8-2023

Semantically constitutive entities in knowledge graphs

Chong Cher CHIA
Singapore Management University, ccchia.2018@phdis.smu.edu.sg

Maksim TKACHENKO
Singapore Management University, mtkachenko@smu.edu.sg

Hady Wirawan LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Semantically Constitutive Entities in Knowledge Graphs

Chong Cher Chia[1][0000−0001−6053−4643], Maksim Tkachenko[0000−0001−6687−0525], and Hady W. Lauw[0000−0002−8245−8677]

School of Computing and Information Systems, Singapore Management University
ccchia.2018@smu.edu.sg; maksim.tkatchenko@gmail.com; hadywlauw@smu.edu.sg

Abstract. Knowledge graphs are repositories of facts about a world. In this work, we seek to distill the set of entities or nodes in a knowledge graph into a specified number of constitutive nodes, whose embeddings would be retained. Intuitively, the remaining accessory nodes could have their original embeddings "forgotten", and yet reconstitutable from those of the retained constitutive nodes. The constitutive nodes thus represent the semantically constitutive entities, which retain the core semantics of the knowledge graph. We propose a formulation as well as algorithmic solutions to minimize the reconstitution errors. The derived constitutive nodes are validated empirically both in quantitative and qualitative means on three well-known publicly accessible knowledge graphs. Experiments show that the selected semantically constitutive entities outperform those selected based on structural properties alone.

Keywords: semantically constitutive · knowledge graph · embeddings.

1 Introduction

Graphs are predominantly used to represent real world data, including social networks, citation network, hyperlink network, etc. One important analysis deals with determining which vertices are the most 'important' in a graph. Because the essential nature of graphs is the very connectivity among its vertices, this notion of 'importance' is frequently formulated in terms of how well a vertex is connected to others in the graph, giving rise to notions such as centrality [4] and influence maximization [25] that would be further explored in related work.

In this work, we are interested in *knowledge graphs*, a machine-friendly way of representing real world facts. These facts are extracted from various sources such as encyclopedic Wikipedia [33], lexical WordNet [14], or even the open Web [34]. The use of knowledge graphs have been extended to applications including question answering [21], recommendations [52], fact-checking [8], etc.

Given its pertinence and myriad applicability, we explore notions of what make a vertex 'important' in a knowledge graph. In addition to the graph-theoretic sense of connectivity, another essential nature of a knowledge graph is its *semantics*. Every triplet instance involving a *head entity*, *relation*, and *tail entity* represents a fact, the totality of which collectively represents our semantic understanding of an underlying 'world'. Suppose we retain only a subset of the entities; which subset best preserves our semantic understanding of the 'world'?

For a concrete representation of semantics, we allude to knowledge graph embeddings [51], which embeds entities and relations into continuous vector spaces.

The plausibility of facts (triplets) can then be assessed from the embeddings of the corresponding entities and relations. In this work, we assume that such embeddings have been derived and specified as input to our problem.

As output, we seek to identify a relatively small subset of ("*constitutive*") entities, whose embeddings would be used to reconstitute the remaining ("*accessory*") entities. To remain true to the raison d'être of a knowledge graph, this reconstitution is faithful to a known fact (triplet) within the graph.

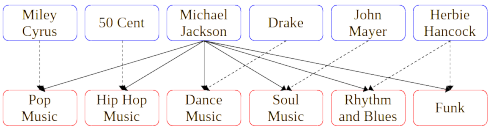


Fig. 1: Subgraph Using 1 or 2 Constitutive Nodes

This concept is illustrated by the knowledge graph subset in Figure 1, where constitutive nodes (top) are connected to accessory nodes (bottom) by relational edges (in this case `Genre`). Embeddings of accessory nodes can potentially be "forgotten", and "reconstituted" by constitutive nodes. For example, we could use a single constitutive node (`Michael Jackson`) to reconstitute all the accessory nodes (solid edges only). While compact, it is not sufficient for distinguishing different musical genres, and using two constitutive nodes produces more informative reconstitutions (`John Mayer`, `Michael Jackson` both produce `Soul Music`).

Contributions. In this work, we make several contributions.

- The problem of identifying semantically constitutive entities in a knowledge graph is novel, and distinct from existing work solely focused on structural connectivity. We propose a reconstitution function consistent with translational embeddings, and produces interpretable reconstitutions by virtue of being supported by actual triplets within the knowledge graph.
- We propose a new algorithmic formulation to identify constitutive nodes, as well as the selection of triplets for each reconstitution. While related to matching or assignment problems, our formulation is novel in allowing up to k constitutive nodes per accessory node. We describe algorithmic solutions based on Integer Linear Programming (ILP), and propose heuristics that speed up the computation particularly for larger graphs.
- We experiment on 3 well-established knowledge graphs, outperforming baselines both quantitatively (downstream tasks), and qualitatively (user study).
- We make our code publicly available[1] for reproducibility.

2 Related Work

Node Centrality Finding the "important" nodes in a graph had previously been approached from structural connectivity. One class of techniques referred to as *point centrality* looks at the quality of an individual node that makes it most central. There are primarily three categories of point centrality measures: local centrality, iterative centrality, or global centrality [48]. *Local centrality* measures centrality by local network topology. A common metric is Degree Centrality,

[1] https://github.com/PreferredAI/semantically-constitutive-entities

which ranks each node based on the number of edges in the graph. For directed graphs, the in-degrees or out-degrees of a node may be used. Another well-known local centrality metric is *h-index* [19,26], where a node has index h if it has at least h neighbouring nodes of h degree. *Iterative centrality* metrics measure a node's centrality through some (possibly fixed) number of iterative calculations. One such metric is Eigenvector Centrality [1,2], which repeatedly updates the centrality for each node based on the centrality of its neighbours. PageRank [40] builds on Eigenvector Centrality by dampening the influence of further neighbours on the centrality of a given node. *Global centrality* measures a node's centrality in the context of the entire network topology, such as Betweenness Centrality [16], derived from the number of shortest paths passing through it.

In the experiments, we compare against representative point centrality metrics, such as degree centrality and PageRank. Such point centrality measures select nodes based on its individual quality. In our problem, we seek to select a *group* of constitutive nodes. Hence, we compare against the group version of these metrics in the experiments. For example, compared to PageRank that selects nodes individually, another formulation of influence maximization seeks to identify a group of "influential" nodes based on their ability to affect other nodes within the graph, in order to maximize social influence [25]. Although NP-hard, algorithms such as SSA guarantee a $(1-1/e-\epsilon)$-approximate solution [20,37,38].

Knowledge Graph A core concept in our work is the representation of semantic information within a knowledge graph. Such representations commonly take the form of Knowledge Graph Embeddings, as discussed in [24]. One class of Knowledge Graph Embeddings are linear/bilinear models, as exemplified by *TransE* [3], which represents relations between two entities as the translation of one point within the embedding space to another. Given a triplet (h, r, t) representing the head entity, relationship, and tail entity respectively, *TransE* minimizes the L_1/L_2 distance between $h + r$ and t. Other Translational Knowledge Graph Embeddings have since been proposed, such as *TransH* [53] which extends the translation operation onto a hyperplane, and *TransD* [23] which uses separate mapping matrices for the head and tail entities, and each projection is defined by both the entity and relation embeddings.

Other classes of embeddings include factorization models (e.g., RESCAL [39], LFM [22]), neural network models (e.g., ConvE [12], ConKB [36]), and transformer-based models (e.g., CoKE [50], KG-BERT [54]).

Inductive Knowledge Graph Completion [11,18,30,49] generates embeddings for unseen entities. This is done from combining embeddings of known entities, and is therefore not comparable with our work.

Another widely studied aspect of knowledge graphs is the summarization of such graphs, typically through the addition and/or removal of nodes (as discussed in [31]). Summaries typically take the form of either a supergraph or a sparsified graph. Supergraphs refer to graphs where the (super)nodes and (super)edges are a collection of nodes/edges from the original graph, and may be obtained by grouping nodes [13,41,42,56] or identifying patterns within the original knowledge graph [6,9,55]. Supergraphs do not retain the entities and edges

of the original graph, and are therefore not comparable with our work. Sparsified graphs are subsets of the original knowledge graph, and reduce the number of nodes and/or edges as compared to the original knowledge graph. This may be accomplished by the introduction of "compressor nodes" [32] or "virtual nodes" [5] to the graph for (edge) dedensification. Other techniques may require a query to base the summary, such as Ontovis [44] or Egocentric Abstraction [28].

3 Semantically Constitutive Entities

Our goal, as stated in Section 1, is to select a (user-specified) number of constitutive nodes from a given knowledge graph. Graph embeddings of constitutive nodes can be used to reconstitute non-selected (i.e., accessory) node embeddings.

Problem Definition A knowledge graph $G = (E, R, T)$ consists of a set of entities E, relations R, and relational triples $T \subset E \times R \times E$. Triple $(h, r, t) \in T$ indicates that relation r is present between head h and tail t entities. Let $\mathrm{H}(\cdot)$ return the corresponding embeddings for entity or relation. For a given target size $\mathcal{P}$, we seek to select a semantically constitutive graph $\hat{G} = (\hat{E}, \hat{R}, \hat{T})$ that ties subset of constitutive entities $\hat{E} \subset E$ with the accessory entities $\hat{E}'$ via relations $\hat{R} \subseteq R$: $\hat{T} \subset \hat{E} \times R \times \hat{E}'$. Formally, we seek to solve the minimization problem:

$$\underset{\hat{G}:|\hat{E}|=\mathcal{P}}{\arg\min} \sum_{e \in \hat{E}'} \mathrm{d}(\mathrm{H}(e), f(e|\hat{G})), \tag{1}$$

where d is a distance function on embeddings (L2 in this work) and $f(e|\hat{G})$ reconstitutes the accessory node e from entities and relations in $\hat{G}$.

To define a particular reconstitution function $f(\cdot\,|\hat{G})$, we draw on the knowledge graph embedding training procedures: a family of related models (Trans*) learn embeddings by treating the relations between entities as translations between two points in a high-dimensional space, which effectively turns into the following equation: $\mathrm{H}(h) + \mathrm{H}(r) \approxeq \mathrm{H}(t)$. A target entity e in principle can be reconstituted using multiple head entities and relations as long as we have an appropriate relation between them:

$$f(e|\hat{G}) = \sum_{(h,r,t)\in\hat{T}} \big[\mathbb{1}_{t=e} \cdot (\mathrm{H}(h) + \mathrm{H}(r))\big] \Big/ \sum_{(h,r,t)\in\hat{T}} \big[\mathbb{1}_{t=e}\big], \tag{2}$$

where $\mathbb{1}_{t=e}$ is 1 when t and e refer to the same node and 0 otherwise.

We also experimented with the use of deep neural networks, such as Multi-Headed Attention [47] encoders, as the basis for an alternate reconstitution function, in order to allow varying levels of reconstitution importance for each constitutive node. However, such networks are challenging to train as modelling reconstruction from an unordered set of constituent node/relation pairs is complex. Furthermore, it is not clear how we can retain the translational embedding relationships in such approaches. As such, we opt to use Equation 2, which is simple and effective for accessory node reconstitution in our experiments, and leave the exploration of alternative reconstruction functions for future work.

Optimization The optimization problem above is similar to the well-known P-Median Problem (PMP) [10], which selects $\mathcal{P}$ facilities such that the total cost of serving all locations is minimized. However, PMP covers only a basic scenario where each accessory entity must be reconstituted with only a single

semantically constitutive node, which is too limiting (as noted in Section 1). It is also not feasible to use a fixed number of constitutive nodes, simply because there may not be sufficient triplets in G to reconstitute each accessory node. Thus, we introduce "phantom" nodes, which "reconstitutes" any accessory node at a higher cost. These phantom nodes serve as padding nodes for the entities with low in-degree and are discarded after selection process is completed.

Given that all nodes in G are both facilities and locations, we allow facilities to serve themselves without cost, mirroring the memorization of retained constitutive node embeddings. We update the constraint on location assignment to allow exactly $\mathcal{G}$ facilities to serve the same location, mirroring accessory node reconstitution with multiple constitutive nodes, and introduce "free" facilities which serve locations that are also facilities at no cost.

Let $\mathcal{P}$ be the desired number of facilities, and $\mathcal{G}$ be the maximum number of constitutive nodes used to reconstitute a given accessory node. Given a set of locations $I = E$, the set of facilities J is defined as $J = I \cup P \cup F$, where $P = \{p_1, p_2, \ldots, p_{\mathcal{G}}\}$ and $F = \{f_1, f_2, \ldots, f_{\mathcal{G}-1}\}$ are the set of phantom and free nodes respectively with I, P, and F being mutually disjoint.

Let X be the facility assignment matrix, such that $X_{ij} = 1$ if location i is served by facility j, and 0 otherwise. Y is the facility opening matrix, such that $Y_j = 1$ if facility j is open, and 0 otherwise. C_{ij} denotes the cost of serving location i from facility j. For the nodes i and j from the knowledge graph G, such that $(i, r, j) \in T$ for some r, we define the cost consistently with the reconstitution function: $\mathrm{d}(\mathrm{H}(i) + \mathrm{H}(r), \mathrm{H}(j))$. If an entity pair has multiple relations, we select the relation that minimizes distance, and denote it as R_{ij}. Free facilities serve locations at no cost (i.e., $C_{ij} = 0$) for any $j \in F$. We arbitrarily set a high cost ($\alpha \geq 1$) for phantom nodes to encourage the preferential selection of real entities, and discard both free and phantom nodes post-selection.

$$C_{ij} = \begin{cases} \min\limits_{(j,r,i)\in T} \mathrm{d}(\mathrm{H}(j) + \mathrm{H}(r), \mathrm{H}(i)) & \text{if} \quad \exists r \in R : (i, r, j) \in T, \\ \alpha \max\limits_{(j,r,i)\in T} \mathrm{d}(\mathrm{H}(j) + \mathrm{H}(r), \mathrm{H}(i)) & \text{if} \quad j \in P, \\ 0 & \text{if} \quad i = j \quad \text{or} \quad j \in F, \\ +\infty & \text{otherwise}, \end{cases} \tag{3}$$

Since PMP is known to be NP-hard, we use an Integer Linear Programming (ILP) solver (i.e., Gurobi [17]) to find an approximate solution:

$$\min \sum_{i\in I} \sum_{j\in J} C_{ij} X_{ij} \qquad \text{subject to} \tag{4}$$

$$\sum_{j\in J} X_{ij} = \mathcal{G} \qquad \forall i \in I \tag{5}$$

$$\sum_{j\in I} Y_j = \mathcal{P} \tag{6}$$

$$X_{ij} \leq Y_j \qquad \forall i \in I, j \in J \tag{7}$$

$$X_{i\hat{j}} \geq Y_i \qquad \forall i \in I, \forall \hat{j} \in F \tag{8}$$

$$Y_j \in \{0, 1\} \text{ and } X_{ij} \in \{0, 1\} \qquad \forall i \in I, j \in J \tag{9}$$

Having a solution to the program above, we can generate the semantically constitutive graph $\hat{G} = (\hat{E}, \hat{R}, \hat{T})$ from X, Y, and R, where $\hat{E} = \{e \in I | Y_e = 1\}$, $\hat{R} = \{R_{ij} | i, j \in I\}$ and $\hat{T} = \{(h, R_{ht}, t) | h \in \hat{E}, t \in \hat{E}'\}$.

Approximation While it is possible to obtain an integer solution directly, we observed that a 2-step procedure achieves slightly better performance at the cost of marginally higher computational costs. We first solve a relaxed version of the problem where the Equation (9) is removed. This results in a partial solution $\bar{Y}$ containing fractional assignment of facilities. We replace the facility set J in the original program with a restricted set $\bar{J} = \{i : \bar{Y}_i \geq \epsilon\}$, and solve the new program directly. In our experiments, we default to the 2-step procedure, and set $\epsilon = 0.01$ to discard non-significant facilities.

Discussion We note that our problem definition is distinct from the Capacitated P Median Problem [15,35,45], which limits the number of locations allowed in each cluster. Our work, conversely, increases the number of clusters each location can belong to, and is therefore not comparable. We also note that phantom (P) and free nodes (F) are artificial constraints, and are removed in Y.

4 Experiments

Our experimental objective is to validate whether paying attention to the semantics in the selection of semantically constitutive entities within a knowledge graph would outperform baselines that focus primarily on structural centrality.

4.1 Experimental Setup

Datasets We experiment on publicly-available datasets (Table 1) which are common benchmarks for evaluating Knowledge Graph Embeddings.

Table 1: Dataset Summary

Dataset	# Entities	# Relations	# Training Triples	# Validation Triples	# Testing Triples
FB15k-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134
CoDEx-L	77,951	69	551,193	30,662	30,662

FB15k-237 FreeBase is a knowledge base containing general facts, and contains reversible (i.e., symmetric) relations. The FB15k-237 dataset [3,46] is a collection of FreeBase triples which retains only a single copy of reversible relation pairs, preventing information leakage during downstream evaluation.

WN18RR WordNet is a knowledge base consisting of different usages of a given word ("senses"), as well as the lexical relations between these "senses". The WN18RR dataset is selected from a collection of WordNet triples [3], where reversible relations have been removed in the same manner as FB15k-237 [12].

CoDEx Wikipedia is a crowdsourced encyclopedia that is openly edited. The CoDEx dataset is sampled from Wikipedia using a selection of seed entities and relations [43]. We use CoDEx-L, the largest version of CoDEx.

Baselines We compare our semantically constitutive nodes to nodes selected by the graph centrality approaches that focus on structural connectivity:

Point Centrality We expect that highly connected nodes are better suited for accessory node reconstitution as compared to low degree nodes, due to the larger number of possible reconstitutions. We calculate the degrees for all nodes

in each dataset, and select the top k nodes as a baseline. We experimented with using in-degrees (Point-In-Centrality) and out-degrees (Point-Out-Centrality) for selection, and observed that the latter generally performs better.

Group Centrality Point Centrality approaches prioritizes nodes within a dense subgraph at the expense of sparser nodes, as they are selected based on local network topology. We attempt to address this by selecting the nodes iteratively in a greedy fashion; after a node e_i is selected, we remove edges to/from e_i from the degree counts of the remaining nodes, stopping after we have selected k nodes or after all edges have been removed. In the latter case, we then randomly select nodes to ensure that there are k facilities. We note that this is similar to the SingleDiscount heuristic [7]. We report the results when using only in-degrees (Group-In-Centrality) and out-degrees (Group-Out-Centrality), as above.

Eigenvector Centrality We observe that the above baselines only consider the centrality of each node (i.e., degree), and places no weight on the influence of their neighbours. We therefore also compare to PageRank[2], which considers both the centrality as well as neighbouring influence when ranking node importance.

Influence Maximization We note that PageRank ranks nodes individually, and may therefore not return the best group of nodes. Our last baseline selects a group of nodes which maximizes the social influence of the group. As this is an NP-Hard problem [25], we use the SSA algorithm (Linear Threshold, $\epsilon = 0.03$, $\delta = 0.01$), which guarantees a $(1 - 1/e - \epsilon)$-approximate solution [20,37,38].

Embedding Models As our focus is on reconstruction, we obtain embeddings from the OpenKE implementation and suggested parameters for TransE, TransH and TransD. We target $\mathcal{P}$ to be a similar proportion (30%) of entities (7K for CoDEx-L, 4K for FB15k-237, 9K for WN18RR) in all following experiments.

4.2 Quantitative Comparisons

A measure of quality is the ability of the selected nodes to retain the semantic meaning of accessory nodes. As we use knowledge graph embeddings to represent node semantics, we turn to knowledge graph embedding evaluation tasks.

Link Prediction Knowledge graph embedding quality is commonly compared via downstream task such as the well-known Link Prediction task. We form embeddings for each node selection by replacing the embedding for discarded entities with the reconstituted embedding. We use the (filtered) Link Prediction Task [3,53]. Given a true testing triple $(\hat{h}, \hat{r}, \hat{t})$, we wish to rank $\hat{t}$ given $(\hat{h}, \hat{r})$ amongst the set of testing entities $\tilde{E}$ (or $\hat{h}$ given $(\hat{r}, \hat{t})$).

Table 2 shows the experimental results for each dataset, where Hit@10% (of entities in the dataset; similar results observed for Hit@5%) is used as metric to facilitate comparison between differently-sized datasets. The first line is the performance of the original (i.e., full-sized) TransE embeddings. Subsequent lines are performances (relative to the original, in percentage) of each selection method. Semantically-Constitutive consistently achieves a higher Hit@10% as compared to the baselines in all cases for FB15k-237 and WN18RR. For CoDEx-L, Semantically-Constitutive outperforms most baselines, tying with one.

[2] Adapted from https://github.com/louridas/pagerank, $a = 0.85, c = 1 \times 10^{-32}$

Table 2: Link Prediction Task Hit@10%, Relative % to Original (TransE Embeddings, Higher is Better)

Model	FB15k-237	WN18RR	CoDEx-L
Original	0.968	0.755	0.989
Point-In-Centrality	80.5	35.0	7.3
Point-Out-Centrality	86.1	37.3	24.9
Group-In-Centrality	78.7	40.0	7.2
Group-Out-Centrality	86.6	41.9	**25.1**
SSA	71.1	25.4	14.3
PageRank	77.1	39.3	6.1
Semantically-Constitutive	**87.9**	**43.2**	**25.1**

Multiple Node Reconstitution We now study the effect of multiple nodes for reconstitution, which is controlled by the parameter $\mathcal{G}$. We expect that larger $\mathcal{G}$ allows reconstitutions to better capture the semantic meaning of the accessory entity, as shown in Section 1. We conduct an ablation study for each dataset, by reducing the number of reconstitution nodes allowed (from $\mathcal{G} = 10$) from the same partial solution $\bar{Y}$ (as described in Section 3), and repeat the Link Prediction task with the resulting reconstitutions (Table 3).

We observe that while the Hit@10% generally remains fairly consistent as $\mathcal{G}$ is reduced for all models, small but noticeable differences in performance can be observed. For example, in FB15k-237, Semantically-Constitutive outperforms all baselines at every $\mathcal{G}$ level. Next, we observe that the best performance for Semantically-Constitutive is at $\mathcal{G}-4$ (88.29%). This suggests that the number of constitutive nodes $\mathcal{G}$ can be tuned to best utilize the selected semantically constitutive nodes, improving downstream performance. WN18RR shows similar improvements ($\mathcal{G}-2$, 43.29%), but CoDEx-L performance is flat across $\mathcal{G}$.

Embedding Models Lastly, although not the focus of our work, we study the generalizability of our approach. We replace the entity embeddings with the encoding representations from translational knowledge graph embedding models such as TransH and TransD, and show the Hit@10% for WN18RR in Table 4(consistent results are observed for other datasets).

First, we observe that among the original embeddings, TransE performs the best (0.755), while TransD (0.738) is able to outperform TransH (0.723). We speculate that this is related to the choice of encoding representation function f, which does not fully capture the translation operation in the TransH and TransD training processes, and leave the selection of a suitable f as future work.

Turning to the baseline approaches, we observe that all models generally perform at similar relative levels across the embedding models. Group-Out-Centrality, the best performing baseline, achieves the best baseline performance on TransE (41.90), similar to the performance of the full embeddings.

Lastly, we observe that while Semantically-Constitutive shows a similar drop in relative performance on TransH (42.86) as compared to TransE (43.21), it was able to achieve a minor improvement on TransD (43.47). We note that the absolute performance of Semantically-Constitutive is still higher on TransE.

Table 3: $\mathcal{G}$ Reduction Hit@10%, Relative % to Original (TransE Embeddings, Higher is Better)

Model	$\mathcal{G} = 10$	$\mathcal{G}-1$	$\mathcal{G}-2$	$\mathcal{G}-3$	$\mathcal{G}-4$
FB15k-237					
Original	0.968				
Point-In-Centrality	80.5	80.5	80.5	80.4	80.4
Point-Out-Centrality	86.1	86.1	86.1	86.1	86.1
Group-In-Centrality	78.7	78.7	78.7	78.7	78.7
Group-Out-Centrality	86.6	86.7	86.7	86.7	86.7
SSA	71.1	70.9	70.9	70.0	70.9
PageRank	77.1	77.1	77.2	77.1	77.1
Semantically-Constitutive	**87.9**	**88.0**	**88.1**	**88.2**	**88.3**
WN18RR					
Original	0.968				
Point-In-Centrality	35.0	35.0	35.0	35.0	35.0
Point-Out-Centrality	37.3	37.3	37.3	37.0	37.0
Group-In-Centrality	40.0	40.0	40.0	40.0	40.0
Group-Out-Centrality	41.9	41.9	41.9	41.9	42.0
SSA	25.4	25.4	25.4	25.5	25.4
PageRank	39.3	39.3	39.3	39.3	39.3
Semantically-Constitutive	**43.2**	**43.2**	**43.3**	**43.2**	**43.1**
CoDEx-L					
Original	0.968				
Point-In-Centrality	7.3	7.3	7.3	7.3	7.3
Point-Out-Centrality	24.9	24.9	24.9	24.9	24.9
Group-In-Centrality	7.2	7.2	7.2	7.2	7.2
Group-Out-Centrality	**25.1**	**25.1**	**25.1**	**25.1**	**25.1**
SSA	14.3	14.3	14.3	14.3	14.3
PageRank	6.1	6.1	6.1	6.1	6.1
Semantically-Constitutive	**25.1**	**25.1**	**25.1**	**25.1**	**25.1**

Effect of Graph Size on Runtime. We study the effects of knowledge graph size on runtime between Semantically-Constitutive and the "Direct" one-step program. We sample (from 23,616 unique) CoDEx-L tail entities to between 23,000 and 14,000 (with intervals of 1,000), and retain only triples containing those sampled tail entities. We then run both "Direct" and Semantically-Constitutive on these sub-graphs, and set $\mathcal{P}$ to 30% of the number of sampled entities to ensure consistent difficulty. We report the mean runtimes on 5 samples in Figure 2 (labelled with the initial sample sizes, from 23K to 14K). As discussed in Section 3, we observe that Semantically-Constitutive achieves a general improvement in model performance at the expense of slightly longer runtimes, particularly at smaller sampled sizes.

4.3 User Study

We conducted a user study to investigate the real-world informativeness of Semantically-Constitutive, and expect Semantically-Constitutive to provide reconstructions (i.e., relation between accessory and constitutive nodes) with higher relevance due to the semantic reconstruction process. We first filter the CoDEx-L dataset to retain only entities that have at least 10 unique edges. We then compare Group-Out-Centrality (best performing baseline) to Semantically-Constitutive

Table 4: Translational Knowledge Graph Embedding Hit@10%, Relative % to Original (WN18RR, Higher is Better)

Model	TransE	TransH	TransD
Original	0.755	0.723	0.738
Point-In-Centrality	35.0	35.8	33.9
Point-Out-Centrality	37.3	38.7	38.2
Group-In-Centrality	40.0	39.2	40.9
Group-Out-Centrality	41.9	41.3	41.5
SSA	25.4	25.0	25.1
PageRank	39.3	42.3	41.2
Semantically-Constitutive	**43.2**	**42.9**	**43.5**

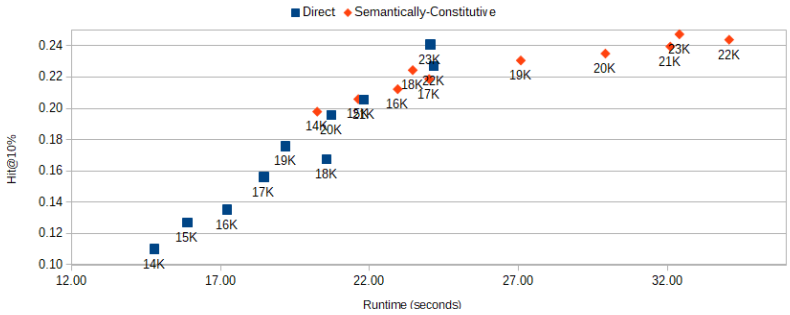


Fig. 2: Model Runtime and Hit@10% on Sampled CODEXL

(TransE embeddings, $\mathcal{G}$=10, $\hat{\mathcal{G}}$=3), and select accessory nodes where the triplet relation is "occupation"[3] and all 3 constitutive nodes differ. We randomly select 20 (from 45 total) such accessory nodes for the user study.

Each user was presented a accessory node (e.g., "Film Actor" in Table 5a) in each question[4], and asked to rank the relevance of all 6 constitutive nodes (supplemented with their Codex-L description) on a five-level Likert Scale [29]. We compare the collected responses by assigning a score between -1 and 1 to each level (Table 5b). Figure 3 shows the average score by 13 users[5] for Group-Out-Centrality (mean = 0.103) and Semantically-Constitutive (mean = 0.458) on each question. We observe that Semantically-Constitutive (in red) generally achieves a higher average score on all questions as compared to Group-Out-Centrality (in blue), showing that the nodes selected by Semantically-Constitutive is better related to the query occupation, and are therefore more informative.

Inter-rater Reliability Next, we wish to study the agreement between different raters. Fleiss' Kappa [27] is commonly used for understanding the inter-rater reliability of ordinal rating data, and range from -1 to 1, with values above 0 indicate agreement (beyond chance) between the raters.

[3] Selected in order to limit the obscurity of triplets in the user study

[4] The order of questions and Likert items were randomized for every user.

[5] This was the number of study participants who agreed to take part in the study. They were neither co-authors, nor aware of the subject of this paper.

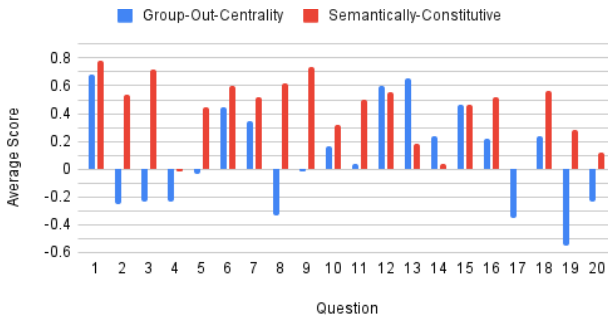


Fig. 3: User Study Scores

Table 5: Example User Study Question ("Film Actor")

(a) Likert Items (Constitutive Node + Codex-L description)

Semantically-Constitutive	Group-Out-Centrality
Robin Williams (American actor and stand up comedian (1951-2014))	John Cale (Welsh composer, singer-songwriter and record producer)
Justin Timberlake (American singer, record producer, and actor)	John Lennon (English singer and songwriter, founding member of The Beatles)
Nicolas Cage (American actor)	A. R. Rahman (Indian singer and composer)

(b) Ranking Options and Associated Score

Option	Relevant	Somewhat relevant	Neither relevant or irrelevant	Somewhat irrelevant	Irrelevant
Score	1	0.5	0	-0.5	-1

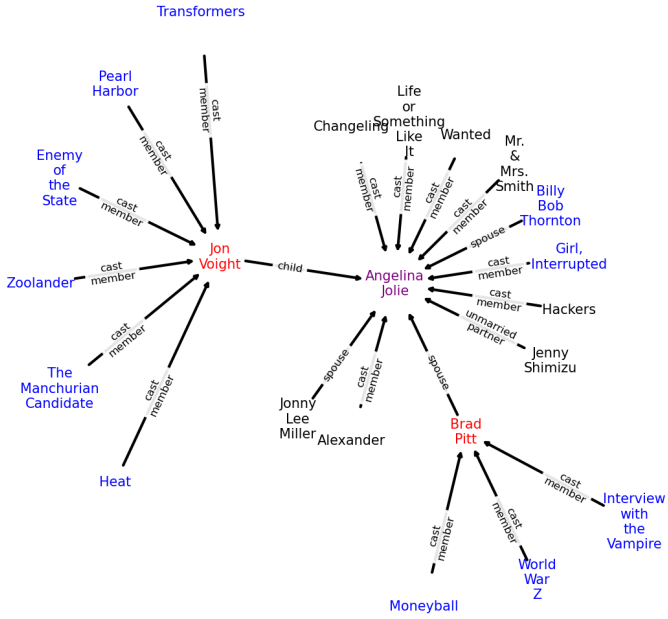
We first combine all "Relevant" and "Somewhat relevant" responses, and do the same for "Neither relevant or irrelevant", "Somewhat irrelevant" and "Irrelevant" classes. Next, we calculate the (2-Rater 2-Class) Fleiss' Kappa for each pair of raters, and average them. The expected Fleiss' Kappa in this setting is 0.245, which suggests that a random pair of raters would likely show fair agreement (as suggested by [27]) on the (binary) relevance of each choice.

Next, we investigate the overall reliability of the User Study. We report that the Multiple-Rater 5-Class Fleiss' Kappa ($0.119 > 0$), indicates that there is likely to be agreement amongst the raters.

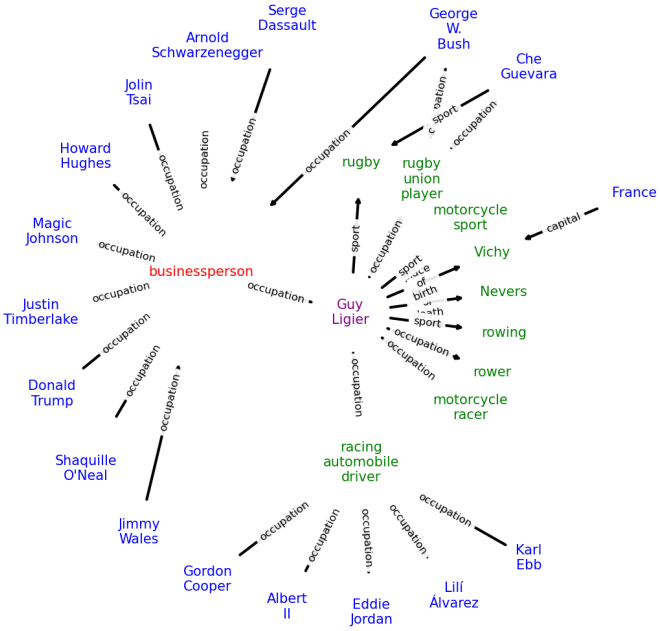
4.4 Case Studies

Next, we show 2 subgraphs generated by Semantically-Constitutive on Codex-L. Figure 4a shows the reconstruction of an accessory node `{Angelina Jolie}`, from two constitutive nodes, (`{Girl, Interrupted}` and `{Billy Bob Thornton}`). Other triplets involving accessory nodes such as `{Mr. & Mrs. Smith}` are discarded from the full knowledge graph. We also show accessory nodes that are reconstituted by other constitutive nodes, such as `{Brad Pitt}` being reconstituted by `{Moneyball}`, `{World War Z}`, and `{Interview with the Vampire}`. This subgraph shows how Semantically-Constitutive reconstitutes (specific) nodes by combining multiple more general relations, such as being cast in a movie.

Figure 4b shows a subgraph from the CoDEx-L dataset, centered on the node representing the constitutive node `{Guy Ligier}`. We also show nodes which are



(a) Subgraph Centered on Accessory Node `Angelina Jolie`



(b) Subgraph Centered on Constitutive Node `Guy Ligier`

Fig. 4: Codex-L Case Studies

reconstituted by {Guy Ligier}, other retained nodes, and reconstituted nodes from these retained nodes. From Figure 4b, we can infer that {Guy Ligier} was probably involved in rowing ({occupation}→{rowing}), racing ({occupation} →{motorcycle racer}), and rugby ({occupation}→{rugby union player}), Note that while {Guy Ligier} is used to reconstitute {racing automobile driver}, this reconstitution is in conjunction with other constitutive nodes such as {Karl Ebb} and {Eddie Jordan}, suggesting that the concept of {racing automobile driver} is not fully captured by a single node. Next, we observe that while a relation exists between {Guy Ligier} and the accessory node {businessperson}, it is reconstituted by {Howard Hughes} and {Donald Trump}, who may be relatively better recognized as businesspersons, instead.

5 Conclusion

In this work, we identify semantically constitutive entities in a knowledge graph (KG). Intuitively, embeddings of "constitutive" nodes can be used to reconstitute "accessory" nodes, and is based on actual KG triples, providing credence and interpretability. Experiments on three knowledge bases validate the proposed methodology in several ways. On the downstream Link Prediction task, our method outperforms structural connectivity baselines. A user study validates our reconstitutions as more consistent with human evaluation.

One limitation of our work is a reliance on pretrained graph embedding as input, as our approach is unable to generate these embeddings from the knowledge graph directly. Next, reconstructions are only as accurate as the KG provided; problematic reconstructions can be avoided by auditing the underlying KG.

One future direction is to adapt our approach to non-translational KG embeddings. Another is to explore how semantically constitutive entities could enhance related tasks such as KG summarization.

References

1. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of mathematical sociology **2**(1), 113–120 (1972)
2. Bonacich, P.: Power and centrality: A family of measures. American journal of sociology **92**(5), 1170–1182 (1987)
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. NeurIPS **26** (2013)
4. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. Social networks **28**(4), 466–484 (2006)
5. Buehrer, G., Chellapilla, K.: A scalable pattern mining approach to web graph compression with communities. In: WSDM. pp. 95–106 (2008)
6. Chen, C., Lin, C.X., Fredrikson, M., Christodorescu, M., Yan, X., Han, J.: Mining graph patterns efficiently via randomized summaries. PVLDB **2**(1), 742–753 (2009)
7. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: KDD. pp. 199–208 (2009)
8. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PloS one **10**(6), e0128193 (2015)
9. Cook, D.J., Holder, L.B.: Substructure discovery using minimum description length and background knowledge. JAIR **1**, 231–255 (1993)

10. Cornuéjols, G., Nemhauser, G., Wolsey, L.: The uncapacitated facility location problem. Tech. rep., Cornell University Operations Research and Industrial Engineering (1983)
11. Dai, D., Zheng, H., Luo, F., Yang, P., Chang, B., Sui, Z.: Inductively representing out-of-knowledge-graph entities by optimal estimation under translational assumptions. arXiv preprint arXiv:2009.12765 (2020)
12. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings. In: AAAI. No. 1 (2018)
13. Dunne, C., Shneiderman, B.: Motif simplification: improving network visualization readability with fan, connector, and clique glyphs. In: CHI. pp. 3247–3256 (2013)
14. Fellbaum, C.: Wordnet. In: Theory and applications of ontology: computer applications, pp. 231–243. Springer (2010)
15. Fleszar, K., Hindi, K.S.: An effective vns for the capacitated p-median problem. European Journal of Operational Research **191**(3), 612–622 (2008)
16. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry pp. 35–41 (1977)
17. Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2022)
18. Hamann, F., Ulges, A., Krechel, D., Bergmann, R.: Open-world knowledge graph completion benchmarks for knowledge discovery. In: IEA/AIE. pp. 252–264 (2021)
19. Hirsch, J.E.: An index to quantify an individual's scientific research output. PNAS **102**(46), 16569–16572 (2005)
20. Huang, K., Wang, S., Bevilacqua, G., Xiao, X., Lakshmanan, L.V.: Revisiting the stop-and-stare algorithms for influence maximization. PVLDB **10**(9), 913–924 (2017)
21. Huang, X., Zhang, J., Li, D., Li, P.: Knowledge graph embedding based question answering. In: WSDM. pp. 105–113 (2019)
22. Jenatton, R., Roux, N., Bordes, A., Obozinski, G.R.: A latent factor model for highly multi-relational data. NeurIPS **25** (2012)
23. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: COLING-IJCNLP. pp. 687–696 (2015)
24. Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y.: A survey on knowledge graphs: Representation, acquisition, and applications. TNNLS **33**(2), 494–514 (2021)
25. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD. pp. 137–146 (2003)
26. Korn, A., Schubert, A., Telcs, A.: Lobby index in networks. Physica A: Statistical Mechanics and its Applications **388**(11), 2221–2226 (2009)
27. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
28. Li, C.T., Lin, S.D.: Egocentric information abstraction for heterogeneous social networks. In: ASONAM. pp. 255–260. IEEE (2009)
29. Likert, R.: A technique for the measurement of attitudes. Arch. psychol (1932)
30. Liu, S., Grau, B., Horrocks, I., Kostylev, E.: Indigo: Gnn-based inductive knowledge graph completion using pair-wise encoding. NeurIPS **34** (2021)
31. Liu, Y., Safavi, T., Dighe, A., Koutra, D.: Graph summarization methods and applications: A survey. CSUR **51**(3), 1–34 (2018)
32. Maccioni, A., Abadi, D.J.: Scalable pattern matching over compressed graphs via dedensification. In: KDD. pp. 1755–1764 (2016)
33. Mahdisoltani, F., Biega, J., Suchanek, F.: Yago3: A knowledge base from multilingual wikipedias. In: CIDR (2014)

34. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., et al.: Never-ending learning. Communications of the ACM **61**(5), 103–115 (2018)
35. Mulvey, J.M., Beck, M.P.: Solving capacitated clustering problems. European Journal of Operational Research **18**(3), 339–348 (1984)
36. Nguyen, D.Q., Nguyen, T.D., Nguyen, D.Q., Phung, D.: A novel embedding model for knowledge base completion based on convolutional neural network. arXiv preprint arXiv:1712.02121 (2017)
37. Nguyen, H.T., Dinh, T.N., Thai, M.T.: Revisiting of 'revisiting the stop-and-stare algorithms for influence maximization'. In: COSN. pp. 273–285 (2018)
38. Nguyen, H.T., Thai, M.T., Dinh, T.N.: Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In: SIGMOD. pp. 695–710 (2016)
39. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: ICML (2011)
40. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
41. Purohit, M., Prakash, B.A., Kang, C., Zhang, Y., Subrahmanian, V.: Fast influence-based coarsening for large networks. In: KDD. pp. 1296–1305 (2014)
42. Riondato, M., García-Soriano, D., Bonchi, F.: Graph summarization with quality guarantees. DMKD **31**(2), 314–349 (2017)
43. Safavi, T., Koutra, D.: CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In: EMNLP. pp. 8328–8350 (Nov 2020)
44. Shen, Z., Ma, K.L., Eliassi-Rad, T.: Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE TVCG **12**(6), 1427–1439 (2006)
45. Stefanello, F., de Araújo, O.C., Müller, F.M.: Matheuristics for the capacitated p-median problem. ITOR **22**(1), 149–167 (2015)
46. Toutanova, K., Chen, D.: Observed versus latent features for knowledge base and text inference. In: CVSC. pp. 57–66 (2015)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS **30** (2017)
48. Wan, Z., Mahajan, Y., Kang, B.W., Moore, T.J., Cho, J.H.: A survey on centrality metrics and their network resilience analysis. IEEE Access **9**, 104773–104819 (2021)
49. Wang, P., Han, J., Li, C., Pan, R.: Logic attention based neighborhood aggregation for inductive knowledge graph embedding. In: AAAI. vol. 33, pp. 7152–7159 (2019)
50. Wang, Q., Huang, P., Wang, H., Dai, S., Jiang, W., Liu, J., Lyu, Y., Zhu, Y., Wu, H.: Coke: Contextualized knowledge graph embedding. arXiv preprint arXiv:1911.02168 (2019)
51. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. TKDD **29**(12), 2724–2743 (2017)
52. Wang, X., He, X., Cao, Y., Liu, M., Chua, T.S.: Kgat: Knowledge graph attention network for recommendation. In: KDD. pp. 950–958 (2019)
53. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: AAAI. vol. 28 (2014)
54. Yao, L., Mao, C., Luo, Y.: Kg-bert: Bert for knowledge graph completion. arXiv preprint arXiv:1909.03193 (2019)
55. Zhang, N., Tian, Y., Patel, J.M.: Discovery-driven graph summarization. In: ICDE. pp. 880–891. IEEE (2010)
56. Zhu, L., Ghasemi-Gol, M., Szekely, P., Galstyan, A., Knoblock, C.A.: Unsupervised entity resolution on multi-type graphs. In: ISWC. pp. 649–667 (2016)