization are taken into account during the generation of the embeddings, specifically, during the translating operation on a hyperplane. Thus, even thought semantics is not explicitly utilized during the computation of the embeddings, considering different types of relations, empowers the embeddings generated by TransH. Moreover, it allows for a more precise encoding of the standards represented in I4.0KG. Figures 7c and 7g illustrate groups of standards in the similarity intervals $[0.9, 1.0]$, and $[0.0, 0.4]$. Inverse Euclidean Distance is able to find in all the Trans* methods a higher density of similar standards than Cosine Similarity. The SemEP algorithm can detect these similarities and represent them in high-precision communities. The other three models embeddings, i.e., TransD, TransE, and TransR do not represent the standards in the best way with either of the two similarity measures. TransD, TransE, and TransR report that most of the standards are in the similarity interval $[0.0, 0.4]$ (cf. Figure 7). This means that no community detection algorithm could be able to discover communities with high quality. Reported results indicate that the presented approach enables – in average– for discovering communities of standards by up to 90%. As an example of a relevant community, we observed a resulting cluster with the standards *IEC 60255 P27 E2, IEC 60255 P151 E1, IEC 60255 2010, IEC 60255 P1 E1, IEC 60255 P149 E1*, and *MTConnect*. All of them are related to product safety requirements and vocabulary for manufacturing equipment. It is important to note that the *MTConnect* standard is in a different equivalent class than the other community standards. However, our approach *I4.0*$\mathcal{RD}$ is able to create a community grouping all of them together. Although these results required the validation of experts in the domain, an initial evaluation suggest that the results are accurate.

7. Conclusions and Future Work

In this paper, we presented the *I4.0*$\mathcal{RD}$ approach that combines knowledge graphs and embeddings to discover associations between I4.0 standards. Our approach resorts to I4.0KG to discover relations between standards; I4.0KG represents relations between standards extracted from the literature or defined according to the classifications stated by the standardization frameworks. Since the relation

between standards is symmetric and transitive, the transitive closure of the relations is materialized in I4.0KG. Different algorithms for generating embeddings are applied on the standards according to the relations represented in I4.0KG. Two similarity measure are applied to asses the similarity of the standards. We employed three community detection algorithms, i.e., SemEP, METIS, and KMeans to identify similar standards, i.e., communities of standards, as well as to analyze their properties. Additionally, by applying the homophily prediction principle, novel relations between standards are discovered. We empirically evaluated the quality of the proposed techniques over 249 standards, initially related through 736 instances of the property *relatedTo*; as this relation is symmetric and transitive, its transitive closure is also represented in I4.0KG with 22,969 instances of *relatedTo*. Furthermore, the equivalent classes induced by the property *relatedTo* were used as baseline in the evaluation process. The Trans* family of embedding models were used to identify a low-dimensional representation of the standards according to the materialized instances of *relatedTo*. Results of a 5-fold cross validation process suggest that our approach is able to effectively identify novel relations between standards. In addition, the Inverse Euclidean Distance enable to identify patterns and links with higher precision. Thus, our work broadens the repertoire of knowledge-driven frameworks for understanding I4.0 standards, and we hope that our outcomes facilitate the resolution of the existing interoperability issues in the I4.0 landscape. As for future work, we envision having a more fine-grained description of the I4.0 standards and improve the embeddings of the standards preserving their semantic characteristics. Furthermore, evaluate other types of embedding methods and other community detection methods.

Acknowledgements

Ariam Rivas is supported by the German Academic Exchange Service (DAAD). This work has been partially funded by the EU H2020 Projects IASIS (GA 727658) and LAMBDA (GA 809965), and the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A).

References

1. Peter Adolphs, Sören Auer, Meik Billmann, Martin Hankel, Roland Heidel, Michael Hoffmeister, Haimo Huhle, Michael Jochem, Markus Kiele, Gunther Koschnick, Heiko Koziolek, Lukas Linke, Reinhold Pichler, Frank Schewe, Karsten Schneider, and Bernd Waser. Structure of the Administration Shell. Status report, ZVEI and VDI, 2016.
2. Shi-Wan Lin, Bradford Miller, Jacques Durand, Graham Bleakley, Amine Chigani, Robert Martin, Brett Murphy, and Mark Crawford. The Industrial Internet of Things Volume G1: Reference Architecture. White Paper IIC:PUB:G1:V1.80:20170131, Industrial Internet Consortium, 2017.
3. Behzad Golshan, Alon Y. Halevy, George A. Mihaila, and Wang-Chiew Tan. Data integration: After the teenage years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 101–106, 2017.
4. Olga Kovalenko and Jérome Euzenat. Semantic matching of engineering data structures. In *Semantic Web for Intelligent Engineering Applications*. Springer, 2016.
5. Michalis Mountantonakis and Yannis Tzitzikas. Large-scale semantic integration of linked data: A survey. *ACM Comput. Surv.*, 52(5):103:1–103:40, September 2019.
6. Sebastian R. Bader, Irlán Grangel-González, Priyanka Nanjappa, Maria-Esther Vidal, and Maria Maleshkova. A knowledge graph for industry 4.0. In *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings*, pages 465–480, 2020.
7. Jack Hodges, Kimberly García, and Steven Ray. Semantic Development and Integration of Standards for Adoption and Interoperability. *IEEE Computer*, 50(11):26–36, 2017.
8. Peter O'Donovan, Kevin Leahy, Ken Bruton, and Dominic T. J. O'Sullivan. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *J. Big Data*,

2:25, 2015.
9. Francesco Lelli. Interoperability of the time of industry 4.0 and the internet of things. *Future Internet*, 11(2):36, 2019.
10. Shi-Wan Lin, Brett Murphy, Erich Clauer, Ulrich Loewen, Ralf Neubert, Gerd Bachmann, Madhusudan Pai, and Martin Hankel. Reference Architectural Model Industrie 4.0 (RAMI 4.0). Technical report, Industrial Internet Consortium and Plattform Industrie 4.0, 2017.
11. Irlán Grangel-González, Paul Baptista, Lavdim Halilaj, Steffen Lohmann, Maria-Esther Vidal, Christian Mader, and Sören Auer. The industry 4.0 standards landscape from a semantic integration perspective. In *22nd IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, Limassol, Cyprus, September 12-15*, pages 1–8, 2017.
12. Irlán Grangel-González, Diego Collarana, Lavdim Halilaj, Steffen Lohmann, Christoph Lange, María-Esther Vidal, and Sören Auer. Alligator: A deductive approach for the integration of industry 4.0 standards. In *20th Int. Conf. on Knowledge Engineering and Knowledge Management, EKAW*, pages 272–287, 2016.
13. Irlán Grangel-González, Lavdim Halilaj, Maria-Esther Vidal, Omar Rana, Steffen Lohmann, Sören Auer, and Andreas W. Müller. Knowledge graphs for semantically integrating cyber-physical systems. In *Database and Expert Systems Applications - 29th International Conference, DEXA, Regensburg, Germany, September 3-6, Proceedings, Part I*, pages 184–199, 2018.
14. Guillermo Palma, Maria-Esther Vidal, and Louiqa Raschid. Drug-target interaction prediction using semantic similarity and edge partitioning. In *Proc. of the 13th Int. Semantic Web Conf. - Part I*, ISWC '14, pages 131–146, NY, USA, 2014. Springer-Verlag New York, Inc.
15. George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, December 1998.
16. David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.
17. Ariam Rivas, Irlán Grangel-González, Diego Collarana, Jens Lehmann, and Maria-Esther Vidal. Unveiling relations in the industry 4.0 standards landscape based on knowledge graph embeddings. In Sven Hartmann, Josef Küng, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil, editors, *Database and Expert Systems Applications. DEXA 2020. Lecture Notes in Computer Science*, pages 179–194, Cham, 2020. Springer International Publishing.
18. Abe Zeid, Sarvesh Sundaram, Mohsen Moghaddam, Sagar Kamarthi, and Tucker Marion. Interoperability in smart manufacturing: Research challenges. *Machines*, 7(2):21, 2019.
19. Paula Monteiro, Marcia Carvalho, Francisco Morais, Monica Melo, Ricardo Machado, and Fernando Pereira. Adoption of architecture reference models for industrial information management systems. In *Int. Conf. on Intelligent Systems (IS)*, pages 763–770. IEEE, 2018.
20. Emmanuel Darmois, Omar Elloumi, Patrick Guillemin, and Philippe Moretto. *Digitising the Industry - Internet of Things Connecting the Physical, Digital and Virtual Worlds*. River Publishers, 2016.
21. Martina Garofalo, Maria Angela Pellegrino, Abdulrahman Altabba, and Michael Cochez. Leveraging knowledge graph embedding techniques for industry 4.0 use cases. *CoRR*, abs/1808.00434, 2018.
22. Christian Galinski. Interoperability of metadata. semantic interoperability. Technical report, International Information Centre for Terminology, Austria, 2005.
23. Sebastian R. Bader, Irlán Grangel-González, Mayesha Tasnim, and Steffen Lohmann. Structuring the industry 4.0 landscape. In *24th IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, Zaragoza, Spain, September 10-13*, pages 224–231, 2019.
24. Nitishal Chungoora, A-F Cutting-Decelle, RIM Young, G Gunendran, Zahid Usman, Jennifer A Harding, and Keith Case. Towards the ontology-based consolidation of production-centric standards. *International Journal of Production Research*, 51(2):327–345, 2013.
25. Yan Lu, K. C. Morris, and Simon Frechette. Standards landscape and directions for smart manufacturing systems. In *IEEE International Conference on Automation Science and Engineering, CASE, Gothenburg, Sweden, August 24-28*, pages 998–1005, 2015.
26. Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*, 29(12):2724–2743, 2017.

27. Mehdi Ali, Hajira Jabeen, Charles Tapley Hoyt, and Jens Lehmann. The keen universe: An ecosystem for knowledge graph embeddings with a focus on reproducibility and transferability. (in press).
28. Gaertler Marco and Erlebach Thomas. *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media, 2005.
29. Aydin Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. Recent advances in graph partitioning. In *Algorithm Engineering - Selected Results and Surveys*, pages 117–158. 2016.
30. M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
31. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research, 2003.
32. Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zhigang Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.