The Measurement of Knowledge in Knowledge Graphs

Jason Xiaotian Dou[1], Haiyi Mao[1], Runxue Bao[1], Paul Pu Liang[2], Xiaoqing Tan[1], Shiyi Zhang[2], Minxue Jia[1], Pengfei Zhou[1], Zhi-Hong Mao[1]

[1] University of Pittsburgh
[2] Carnegie Mellon University
jason.dou@pitt.edu

Abstract

Knowledge graphs (KG) have emerged as comprehensive tools to enable knowledge discovery. Knowledge representation learning learns knowledge graph embeddings, which are extremely useful feature inputs for a wide variety of prediction and graph analysis tasks. Due to the diverse and subjective nature of KG, KG evaluation becomes an important and open problem. Existing evaluation methods are often based on case studies or downstream tasks like information retrieval and question answering systems, which is either lack generalizability or hard to implement. To address the challenges above, we propose three direct metrics for the knowledge graph, the K_Score, I_Score, and C_Score, derived from the science of science, information theory, and causality perspectives, respectively. We propose a human-centered approach to evaluate the effectiveness of our metrics. Through a pilot study, we share insights on the complications of human-centered evaluation and motivate future work.

Introduction

Knowledge graphs (KG) encode the real world to support domains and applications like computer vision, natural language processing, recommendation system, information retrieval, and so forth. KG can provide a holistic view of knowledge (Kamdar 2019; Hegel 2018) to enable scientific discovery (Chandak, Huang, and Zitnik 2022; Sun et al. 2019; Du et al. 2017; Liu et al. 2022; Ren et al. 2022; Mao, Broerman, and Benos 2020; Rosas et al. 2022; Paudel et al. 2022; Mao and Dou 2022), and on the other hand, incorporates expert knowledge into machine learning framework (Li et al. 2022; Qian et al. 2021; Mao et al. 2022a; Dou et al. 2022; Dou, Luo, and Yang 2022; Mao et al. 2022b; Tan et al. 2022a; Yang et al. 2022; Zeng et al. 2022; Mao and Dou 2022; Tang et al. 2022; Shi et al. 2022; Tang et al. 2023).

WordNet (Miller 1998) (Figure 1), VisualGenome (Krishna et al. 2017) (Figure 2), and PrimeKG (Chandak, Huang, and Zitnik 2022) (Figure 3) are three representative knowledge graphs we consider in this paper. WordNet is a large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet's structure makes it a powerful tool for downstream tasks in natural language processing (Majewska et al. 2021).

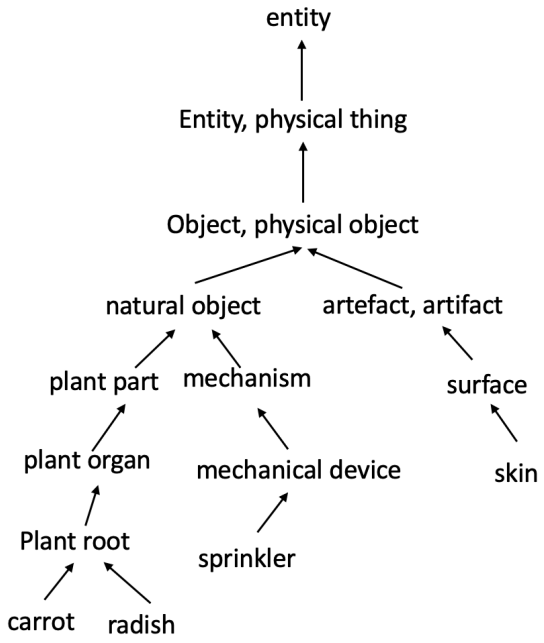

Figure 1: A WordNet example (Miller 1998).

VisualGenome is among the first to provide detailed labeling of object interactions and attributes, grounding visual concepts in language. It enables the modeling of relations between objects and images. Question answering task is used to evaluate VisualGenome's performance. PrimeKG is a precision medicine-oriented knowledge graph that provides a holistic view of diseases. It supports artificial intelligence analyses of how drugs target disease-related molecular perturbations by unique drug-disease edges. Multimodal analyses (Tsai et al. 2019) can be enabled by PrimeKG's graph structure. Figure 3 shows a PrimeKG example. A case study in autism is conducted to evaluate the relevance of PrimeKG to the clinical presentation of autism. Computational approaches are used to group disease nodes. Although there is considerable interest in the construction and applications of KG, the evaluation is challenging. As we mentioned in the last paragraph, downstream tasks and case studies are

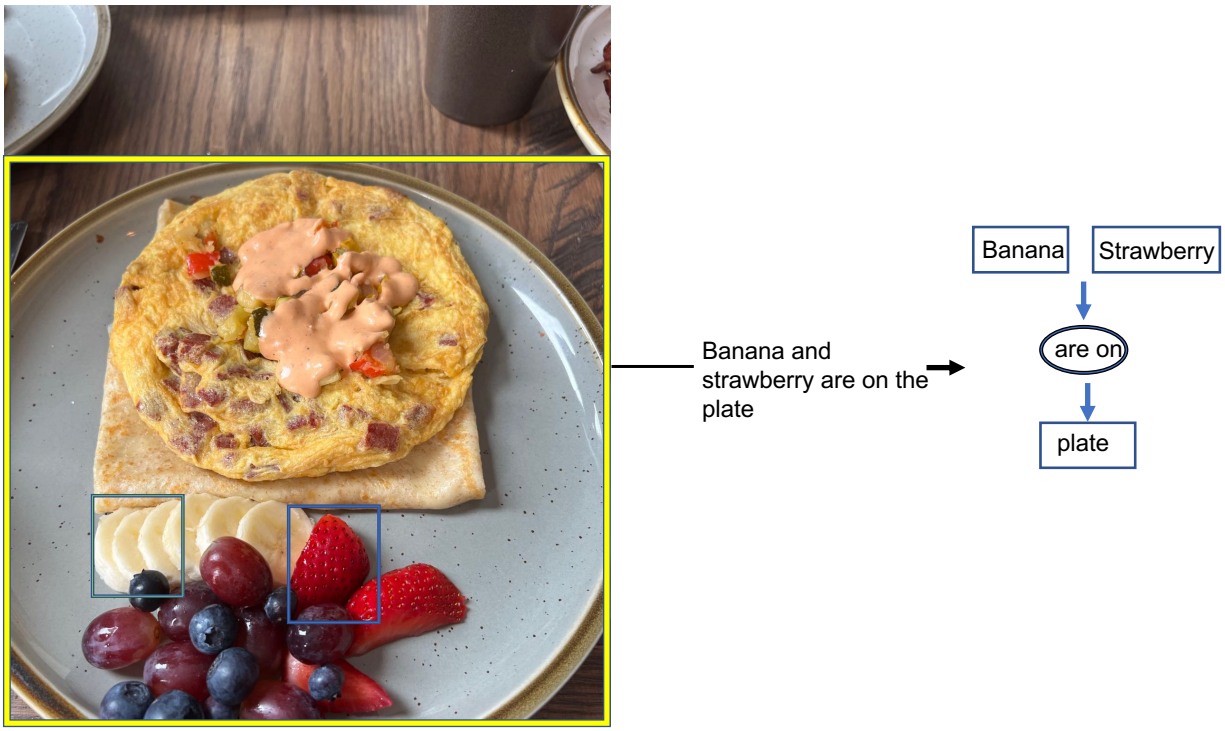

Figure 2: A VisualGenome example inspired by (Krishna et al. 2017).

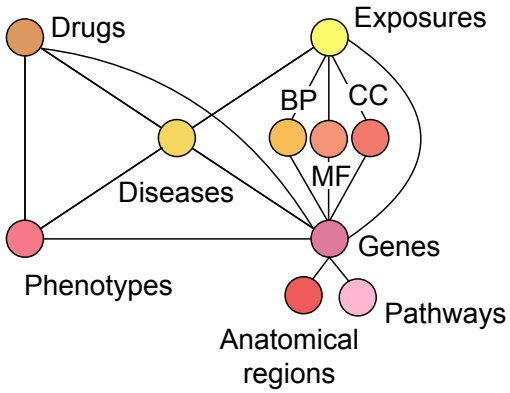

Figure 3: A biomedical knowledge graph example from PrimeKG (Chandak, Huang, and Zitnik 2022).

used to evaluate knowledge graphs like VisualGenome and PrimeKG, while there are no principled metrics to measure the "knowledge" in knowledge graphs. Thus we explore this space and propose three metrics: the K_Score, based on the literature of Science of Science; the I_Score originated from the information theory; and the C_Score from the causal perspective.

The K_Score is our initial attempt to measure knowledge in knowledge graphs. The insightful human-centered intuition mediates the current research landscape that human-centric metrics have been largely under-served in terms of research and development of machine learning, and specifically representation learning. Then based on the limitations of the K_Score, we propose a more mathematically rigorous I_Score to address issues the K_Score may encounter. Last but not least, the C_Score measures the causal information of a knowledge graph by a corresponding causal graph. This is specifically impactful in biology and medicine since the causal mechanism of biological processes is the holy grail of biomedical inquiry.

In the rest of the paper, we introduce the proposed metrics in detail in Section 2. A human-centered pilot study to evaluate the three metrics is summarized in Section 3. The conclusion and future work are provided in Section 4.

Proposed Metrics

Science of Science Approach: the K-Score

The science of science as an academic discipline offers a quantitative understanding of the interactions among scientific agents across diverse geographic and temporal scales (Fortunato et al. 2018; Dou 2017; Dou et al. 2017; Dou, Sun, and Zou 2016; Tan et al. 2022b; Mo et al. 2022). On a relevant note, Weis and Jacobson (2021) use knowledge graph dynamics to provide an early-warning signal for impactful research by learning high-dimensional relationships among features calculated across time from the scientific literature.

Wu, Wang, and Evans (2019) proposes a disruption measurement D to assess the difference between the observed number of papers divided by the number of all subsequent works. Subsequently, Xu, Wu, and Evans (2022) explores the relation between team structure and the character of

knowledge they produce. It uses a lead (or L)-ratio to characterize the fraction of Lead authors in a team.

Inspired by the two metrics above, we propose the K_Score measure the "knowledge" a knowledge graph conveys. The initial idea of the personalized measurement of knowledge in KG is as the following:

[The Knowledge Score] Suppose we have a total of n nodes with n_u unknown nodes, and e edges with e_u unknown edges. The context parameter of the node and the edge is defined as c. Then the Knowledge Score, K_Score is formulated as:

$$K\_Score = c * (\frac{n_u}{n} + \frac{e_u}{e}). \tag{1}$$

Here we explain why we call the K_Score a personalized knowledge score. When we look at the same knowledge graph with 40 nodes and 80 edges, different people have different existing knowledge and different context argument. For example, we assume, for a PhD student in machine learning, there are 30 unknown nodes, 60 unknown edges, and the context parameter as 2. Then we have the K_Score as the following:

$$K\_Score = 2 * (\frac{30}{40} + \frac{60}{80}) = 3.00. \tag{2}$$

While for a biomedical expert, the unknown nodes 5 and edges 10 can be much less, while the existing knowledge base will enable a much bigger context parameter 10. Then we have the K_Score as the following:

$$K\_Score = 10 * (\frac{5}{40} + \frac{10}{80}) = 2.50. \tag{3}$$

Although carrying intuitive insights, the above formulation is very preliminary. One major drawback is that we did not consider the size of the nodes and edges of the KG, which could be an important factor of "knowledge". Another challenge is to take into account the dynamic interactions between edges and nodes. Also, we develop the K_Score mainly having the knowledge graphs like PrimeKG in mind. How to interpret multi-modality knowledge graphs like VisualGenome within the K_Score is also among the following challenges we want to tackle.

Information Theory Approach: the I-Score

Although the K_Score is intuitive and interpretable, as mentioned above, it is far from complete. We start this subsection by giving an example where the K_Score fails. Recall the Knowledge Score K_Score above and consider the following case, where we have two biomedical knowledge graphs with edges and nodes: KG_1 and KG_2.

- For the KG_1 with 20 nodes and 60 edges, we have 15 unknown nodes and 45 unknown edges.
- For the KG_2 with 40 nodes and 80 edges, we have 30 unknown nodes and 60 unknown edges.

Although we gain more knowledge from KG_2 than KG_1, we notice that the K_score is the same for the two cases. This demonstrates that K_score can evaluate the intra-knowledge of the knowledge graphs but is limited in generating comparable measurements across the graphs.

The information theory (Shannon 1948) is the mathematical treatment of the concepts, parameters and rules governing the transmission of messages through communication systems. it studies the quantification and communication of digital information and thus it can be used to measure the knowledge of knowledge graphs. To evaluate the inter-knowledge of different graphs, inspired by the information theory, we propose a method to assess the uncertainty of the graph to measure the knowledge.

Specifically, let X be a random variable of the nodes taking on a finite number M of different values $x_1, \cdots, x_M$ with probability $p_1, \cdots, p_M, p_i > 0$, such that

$$\sum_{i=1}^{M} p_i = 1. \tag{4}$$

and Y be a random variable of the edges taking on a finite number N of different values $y_1, \cdots, y_N$ with probability $q_1, \cdots, q_N, q_i > 0$ such that

$$\sum_{i=1}^{N} q_i = 1. \tag{5}$$

The knowledge we gain from the KG is desired to be a function of $p_1, \cdots, p_M$ and $q_1, \cdots, q_N$.

[The Informative Score] Suppose we have total n nodes with the set S_n of unknown nodes with probability

$$p_i > 0, i \in S_n. \tag{6}$$

and total e edges with the set S_e of unknown edges with probability

$$q_i > 0, i \in S_e. \tag{7}$$

the importance of the node and the edge is defined as c_n and c_e, then the Informative Score, the I_Score can be formulated as:

$$I\_Score = c_n * \sum_{i \in S_n} p_i \log \frac{1}{p_i} + c_e * \sum_{i \in S_e} q_i \log \frac{1}{q_i}. \tag{8}$$

With such a definition, our score has the following properties:

[Additivity] The I_Score can be divided into two parts as

$$I\_Score = H(p_1, \cdots, p_M) + H(q_1, \cdots, q_N), \tag{9}$$

where each $H(\cdot)$ is an entropy-like function. Hence, the I_Score can be further represented as

$$I\_Score = \sum_{i=1}^{M} H(p_i) + \sum_{i=1}^{N} H(q_i). \tag{10}$$

[Monotonicity] The I_Score takes the size of the KG into account and thus can measure the inter-knowledge of the graphs. Specifically, let

$$f(M) = H(\frac{1}{M}, \cdots, \frac{1}{M}). \tag{11}$$

If $M < M'$, then

$$f(M) < f(M'). \tag{12}$$

Consider the case that the K_Score cannot tell the difference above, we choose

$$c_n = c_e = 0.5. \tag{13}$$

and assume the probability to be equal here for simplicity. Then back to the KG_1 and KG_2 in the case we mentioned at the beginning of the subsection, we have the I_Scores as as the following:

$$I\_Score_{KG_1} = 0.5 * \frac{15}{20}\log 20 + 0.5 * \frac{45}{60}\log 60 = 1.16, \tag{14}$$

$$I\_score_{KG_2} = 0.5 * \frac{30}{40}\log 40 + 0.5 * \frac{60}{80}\log 80 = 1.31. \tag{15}$$

With the monotonicity, the I_Score can effectively measures the knowledge across the graphs with different sizes. From this angle, the I_Score has better properties and measures knowledge in knowledge graphs more precisely over the K_Score.

Measuring Causal Informative Score: the C-Score

Finding underlying causal relations is a fundamental task in many disciplines, including economics, biology, and medicine. Knowledge graphs organically comprise causal information (Nordon et al. 2019). Recent work (Domingo-Fernández et al. 2022) has shown that causal reasoning over knowledge graphs can accelerate the drug discovery process. However, most evaluation metrics do not take how much causal information a knowledge graph may contain into account. In this section we take a causal perspective (Pearl 2009) to the knowledge measurement problem. We first bridge the knowledge graphs and causal graphs (Pearl 2009; Spirtes et al. 2000) with overlapping edges. Then we introduce the concepts of the causal informative score (C_Score) and $\mathcal{P}-$ causal informative score ($\mathcal{P}-$C_Score) to evaluate the causal information in the learned knowledge graph.

In the beginning, we present notations and definitions of our approach. We assume the causal relationships can be encoded as a graph $\mathcal{G} = \{V_C, E_C\}$. Each vertex $n \in V_C$ in $\mathcal{G}$ represents a random variable. Furthermore, the causal graphs could be derived from observational data with PC algorithm (Spirtes et al. 2000) or expert knowledge. We do not limit the causal graphs as a Directed Acyclic Graph (DAG) in this paper. Meanwhile, we have a *directed* or *partial directed* knowledge graph $KG = \{V_{KG}, E_{KG}\}$. Then we define *causal sufficient* KG, *causal efficient* KG, and *causal equivalent* KG. It's worth noting that in the knowledge graph there are different edges, but we only examine the direct edges which could be positive or negative correlations. In order to illustrate the concepts we propose, we first have the following definitions.

[Causal Sufficient Knowledge Graph] If a knowledge graph KG,

$$\forall e \in E_{KG} \implies \exists e' \in E_C. \tag{16}$$

then we say KG is causal sufficient.

[Causal Efficient Knowledge Graph] If a knowledge graph KG,

$$\forall e \in E_C \implies \exists e' \in E_{KG}. \tag{17}$$

then we say KG is causal efficient.

[Causal Equivalent Knowledge Graph] If a knowledge graph KG is simultaneously causal sufficient and efficient regarding a casual graph $\mathcal{G}$, we say KG is causal equivalent to $\mathcal{G}$. Next, we define similar concepts on sub graphs.

[Causal Sufficient Knowledge Sub-Graph] if a sub knowledge graph

$$KG_s \subset KG. \tag{18}$$

are causal sufficient to a sub graph $\mathcal{G}$, we say KG_s is causal sub-sufficient to $\mathcal{G}$.

[Causal Efficient Knowledge Sub-Graph] if a sub knowledge graph

$$KG_s \subset \mathcal{KG}. \tag{19}$$

are causal efficient to a sub graph $\mathcal{G}$, we say KG_s is causal sub-efficient to $\mathcal{G}$. Knowledge graphs provide a large number of covariates that are not contained in causal graphs(Nordon et al. 2019). Thus, an edge in a causal graph could correspond to a path in the knowledge graph (Nordon et al. 2019). For example, when we apply PC algorithm (Spirtes et al. 2000) to observational data to obtain a causal graph where exists an edge $A \to C$. Meanwhile, there might be $A \to B \to C$ in the knowledge graph. So the path $A \to B \to C$ refers to the edge $A \to C$. In order to tackle this case, we further define $\mathcal{P}$-*Causal Knowledge Graph*.

[$\mathcal{P}-$Causal Sufficient Knowledge Graph] If a knowledge graph KG

$$\forall \text{ path } p \text{ with at most } P \text{ distinct vertices } \implies \exists e' \in E_C. \tag{20}$$

then we say KG is $\mathcal{P}$ causal sufficient.

[$\mathcal{P}-$Causal Efficient Knowledge Graph] If a knowledge graph KG,

$$\forall e \in E_C \implies \exists \text{ path } p \text{ with at most } P$$

distinct vertices $\in E_{KG}$.(21)

then we say KG is $\mathcal{P}$ causal efficient.

[$\mathcal{P}-$Causal Equivalent Knowledge Graph] If a knowledge graph KG are both $\mathcal{P}-$causal sufficient and efficient regarding a casual graph $\mathcal{G}$, we say KG is $\mathcal{P}$ causal equivalent to $\mathcal{G}$.

With previous definitions, we introduce the definition of *causal sufficient score*, *causal efficient score*, *causal informative score*.

[Causal Informative Score] Given a knowledge graph KG, ground truth causal graph $\mathcal{G}$, and a set of causal equivalent sub-graph $\{KG_i = \{V_i, E_i\}\}$, the causal informative score is defined as

$$C\_Score = \frac{\max|E_i|}{|E_{KG}|}. \tag{22}$$

From the definition, it is easy to see the causal informative score illustrates how much causal information a KG contains. Moreover, we can derive the $\mathcal{P}-$causal informative score based on the definitions of $\mathcal{P}-$causal equivalent KG.

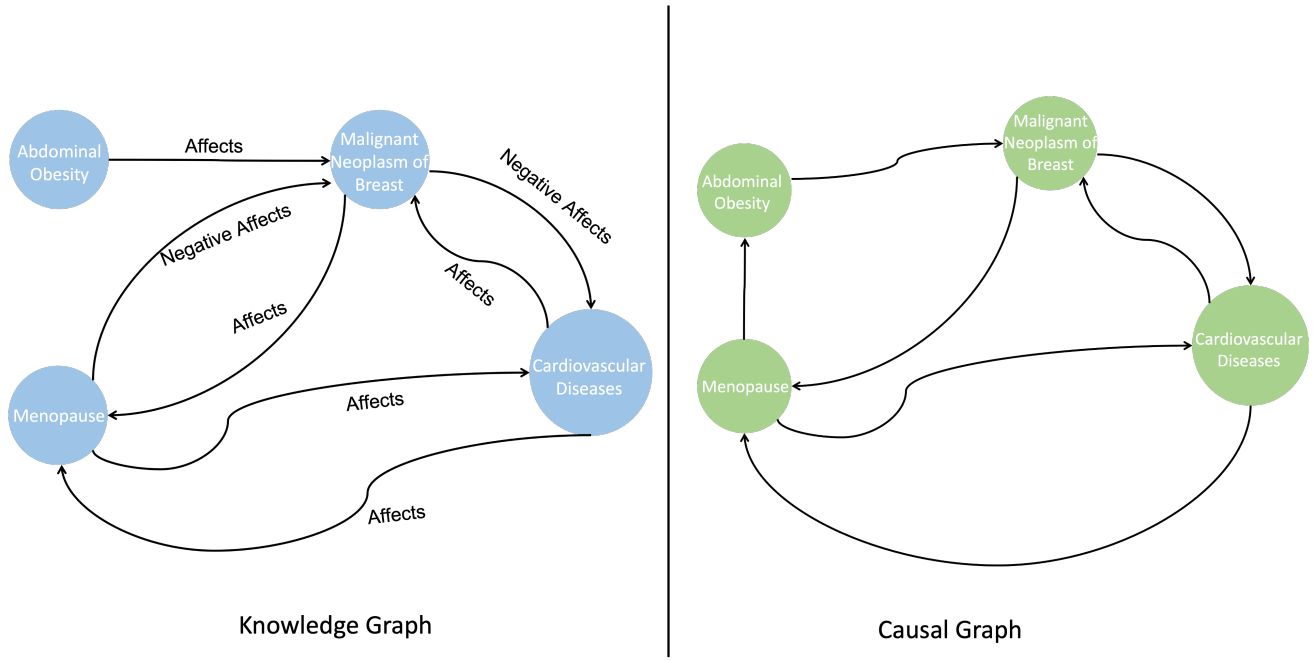

Figure 4: Knowledge Graph examples and corresponding causal graphs.

Knowledge Graph	P1	P2	P3	P1 K_Score	P2 K_Score	P3 K_Score	P1 I_Score	P2 I_Score	P3 I_Score
WordNet (Fig. 1)	0.95	0.75	0.90	0.70	0.47	0.53	0.42	0.69	0.84
VisualGenome (Fig. 2)	0.97	0.78	0.60	0.67	0.50	0.81	0.16	0.35	0.08
PrimeKG (Fig. 3)	0.99	0.88	0.95	0.90	0.94	0.41	0.94	0.89	0.69

Table 1: Human Subjective Evaluation vs K_Score and I_Score Evaluation.

[$\mathcal{P}-$Causal Informative Score] Given a knowledge graph KG, ground truth causal graph $\mathcal{G}$, and a set of $\mathcal{P}-$ causal equivalent sub-graph

$$\{KG_i^{\mathcal{P}} = \{V_i^{\mathcal{P}}, E_i^{\mathcal{P}}\}\}. \tag{23}$$

the $\mathcal{P}-$causal informative score is defined as

$$C\_Score_{\mathcal{P}} = \frac{\max|E_i^{\mathcal{P}}|}{|E_{KG}|}. \tag{24}$$

We can see that the $\mathcal{P}-$causal informative score reveals how much causal information the learned KG contains based on $\mathcal{P}-$causal equivalent knowledge graphs. The main difference between $\mathcal{P}-$Causal Informative Score and Causal Informative Score is that we consider a large number of covariates (confounders and mediators) in KG which do not appear in a causal graph $\mathcal{G}$. As the previous example shows, there might be $A \to B \to C$ in the knowledge graph but only $A \to C$ appears in the causal graph.

Examples Fig. 4 shows an example of a knowledge graph and corresponding causal graph of obesity, heart disease, menopause, and breast cancer. Thus, we can calculate the following C_Score and $\mathcal{P}-$C_Score based on above definitions and illustrations:

1. C_Score = $\frac{4}{7}$,
2. $2-$C_Score = $\frac{4}{7}$.

We can see that, in this case, the causal informative score and $2-$Causal Informative Score are equivalent. One limitation of the above framework is that a ground truth causal knowledge graph is required to be compared. Indeed in current causal discovery literature, simulated studies based on ground truth causal graphs are often needed to validate causal discovery algorithms' performance.

Evaluation

Comparison with Existing Approaches

The major advantages of the three metrics we propose are as the following:

First, compared with the downstream tasks approach like evaluating by question answering (QA) and information retrieval (IR) (Nair et al. 2018), one direct benefit of applying our metrics is we don't need to implement a QA or IR systems to evaluate the knowledge graphs/knowledge bases, which may require additional engineering efforts and technical expertise. Our metrics, for example, the K_Score, originated from human-centered evaluation, are much more convenient to use and direct. Furthermore, given we offer a set of three metrics that have different characteristics and strengths, users can choose whatever they think is the most

suitable for specific purposes.

On the other hand, compared to the case study approach (Chandak, Huang, and Zitnik 2022), our methods are more general and have broader applicability. One example of case study evaluation is (Wang et al. 2020), which uses a drug repurposing report generation task to evaluate the COVID-19 literature knowledge graph they developed. Although it's very practical and useful, it's relatively specific and is not as versatile as the metrics we propose.

Last but not least, our personalized measurement approach is a great and concrete example concurring with the human-centered AI initiative advocated by (Shneiderman 2022) and the causality score can be linked with the data generation process (Pearl et al. 2000). We will illustrate the two points further in follow-up research with more details.

Quantitative Evaluation

Anderson et al. (2016) proposed an automated evaluation metric SPICE for image caption. It computes a score that captures the similarity between a candidate caption and a set of reference captions associated with an image. It compares SPICE to existing metrics by correlation with human judgments, It inspires us to design a study to study the correlation between the three metrics we propose and human judgments.

We start by conducting a human-centered study by inviting three interviewees (P1, P2, and P3) to look at the three typical knowledge graphs (Figure 1, 2, and 3) and give a subjective evaluation of the knowledge gain. Then we calculate the personalized K_Score and I_Score for the three-person respectively. The results are summarized in Table 1. In the above pilot evaluation study, we identify multiple interesting phenomena:

First, human subjective evaluation can be quite arbitrary, the current three samples cannot provide stable ground truth. Second, Indeed different persons intrinsically equip with different context knowledge (a.k.a different c) when facing different knowledge graphs, but how to estimate the parameter c is a question. Third, what is unknown is not always clear to humans and it can be elevated to a philosophical level.

In the following steps, we plan to use Pearson's ρ (Benesty et al. 2009) to measure linear association and Kendall's τ rank correlation coefficient (Abdi 2007) to evaluate the ordinal association between two measured quantities.

Challenges of Computing the C_Score and $\mathcal{P}-$C_Score

There are two main challenges in computing the C_Score and $\mathcal{P}-$C_Score. First there may be multiple nodes in KG referring to one random variable in the causal graph. For example *Malignant ovarian surface Epithelial-Stromal Tumor* and *Epithelial ovarian cancer* refer to the same concept (node) in the causal graph. How to map a knowledge graph's vertices to causal graphs remains an open problem. Second, there are various kinds of edges that represent different relations in KG. It's still under exploration what edges should be classified as "causal" edges. Solving these two main challenges is one of the future works of the paper.

Conclusion and Future Work

Knowledge graphs represent networks of real-world entities: objects, events, situations, or concepts, and illustrate the relationship between them. In this paper, we present three new principled metrics K_Score based on the science of science, I_Score based on information theory, and C_Score based on causality to evaluate the knowledge in knowledge graphs. In the next step, we plan to carefully design a survey study to collect human judgment of knowledge gained from a collection of knowledge graph examples, then conduct the correlation study mentioned above. This work provides novel and unique perspectives for pressing knowledge graph evaluation problems.

References

Abdi, H. 2007. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, 508–510.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, 1–4. Springer.

Chandak, P.; Huang, K.; and Zitnik, M. 2022. Building a knowledge graph to enable precision medicine. *bioRxiv*.

Domingo-Fernández, D.; Gadiya, Y.; Patel, A.; Mubeen, S.; Rivas-Barragan, D.; Diana, C. W.; Misra, B. B.; Healey, D.; Rokicki, J.; and Colluru, V. 2022. Causal reasoning over knowledge graphs leveraging drug-perturbed and disease-specific transcriptomic signatures for drug discovery. *PLoS computational biology*, 18(2): e1009909.

Dou, J. 2017. Impartial redistricting: a Markov chain approach to the" Gerrymandering problem". *arXiv preprint arXiv:1711.04618*.

Dou, J.; Liu, M.; Muneer, H.; and Schlussel, A. 2017. What Words Do We Use to Lie?: Word Choice in Deceptive Messages. *arXiv preprint arXiv:1710.00273*.

Dou, J.; Sun, N.; and Zou, X. 2016. " Draw My Topics": Find Desired Topics fast from large scale of Corpus. *arXiv preprint arXiv:1602.01428*.

Dou, J. X.; Jia, M.; Zaslavsky, N.; Bao, R.; Zhang, S.; Ni, K.; Liang, P. P.; Mao, H.; and Mao, Z. 2022. Learning More Effective Cell Representations Efficiently. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*.

Dou, J. X.; Luo, L.; and Yang, R. M. 2022. An optimal transport approach to deep metric learning (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12935–12936.

Du, J.; Yan, X.; Liu, Z.; Cui, L.; Ding, P.; Tan, X.; Li, X.; Zhou, H.; Gu, Q.; and Xu, J. 2017. cBinderDB: a covalent binding agent database. *Bioinformatics*, 33(8): 1258–1260.

Fortunato, S.; Bergstrom, C. T.; Börner, K.; Evans, J. A.; Helbing, D.; Milojević, S.; Petersen, A. M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; et al. 2018. Science of science. *Science*, 359(6379): eaao0185.

Hegel, G. W. F. 2018. *Hegel: The phenomenology of spirit*. Oxford University Press.

Kamdar, M. R. 2019. *A web-based integration framework over heterogeneous biomedical data and knowledge sources*. Stanford University.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Li, S.; Feng, M.; Wang, L.; Essofi, A.; Cao, Y.; Yan, J.; and Song, L. 2022. Explaining Point Processes by Learning Interpretable Temporal Logic Rules. In *International Conference on Learning Representations*.

Liu, J.; Lian, J.; Sprott, J. C.; Liu, Q.; and Ma, Y. 2022. The Butterfly Effect in Primary Visual Cortex. *IEEE Transactions on Computers*.

Majewska, O.; Collins, C.; Baker, S.; Björne, J.; Brown, S. W.; Korhonen, A.; and Palmer, M. 2021. BioVerbNet: a large semantic-syntactic classification of verbs in biomedicine. *Journal of Biomedical Semantics*, 12(1): 1–13.

Mao, H.; Broerman, M. J.; and Benos, P. V. 2020. Interpretable Factors in scRNA-seq Data with Disentangled Generative Models. In *(BIBE)*, 85–88. IEEE.

Mao, H.; and Dou, J. X. 2022. Decomposable Sparse Tensor on Tensor Regression. *arXiv preprint arXiv:2212.05024*.

Mao, H.; Jia, M.; Dou, J. X.; Zhang, H.; and Benos, P. V. 2022a. COEM: cross-modal embedding for metacell identification. *ICML Workshop on Computational Biology*.

Mao, H.; Liu, H.; Dou, J. X.; and Benos, P. V. 2022b. Towards Cross-Modal Causal Structure and Representation Learning. In *Machine Learning for Health*.

Miller, G. A. 1998. *WordNet: An electronic lexical database*. MIT press.

Mo, S.; Xia, J.; Tan, X.; and Raj, B. 2022. Point3D: tracking actions as moving points with 3D CNNs. *arXiv preprint arXiv:2203.10584*.

Nair, A.; Pong, V.; Dalal, M.; Bahl, S.; Lin, S.; and Levine, S. 2018. Visual reinforcement learning with imagined goals. *arXiv preprint arXiv:1807.04742*.

Nordon, G.; Koren, G.; Shalev, V.; Kimelfeld, B.; Shalit, U.; and Radinsky, K. 2019. Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1102–1109.

Paudel, S.; Warner, B. E.; Wang, R.; Adams-Haduch, J.; Reznik, A. S.; Dou, J. X.; Huang, Y.; Gao, Y.-T.; Koh, W.-P.; Bäckerholm, A.; et al. 2022. Serologic Profiling Using an Epstein-Barr Virus Mammalian Expression Library Identifies EBNA1 IgA as a Prediagnostic Marker for Nasopharyngeal Carcinoma. *Clinical Cancer Research*, OF1–OF10.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19: 2.

Qian, K.; Beirami, A.; Kottur, S.; Shayandeh, S.; Crook, P.; Geramifard, A.; Yu, Z.; and Sankar, C. 2021. Database Search Results Disambiguation for Task-Oriented Dialog Systems. *arXiv preprint arXiv:2112.08351*.

Ren, Y.; Senarathna, J.; Grayson, W. L.; and Pathak, A. P. 2022. State-of-the-art techniques for imaging the vascular microenvironment in craniofacial bone tissue engineering applications. *American Journal of Physiology-Cell Physiology*, 323(5): C1524–C1538.

Rosas, L.; Jia, M.; Cruz, T.; Kapetanaki, M.; Tabib, T.; Sembrat, J.; Stacey, S.; Reader, B.; Peters, V.; Riley, M.; et al. 2022. Selective Expression of MOXD1 in IPF Fibroblast Correlates with Cell Senescence. In *A60. PULMONARY FIBROSIS: ANIMAL AND CELL CULTURE MODELS*, A1967–A1967. American Thoracic Society.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Shi, J.; Wu, Y.; Zeng, D.; Hu, J.; and Shi, Y. 2022. FedCoCo: A Memory Efficient Federated Self-supervised Framework for On-Device Visual Representation Learning. *arXiv preprint arXiv:2212.01006*.

Shneiderman, B. 2022. *Human-Centered AI*. Oxford University Press.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Tan, X.; Chang, C.-C. H.; Zhou, L.; and Tang, L. 2022a. A Tree-based Model Averaging Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 21013–21036. PMLR.

Tan, X.; Qi, Z.; Seymour, C. W.; and Tang, L. 2022b. RISE: Robust Individualized Decision Learning with Sensitive Variables. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Tang, H.; Guo, L.; Fu, X.; Wang, Y.; Mackin, S.; Ajilore, O.; Leow, A. D.; Thompson, P. M.; Huang, H.; and Zhan, L. 2023. Signed graph representation learning for functional-to-structural brain network mapping. *Medical Image Analysis*, 83: 102674.

Tang, H.; Ma, G.; Guo, L.; Fu, X.; Huang, H.; and Zhan, L. 2022. Contrastive brain network learning via hierarchical signed graph pooling model. *IEEE Transactions on Neural Networks and Learning Systems*.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Wang, Q.; Li, M.; Wang, X.; Parulian, N.; Han, G.; Ma, J.; Tu, J.; Lin, Y.; Zhang, H.; Liu, W.; et al. 2020. COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv:2007.00576*.

Weis, J. W.; and Jacobson, J. M. 2021. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology*, 39(10): 1300–1307.

Wu, L.; Wang, D.; and Evans, J. A. 2019. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744): 378–382.

Xu, F.; Wu, L.; and Evans, J. 2022. Flat Teams Drive Scientific Innovation. *arXiv preprint arXiv:2201.06726*.

Yang, Z.; Yu, H.; He, Y.; Sun, W.; Mao, Z.-H.; and Mian, A. 2022. Fully Convolutional Network-Based Self-Supervised Learning for Semantic Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zeng, H.; Ng, Z. W.; Zhou, P.; Lou, X.; Yau, D. K.; and Winslett, M. 2022. Detecting Cyber Attacks in Smart Grids with Massive Unlabeled Sensing Data. In *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 1–7. IEEE.