How to Avoid Jumping to Conclusions: Measuring the Robustness of Outstanding Facts in Knowledge Graphs

ABSTRACT

An *outstanding fact* (OF) is a striking claim by which some entities stand out from their peers on some attribute. OFs serve data journalism, fact checking, and recommendation. However, one could *jump to conclusions* by selecting truthful OFs while intentionally or inadvertently ignoring lateral contexts and data that render them less striking. This *jumping conclusion bias* from unstable OFs may disorient the public, including voters and consumers, raising concerns about fairness and transparency in political and business competition. It is thus ethically imperative for several stakeholders to measure the robustness of OFs with respect to lateral contexts and data. Unfortunately, a capacity for such inspection of OFs mined from knowledge graphs (KGs) is missing. In this paper, we propose a methodology that inspects the robustness of OFs in KGs by perturbation analysis. We define (1) entity perturbation, which detects outlying contexts by perturbing context entities in the OF; and (2) data perturbation, which considers plausible data that render an OF less striking. We compute the *expected* strikingness scores of OFs over *perturbation relevance distributions* and assess an OF as robust if its *measured* strikingness does not deviate significantly from the expected. We devise a suite of exact and sampling algorithms for perturbation analysis on large KGs. Extensive experiments reveal that our methodology accurately and efficiently detects frail OFs generated by existing mining approaches on KGs. We also show the effectiveness of our approaches via case and user studies.

CCS CONCEPTS

• **Computing methodologies → Semantic networks**.

KEYWORDS

Outstanding fact discovery; Robustness Measuring; Perturbation analysis; Knowledge graphs

ACM Reference Format:

. 2024. How to Avoid Jumping to Conclusions: Measuring the Robustness of Outstanding Facts in Knowledge Graphs. In . ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The discovery of *outstanding facts* (OFs) is central to a wide range of application domains, including data journalism [8, 23], fact checking [2, 18–22], and recommendation [29, 46, 49, 50]. An OF is a striking statement that an attribute of one or a few entities stands out among their *peer entities*. For instance, the news media often use OFs as leads to draft news stories and attract audiences: *"Kamala Harris is the first woman to hold the position of Vice President of the United States"* [37]. This statement reveals an OF regarding gender (*attribute*) over all politicians (*peer entities*) who have held the position of US vice president. Manual discovery of OFs can be tedious and time-consuming, requiring extensive human effort to sift through vast amounts of data. As such, there have been many studies on automatic OF mining [5, 10, 25, 43, 46, 49].

Despite the traction of OFs in engaging with a diversity of audiences, from journalists in the first place to citizens as voters or consumers, their striking and captivating nature also entails a risk that they may induce misinformation by misleading people to *jump to conclusions* [26], i.e., craft rushed interpretations and reach biased decisions [2, 28, 45]. For example, in sales campaigns, rare product attributes may be framed as OFs by strategically placing products in specific contexts, potentially misinforming customers [11]. Such OF-induced misinformation may violate legislation; in the UK, for instance, it is classified as *misleading action and omission*, a trading practice that violates the *Consumer Protection from Unfair Trading Regulations* (CPRs) [34]. Likewise, advances in data journalism come with the perils of misinformation; whilst media scholars have long shown the power of media in shaping public opinion [9], legal scholars have been debating the legal implications of disinformation as a fundamental rights violation [39]. To address the perils of OF-induced misinformation in particular, prior works conduct *perturbation analysis*, i.e., assess OFs by a context-aware robustness measure while perturbing numerical parameters [45], group-by attributes [28], and time intervals [2] in the query that defines the OF context. However, existing robustness measures and perturbation analysis methods overlook two types of hasty generalizations: (1) *context entity generalization* and (2) *limited data generalization*.

Context Entity Generalization. *The cause of gender parity in academia engenders extensive and controversial discourse within online communities and academic circles [30, 33]. To seize public attention, one may generate eye-catching OFs using existing OF discovery methods [46, 49]: for instance, the American Council on Education (ACE), a prominent US higher education association, boasts a diverse membership comprising numerous accredited colleges and universities. Notably, among the individuals employed by ACE member institutions, the male demographic constitutes 31%.*

This fact might mislead the public to generalize about a perceived disparity of male faculty members in US universities. However, considering other associations, such as the American Association of State Colleges and Universities, the male percentage rises to 50%. In fact, the *aggregate* male percentage among US academics is 66%. We conclude that ACE is an untypical *context entity* that deliberately or inadvertently provokes generalizations about the gender gap.

Limited Data Generalization. *Seed investing in startups generates returns from the capital market once a company is publicly listed. Reddit, a US social media site, has filed for its initial public offering*

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

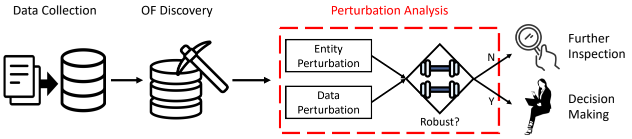


Figure 1: Workflow of the OF mining process.

(IPO) [13]. An OF generated by [46, 49] on Wikidata yields Uber as the ***only*** *success case of an IPO backed by Reddit's investors.*

The noted OF regarding Reddit's investors could undermine confidence in Reddit's IPO due to an apparent lack of investor experience in IPO cases. However, startup enterprises often progress through multiple funding rounds, rendering such a generalization dubious, as Reddit has the potential to attract more investors in the future. Remarkably, Sequoia Capital invested in Reddit during its 2017 Series C funding round [6], a fact *undocumented* in Wikidata at the time of writing, and has also invested in several publicly traded companies, including Apple, Airbnb, and Nvidia.

Perturbation Analysis of OFs. To mitigate such potential ramifications, we propose a novel *perturbation analysis* methodology that evaluates OF robustness. We focus on OFs mined from knowledge graphs (KGs) [46, 49], as there is currently no method to assess their robustness, while the richness of facts available in large KGs, such as DBpedia [27], Freebase [4], and Wikidata [40], allows semantically relevant perturbations to curb both context entity and limited data generalizations. We define (1) *entity perturbation*, which evaluates the strikingness of an OF by substituting the context entity with a similar entity from the KG (e.g., replacing "ACE" with a similar association), and (2) *data perturbation*, which introduces a plausible edge into the KG to evaluate the continued validity of the OF claim (e.g., including another investor, Sequoia Capital, to Reddit).

Moreover, we calculate an OF's *expected* strikingness over a *perturbation relevance distribution* (PRD) that assigns weights to (entity or data) perturbations by their relevance to the OF. An OF whose strikingness is significantly higher than the expected value derived from the respective PRD is less robust than others and should be inspected by domain experts. Our methodology can be seamlessly integrated into the overall OF mining process, as the dashed-box component in Figure 1 illustrates. Using our methodology, stakeholders such as data journalists, fact-checkers, civil society organizations, and other interested parties can cross-check, evaluate, and verify the robustness of OFs on which they intend to base decisions. We summarize our main contributions as follows:

- We formalize the perturbation analysis of OFs from KGs by *entity* and *data perturbation*. To our best knowledge, this is the first attempt to measure the robustness of OFs extracted from KGs.
- We devise efficient exact algorithms for entity and data perturbation analysis. We further propose sampling-based approximation algorithms that scale to large perturbation spaces.
- We conduct extensive experiments to evaluate our approaches. The results demonstrate that our analysis discovers frail OFs generated by existing OF mining approaches [46]. Furthermore, our approximation algorithms complete the perturbation analysis within 5 minutes for over 90% of OFs, with a mean estimation error around 2.7% and 1.3% for entity and data perturbation, respectively. Finally, we validate the effectiveness of our system with a crowdsourced user study.

2 BACKGROUND

A knowledge graph (KG) is denoted as $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{L})$, where $\mathcal{V}$ and $\mathcal{E}$ represent the sets of node and edge instances, respectively. An edge instance $e \in \mathcal{E}$ connects two node instances $u, v \in \mathcal{V}$. The function $\mathcal{L}$ maps each node or edge instance v or e to its type $\mathcal{L}(v)$ or $\mathcal{L}(e)$. We define a node variable $\tilde{v} \in \{v, \mathcal{L}(v)\}$ to be either a node instance v or a node type $\mathcal{L}(v)$. $\mathcal{N}(v, \mathcal{L}(e))$ and $\mathcal{N}(V, \mathcal{L}(e))$ represent the neighbors of node v and node set V via edge type $\mathcal{L}(e)$, respectively. $\mathcal{N}(v)$ and $\mathcal{N}(V)$ are the neighbors of node v and node set V without the edge type constraint. For ease of presentation, we simply use $\mathcal{G}(\mathcal{V}, \mathcal{E})$ or $\mathcal{G}$ to refer to the KG.

We study OFs discovered by path queries, as path queries can be easily and intuitively translated into natural language claims [46]. Nonetheless, our perturbation analysis can be extended to support general KG queries. Formally, we define path queries as follows:

Definition 2.1 (Path Query). A path query is a sequence of node variables and edge types. We use $P(\tilde{v_0}, E_0, \tilde{v_1}, \ldots, \tilde{v}_{k-1}, E_{k-1}, \tilde{v_k})$, or simply $P(\tilde{v_0}, \tilde{v_k})$, to denote a k-hop path query, where E_i is the edge type of the i^{th} edge and $i \in \{0, ..., k-1\}$. E_i^{-1} represents the *syntactic reverse* edge type of E_i, i.e., for an edge $e = (u, v)$ where $\mathcal{L}(e) = E_i$, we denote by $\mathcal{L}(e') = E_i^{-1}$ the type of its reverse edge $e' = (v, u)$.

Definition 2.2 (Matching Instance). A *matching instance* to a path query $P(\tilde{v_0}, \tilde{v_k})$ is a sequence of node and edge instances in $\mathcal{G}$ denoted by $p(v_0, e_0, v_1, \ldots, v_{k-1}, e_{k-1}, v_k)$ or $p(v_0, v_k)$ satisfying:

- $\forall i \in \{0, \ldots, k\}$, $\mathcal{L}(v_i) = \tilde{v_i}$, where $\tilde{v_i}$ is a node type $\mathcal{L}(v)$;
- $\forall i \in \{0, \ldots, k\}$, $v_i = \tilde{v_i}$, where $\tilde{v_i}$ is a node instance v; and
- $\forall i \in \{0, \ldots, k-1\}, e_i = (v_i, v_{i+1}) \in \mathcal{E}$ and $\mathcal{L}(e_i) = E_i$.

We use $p \triangleright P$ to represent that path p is a matching instance of path query P. For each node variable $\tilde{v_i}$ in path query P, we denote the *matching node set* of $\tilde{v_i}$ (i.e., the set of all node instances that correspond to $\tilde{v_i}$ in P) as $V_i \subseteq \mathcal{V}$ for all $i \in \{0, \ldots, k\}$.

Most existing studies [46, 49] discover attribute-value pairs (e.g., *Gender-Female*) in OFs by defining *peer entities*. To generate diversified OFs, e.g., news stories in different contexts, they acquire peer entities from *context path queries* by specifying a node constraint as context information. Formally:

Definition 2.3 (Peer Entities). Given a *target attribute* $\mathcal{A}$ and a path query $P(\tilde{v_0}, \tilde{v_k} = c)$ where c is a *context entity*, an instance v_0 is a *peer entity* of c under P and $\mathcal{A}$ if $\exists\, p(v_0, c) \triangleright P(\tilde{v_0}, c)$ and v_0 has the target attribute $\mathcal{A}$. We denote by V_0 the set of peer entities of c under a path query P and attribute $\mathcal{A}$.

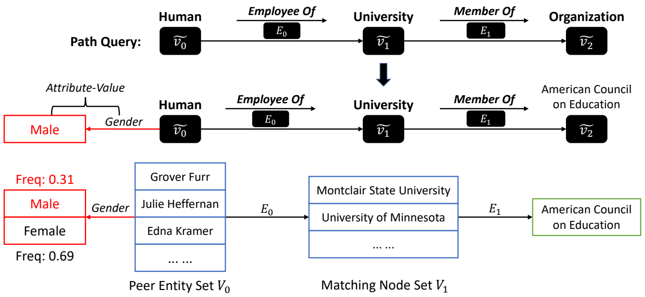


Figure 2: An example of OF discovery.

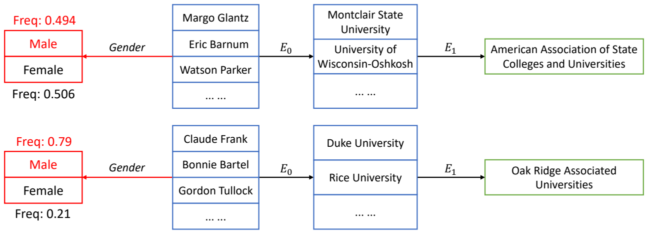


Figure 3: An example of entity perturbation.

Example 2.4. Consider the path query in Figure 2 and the target attribute *Gender*. The query $P(\tilde{v_0}, E_0, \tilde{v_1}, E_1,$ 'American Council on Education') uses the ACE as a context entity for OF discovery, where E_0 = 'Employee Of' and E_1 = 'Member Of'. Under query P, set V_1 contains all universities with ACE membership, while the set of peer entities V_0 includes all people who have worked with at least one of these universities and have the *Gender* attribute.

Definition 2.5 (Candidate OF). A candidate OF is a triplet $Q = (\mathcal{A}, \mathcal{X}, P)$, where $\mathcal{A}$, $\mathcal{X}$ and P denote the following:

- $\mathcal{X}$ is a *value* in the domain of *attribute* $\mathcal{A}$; and
- P is a path query $P(\tilde{v_0}, c)$ that implicitly defines *peer entities* of the *context entity* c under attribute $\mathcal{A}$.

The value $\mathcal{X}$ in a candidate OF is considered *striking* if it stands out among the values of the same attribute $\mathcal{A}$ in the peer entity set defined by P. Specifically, we adopt the OF *strikingness measure*, i.e., the measure used by existing OF discovery approaches [44, 46, 49] to rank the candidate OFs and identify the most striking ones.

Definition 2.6 (Strikingness Measure). Given a candidate OF $Q = (\mathcal{A}, \mathcal{X}, P)$, $F(\mathcal{A}, \mathcal{X}, V_0)$ denotes the *frequency* of value $\mathcal{X}$ for attribute $\mathcal{A}$ among peer entities in V_0. The strikingness of Q is:

$$\mathcal{I}(Q) = \sum_{\mathcal{X}' \in \overline{\mathcal{X}}} F(\mathcal{A}, \mathcal{X}', V_0),$$

where $\overline{\mathcal{X}} = \{\mathcal{X}' | F(\mathcal{A}, \mathcal{X}', V_0) > F(\mathcal{A}, \mathcal{X}, V_0)\}$, i.e., the set of values X' whose frequency for attribute $\mathcal{A}$ is higher than that of $\mathcal{X}$ among the peer entities.

Example 2.7. Figure 2 presents a candidate OF regarding the gender gap among US university employees. As elaborated in Example 2.4, the peer entity set V_0 includes all people who have the attribute *Gender* and have worked with an ACE member. The frequency of appearance of the attribute-value pair *Gender-Male* in V_0 is 0.31. Given that the sum of frequencies for *Gender* values **higher** than that of *Male* (i.e., in this case, for value *Female*) is 0.69, by Definition 2.6, the OF's strikingness score for attribute-value pair *Gender-Male* is 0.69. *In simple terms, there are fewer male employees among ACE universities, with a strikingness of 0.69.*

We emphasize that OF discovery is independent of the challenge we address in this paper. Our task is rather to help assess whether *a given OF* is robust, as the workflow in Figure 1 depicts.

3 PERTURBATION MODEL

In this section, we first introduce the two perturbation types we examine and define their respective perturbation spaces. Then, we formalize the perturbation relevance distribution in these perturbation spaces, which we use to assess the robustness of OFs.

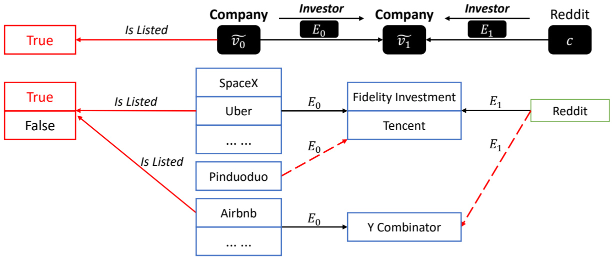


Figure 4: An example of data perturbation.

3.1 Perturbation Space

We introduce two types of perturbation: *entity perturbation* and *data perturbation*. An **entity perturbation** assesses the robustness of an OF-generating query by changing its context entity. Formally,

Definition 3.1 (Entity Perturbation Space). An *entity perturbation* Q' of an OF $Q = (\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$ replaces the context entity path constraint c with another entity $c' \in \mathcal{V}$ such that:

- $\exists\, p(v_0, c') \triangleright P' = P(\tilde{v_0}, c')$; and
- v_0 has the same value $\mathcal{X}$ on attribute $\mathcal{A}$ as the OF Q.

We denote the perturbed OF by $Q' = (\mathcal{A}, \mathcal{X}, P')$. The *entity perturbation space* of Q is denoted as $\mathcal{Q}$ and contains all possible substitute OFs Q' for Q with different context entities c'.

Intuitively, a valid entity perturbation changes the context entity path constraint c to c' such that the perturbed peer entities still contain the attribute-value pair $(\mathcal{A}, \mathcal{X})$ of the original OF.

Example 3.2. Figure 3 shows two examples of entity perturbation for the OF in Figure 2. When we replace the context entity 'American Council on Education' in the original OF with the 'American Association of State Colleges and Universities' (AASCU), the matching instances for the resulting path query differ significantly. The peer entity set now includes faculty members in other universities, e.g., the University of Wisconsin–Oshkosh, bringing the strikingness of the *Gender-Male* pair down to 0.506 (from 0.69 in the original OF). The Oak Ridge Associated Universities (ORAU), a consortium of US universities, is another possible entity perturbation. With ORAU as the context entity, the strikingness of "male employees" drops even lower, to 0, since no other *Gender* value has **higher** frequency than *Male* in the respective V_0.

A **data perturbation** adds a plausible edge to the KG that may render an OF less striking. While many plausible edges may be added to the KG, most would not affect the OF at hand. We only consider *admissible* data perturbations, i.e., edges whose addition results in a different peer entity set for path query $P(\tilde{v_0}, c)$:

Definition 3.3 (Data Perturbation Space). An *admissible data perturbation* Q' of an OF $Q = (\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$ adds an edge $e' = (u, v)$ to the KG $\mathcal{G}$ such that $V_0' \neq V_0$, where V_0' is the new peer entity set of the original path query $P(\tilde{v_0}, c)$ after adding e' to $\mathcal{G}$.

For a perturbation to be admissible, the node variable corresponding to u in the path query $P(\tilde{v_0}, c)$ should be a node type.

Example 3.4. Figure 4 illustrates two admissible data perturbations under the Reddit OF mentioned in Section 1. The first adds a plausible edge e = ('Pinduoduo', 'Tencent') for E_0. Consequently, Pinduoduo is added to the peer entity set V_0, rendering the *Is_listed-True* pair less striking in the updated V_0 since Pinduoduo has already been publicly listed on NASDAQ in July 2018. Another plausible edge $e' = $ (‘Y Combinator’, ‘Reddit’) is also an admissible perturbation, for E_1 this time. Recovering e' reveals more publicly listed companies invested by Y Combinator, including Airbnb, CoinBase, Twillo, PagerDuty, and Rackspace Technology.

3.2 Perturbation Relevance Distribution

Naturally, some perturbations are more relevant to an OF than others. For instance, if one values the attributes of the association, ORAU presents a compelling alternative to ACE in Example 2.4 since both associations welcome non-state-supported universities. Conversely, if one values members’ constitution, AASCU emerges as a convincing substitute since Montclair State University (under ACE) is also affiliated with AASCU. Similarly, between the data perturbations (‘Pinduoduo’, ‘Tencent’) and (‘Y Combinator’, ‘Reddit’) in Example 3.4, the more relevant to the Reddit OF in the KG would be a more convincing perturbation. To formalize the relevance of entity and data perturbations, we introduce *relevance distributions*.

Entity Perturbation Relevance. For a relevant entity perturbation, the context entity c' in the perturbed OF Q' should be semantically relevant to the context entity c in the original OF Q. Given an OF Q produced by path query $P(\tilde{v_0}, c)$, we model the *relevance* of an entity perturbation $Q' \in \mathcal{Q}$ from c to c' as the *node similarity score* between c and c'. The choice of similarity function is orthogonal to our method and any measure applicable to KGs [31, 36] can be used. Here, we opt for the Jaccard distance $\mathcal{S}(c, c')$ between the neighbor sets of c and c', i.e., between $\mathcal{N}(c)$ and $\mathcal{N}(c')$.

Data Perturbation Relevance. For a data perturbation to be relevant to an OF Q, the added edges should be (*i*) semantically related to Q and (*ii*) plausibly existing now or in the near future.

To address both desiderata, we propose an efficiently computed function, *head-tail relevance*. Conceptually, *head relevance* measures the semantic relevance of an added edge $e = (u, v)$ to OF Q through the similarity of e to edges in the matching path instances of Q, in particular, the similarity of u to the neighbors of v that appear in the matching node set V of the node type of u. Nevertheless, head relevance alone may spawn relevant yet implausible edges. To account for the plausibility of edge e, we augment the function with *tail relevance*, which measures the similarity between u’s neighbors of type E and the tail node v.

Given an admissible data perturbation $e = (u, v)$ for edge type E of OF Q, we formally define the head-tail relevance as:

$$\mathcal{S}(e) = \underbrace{\left(\sum_{u' \in \mathcal{N}(v, E^{-1}) \cap V} \mathcal{S}(u, u')\right)}_{\text{head relevance}} \underbrace{\left(\sum_{v' \in \mathcal{N}(u, E)} \mathcal{S}(v, v')\right)}_{\text{tail relevance}}$$

To accommodate space constraints, we present an example that illustrates the rationale by which we define the head-tail relevance in Appendix A.

Discussion. Existing *link prediction* methods can suggest plausible edges in a KG [3, 7, 47]; such methods may provide orthogonal validation to strengthen the perturbations identified by our methods. However, these methods target the prediction of any edge in a graph as an end in itself; contrariwise, we are interested in suggesting plausible edges with respect to an OF Q for the larger objective of building a distribution over them, as we explain next.

Based on the above concepts, we now formally define the *perturbation relevance distribution* (PRD) for an OF Q:

Definition 3.5 (Perturbation Relevance Distribution (PRD)). Given an OF Q, we denote as $\mathcal{D}(Q)$ the PRD for entity/data perturbation. The probability mass function of $\mathcal{D}(Q)$ is proportional to the corresponding perturbation relevance function, i.e., $\mathbb{P}_{\mathcal{D}}(\cdot) \propto \mathcal{S}(c, c')$ for entity perturbation and $\mathbb{P}_{\mathcal{D}}(\cdot) \propto \mathcal{S}(e)$ for data perturbation.

Deviation Measure. Intuitively, an OF is not robust if many highly relevant perturbations thereof have substantially smaller strikingness scores. In effect, we may compute the mean scores under both PRDs of entity and data perturbations, and if the strikingness score of an OF Q substantially deviates from the means, we regard Q as a less robust claim. We thus define the *deviation measure* of Q as:

$$\Delta(Q) = \frac{\mathbb{E}_{Q' \sim \mathcal{D}}[\mathcal{I}(Q) - \mathcal{I}(Q')]}{\mathbb{E}_{Q' \sim \mathcal{D}}[\mathcal{I}(Q')]} \tag{1}$$

for a PRD $\mathcal{D}$ defined on Q, in a manner resembling robustness scores used in other contexts [28]. In addition, as a *counterargument* to OF Q, we return the perturbation Q' that maximizes $(I(Q) - I(Q')) \cdot \mathbb{P}_{\mathcal{D}}(Q')$. $\mathcal{I}(Q) - \mathcal{I}(Q')$ measures the extent to which a perturbation Q' deviates from the original fact in terms of strikingness, while the weight for each perturbation is determined by the PRD $\mathcal{D}$.

Before presenting our algorithms for perturbation analysis, we summarize the frequently used notation in Table 1.

Table 1: Frequently used notation.

Symbol	Description
$\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{L})$	A knowledge graph (KG)
$\tilde{v}$	A node variable, instantiated to either a node or a node type
$\mathcal{L}(v), \mathcal{L}(e)$	Node type and edge type
$\mathcal{N}(v), \mathcal{N}(V)$	All neighbors of a node v and a node set V
$\mathcal{N}(v, \mathcal{L}(e))$	Neighbors of node v via edge type $\mathcal{L}(e)$
$\mathcal{N}(V, \mathcal{L}(e))$	Neighbors of node set V via edge type $\mathcal{L}(e)$
$P(\tilde{v_0}, \tilde{v_k})$	A k-hop path query from $\tilde{v_0}$ to $\tilde{v_k}$
$Q(\mathcal{A}, \mathcal{X}, P)$	An OF with attribute $\mathcal{A}$ and value $\mathcal{X}$ based on query P
V_i	The matching node set w.r.t. $\tilde{v_i}$ of a path pattern
V_0	The peer entity set
$\mathcal{I}(\cdot)$	The strikingness measure
$\mathcal{Q}$	The entity perturbation space
$\mathbb{P}_{\mathcal{D}}$	The probability mass function of the PRD
$\Delta(Q)$	The deviation measure

4 ENTITY PERTURBATION ALGORITHMS

In this section, we first present an exact entity perturbation algorithm and then a sampling-based estimation algorithm.

Exact Algorithm. Given an OF $Q(\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$, the exact entity perturbation algorithm runs in two stages to produce the deviation measure for Q, as outlined in Algorithm 1.

ForwardPass (Lines 1–3): At this stage, we aim to find the perturbation space $\mathcal{Q}$ for the given OF. Recall that $\mathcal{Q}$ is the set of all feasible substitutes of the context entity c in the given query $P(\tilde{v_0}, c)$ of length k. We replace the original context entity c with a node variable $\tilde{v_k}$. Then, by Definition 3.1, we utilize the attribute-value pair $(\mathcal{A}, \mathcal{X})$ in the given OF Q to retrieve nodes $\tilde{v_0}$ that match the query $P(\tilde{v_0}, \tilde{v_k})$ and form a candidate set C_0 with them. Thereafter, at each level i, we find all neighbors of nodes in set C_{i-1} via edge type E_{i-1} to form set V_i. The set C_k we eventually derive at level k

constitutes the perturbation space Q. We call this stage "ForwardPass" as it is performed progressively from level 0 to level k.

BackwardPass (Lines 4–15): To evaluate the deviation measure of the given OF Q, we need to enumerate all nodes in the perturbation space $Q = C_k$. For each alternative $c' \in C_k$, we replace the context entity c in the given query P by c' and represent the perturbed query as $P(\tilde{v_0}, c')$. Then, starting from c', we find the peer entity set V'_0 for $P(\tilde{v_0}, c')$ by performing the reverse operation of ForwardPass from c' to $\tilde{v_0}$. Hence, this stage is called "BackwardPass". Thereafter, we calculate the strikingness score $\mathcal{I}(Q')$ for each perturbed OF Q' and the associated relevance $S(c, c')$. The expected strikingness score under the entity PRD is the normalized weighted sum of $\mathcal{I}(Q')$ following (c, c') in the space Q. Lastly, we compute the deviation measure $\Delta(Q)$ by Equation 1. Appendix C provides the time complexity of the whole process.

Algorithm 1: Exact Algorithm for Entity Perturbation

```
   Input: KG $\mathcal{G}$; OF $Q(\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$; strikingness score $\mathcal{I}(Q)$
   Output: Deviation measure $\Delta(Q)$
1  $C_0 \leftarrow$ AttributeValueFilter($\mathcal{A}, \mathcal{X}, \mathcal{G}$);
2  for $i \in \{1, \ldots, k\}$ do
3      $C_i \leftarrow \mathcal{N}(C_{i-1}, E_{i-1})$;
4  $\mu \leftarrow 0, Z \leftarrow 0$;
5  foreach $c' \in C_k$ do
6      $Q' \leftarrow (\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c'))$;
7      $V'_k \leftarrow \{c'\}$;
8      for $i \in \{k-1, \ldots, 0\}$ do
9          $V'_i \leftarrow \mathcal{N}(V'_{i+1}, E_i^{-1})$;
10     $\mathcal{I}(Q') \leftarrow$ Strikingness($V'_0, \mathcal{A}, \mathcal{X}$);
11     $\mu \leftarrow \mu + \mathcal{I}(Q') \cdot \mathcal{S}(c, c')$;
12     $Z \leftarrow Z + \mathcal{S}(v, v')$;
13 $\mathbb{E}_{Q' \sim \mathcal{D}} \leftarrow$ Normalization($\mu, Z$);
14 $\Delta(Q) \leftarrow$ DeviationMeasure($\mathbb{E}_{Q' \sim \mathcal{D}}, \mathcal{I}(Q)$);
15 return $\Delta(Q)$;
```

Sampling-based Approximation. The exact algorithm enumerates all possible entity perturbations to compute $\mathbb{E}_{Q' \sim \mathcal{D}}$. This enumeration is expensive, especially when the entity perturbation space Q is large. Therefore, we propose a sampling strategy to get the point estimation $\hat{\mathbb{E}}_{Q' \sim \mathcal{D}}$. Given an OF $(\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$, we use the same ForwardPass process as the exact algorithm to obtain the perturbation space. We then sample a node c' as the perturbation according to the relevance distribution $\mathbb{P}_{\mathcal{D}}$. For each sample, we evaluate the strikingness of the associated OF. The mean of n samples is a point estimator $\hat{\mathbb{E}}_{Q' \sim \mathcal{D}}$. The confidence interval (CI) for the point estimate is defined as $[\hat{\mathbb{E}}_{Q' \sim \mathcal{D}} - \epsilon, \hat{\mathbb{E}}_{Q' \sim \mathcal{D}} + \epsilon]$. Since all samples are independent and identically distributed (i.i.d.) with finite variance, the half width ϵ is $O(\sqrt{\sigma^2/n})$ by the Central Limit Theorem, where σ^2 is the population variance. Further, $\mathbb{E}_{Q' \sim \mathcal{D}} \in [0, 1]$ since $\mathcal{I}(Q) \in [0, 1]$, hence the upper bound for the population variance σ^2 is 1/4. In effect, the half width ϵ is bounded by the sample size n. Thus, ϵ converges at a rate of $O(1/\sqrt{n})$ with increasing n.

5 DATA PERTURBATION ALGORITHMS

A simple strategy for data perturbation is to consider any possible combination $e = (u, v) \notin \mathcal{E}$ and check whether it satisfies Definition 3.3. However, since this strategy treats both u and v as free variables, the search space can be prohibitively large on real-world KGs with millions of nodes. Thus, we first propose methods to effectively prune the search space. Then, we develop efficient exact and sampling algorithms based on the pruned search space.

Admissibility-based Reduction. By Definition 3.3, the peer entity set V'_0 resulting from an admissible data perturbation should be different from the original peer entity set V_0. Therefore, each admissible data perturbation should add at least *one* new peer entity to V_0. It follows that an edge $e' = (u, v)$ for an admissible data perturbation should have a node u that *partially* matches the path query associated with the OF. We define the notion of a *partial matching node* as follows:

Definition 5.1 (Partial Matching Node). Given an OF $Q = (\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$, a node u is a *partial match* at level $i \in \{0, \ldots, k-1\}$ if:

- $u \notin V_i$ and $\exists\, p(v_0, \ldots, u) \triangleright P' = P(\tilde{v_0}, \tilde{v_i})$; and
- $v_0 \notin V_0$, where v_0 has attribute $\mathcal{A}$.

Here, P' is a part of P from level 0 to level i. We denote the set of partial matching nodes at level i as $\overline{V_i}$.

Lemma 5.2. *An edge $e = (u, v)$ is an admissible data perturbation for edge type E_i of an OF $Q = (\mathcal{A}, \mathcal{X}, P)$* only if *$u \in \overline{V_i}$ and for a partial matching node $v \in \overline{V_i}$ where $i \in \{1, \ldots, k\}$, there exists $u \in \overline{V_{i-1}}$ such that $v \in \mathcal{N}(u, E_{i-1})$.*

Proof. First, the necessary condition for an admissible data perturbation holds by Definition 3.3 and 5.1. Second, since $v \in \overline{V_i}$, $\exists\, p(v_0, \ldots, u, v) \triangleright P(\tilde{v_0}, \tilde{v_i})$, where $v_0 \notin V_0$ and v_0 has attribute $\mathcal{A}$. Thus, $p(v_0, \ldots, u) \triangleright P(\tilde{v_0}, \tilde{v}_{i-1})$ and u is a match for $\tilde{v}_{i-1}$. It suffices to prove that $u \notin V_{i-1}$. If $u \in V_{i-1}$, then one can find a path instance $p(v_0, \ldots, u, \ldots, c) \triangleright P(\tilde{v_0}, c)$, which contradicts that $v_0 \notin V_0$. Hence, we conclude that $u \in \overline{V_{i-1}}$. □

By Lemma 5.2, we can reduce the search space for data perturbations to the partial matching set. Also, any partial matching node at level i is linked to a partial matching node at level $i - 1$ via edge type E_{i-1}. Hence, we compute the partial matching sets as follows: we initialize $\overline{V_0}$ as the nodes with attribute $\mathcal{A}$ that are not in V_0, i.e., $\overline{V_0} = \mathcal{V}_{\mathcal{A}} - V_0$ and, at each subsequent level $i > 0$, we obtain $\overline{V_i} = \mathcal{N}(\overline{V_{i-1}}, E_{i-1}) - V_i$.

Relevance-based Reduction. Admissibility-based reduction guarantees that each perturbation from the new search space is admissible. However, if the relevance of a data perturbation is 0, the perturbation does not contribute to the final deviation measure. Therefore, we define relevant perturbations as:

Definition 5.3 (Relevant Data Perturbation). An edge $e = (u, v)$ is a *relevant data perturbation* at level i if e is an admissible data perturbation and $\mathcal{S}(e) > 0$.

Lemma 5.4. *An admissible perturbation $e = (u, v)$ is relevant for edge type E_i of an OF $Q = (\mathcal{A}, \mathcal{X}, P(\tilde{v_0}, c))$ only if $v \in V_{i+1}$.*

Proof. For an admissible perturbation $e = (u, v)$ of type E_i to be head-relevant, there must exist a matching instance $p(v_0, \ldots, u, v, \ldots, c) \triangleright P(\tilde{v_0}, c)$ after adding e, which implies that $\exists\, p(v, \ldots, c) \triangleright P(\tilde{v}_{i+1}, c)$. Note that if $\exists\, u' \in \mathcal{N}(v, E_i^{-1}) \cap V_i$, we can build a path instance p^* by concatenating three path instances, $p_1(v_0, \ldots, u') \oplus p_2(u', E_i, v) \oplus p_3(v, \ldots, c)$. Ergo, $p^* \triangleright P$ and thus $v \in V_{i+1}$; otherwise, the head relevance would be 0. □

Still, to check whether a perturbation $e = (u, v)$ is relevant, we need to calculate if $\mathcal{S}(e) > 0$, which involves finding cycles of length 6 that contain u and v, a computationally expensive task [17, 38]. Besides, there is a massive number of node pairs to check for relevant data perturbation, even for a single level. To enable efficient pruning without cycle enumeration, we decompose the relevance of a data perturbation into head and tail relevance as follows:

Definition 5.5 (Relevant Matching Set V_i^*). $u \in \overline{V_i}$ is *head-relevant* if there exists $r \in V_{i+1}$ and $n \in \mathcal{N}(r, E_i^{-1}) \cap V_i$ with $\mathcal{S}(u, n) > 0$. $u \in \overline{V_i}$ is *tail-relevant* if there exists $r' \in V_{i+1}$ and $n \in \mathcal{N}(u, E_i)$ with $\mathcal{S}(n, r') > 0$. A *relevant matching set* V_i^* for level i contains all head- and tail-relevant nodes.

Remark. Even though $u \in V_i^*$ is not a sufficient condition for a relevant data perturbation $e = (u, v)$ at level i, it reduces the search space from $|\overline{V_i}| \cdot |\mathcal{L}(\tilde{v}_{i+1})|$ to $|V_i^*| \cdot |V_{i+1}|$, where $|\mathcal{L}(\tilde{v}_{i+1})|$ is the number of node instances with the same type as the node variable $\tilde{v}_{i+1}$. This leads to significant computation savings, as we show experimentally in Section 6. Putting everything together, we obtain the following theorem for pruning the search space.

Theorem 5.6. *An edge $e = (u, v)$ is an admissible and relevant data perturbation at level i of an OF Q only if $u \in V_i^*$ and $v \in V_{i+1}$.*

Algorithm for Data Perturbation Space Reduction. Apart from reducing the search space, V_i^* can be efficiently computed from the matching node sets V_i, V_{i+1} and the partial matching set $\overline{V_i}$, in **linear** time. For a given level i with edge type E_i, we compute three sets, namely $\widetilde{V_i}$, $\widehat{V_{i+1}}$ and V_i^*. Figure 5 depicts the relationship among $\widetilde{V_i}$, $\widehat{V_{i+1}}$ and V_i^*. $\widetilde{V_i}$ denotes the set of *head-relevant* nodes, which are computed from the partial matching set $\overline{V_i}$. Directly computing the relevant matching set V_i^* from $\widetilde{V_i}$ is expensive, as it requires enumerating 3-hop paths between nodes from $\widetilde{V_i}$ to V_{i+1}. Hence, we compute the intermediate set $\widehat{V_{i+1}}$, which is obtained by first finding neighbors of $\widetilde{V_i}$ via edge type E_i, i.e., $\mathcal{N}(\widetilde{V_i}, E_i)$, and including a node $u \in \mathcal{N}(\widetilde{V_i}, E_i)$ to $\widehat{V_{i+1}}$ if u shares at least a common neighbor with the nodes in V_{i+1}. As such, any node $u \in \widehat{V_{i+1}}$ serves as a bridge between $\widetilde{V_i}$ and V_{i+1} for the purpose of finding *tail-relevant* nodes. Finally, if a node $u \in \widetilde{V_i}$ connects to another node $n \in \widehat{V_{i+1}}$ via edge type E_i, u is added to V_i^* as it is both head- and tail-relevant.

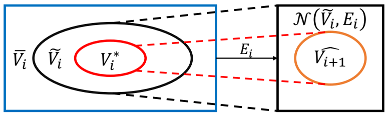


Figure 5: Venn diagram of perturbation space reduction.

Exact Algorithm. The exact algorithm for data perturbation also executes in two stages, as shown in Algorithm 2. In the first stage (Lines 1–4), we obtain the perturbation space by the efficient space reduction just described (and outlined in Appendix B). In the second stage (Lines 5–20), we evaluate all possible data perturbations of $e = (u, v)$ where $u \in V_i^*, v \in V_{i+1}$ at all levels. We calculate the head relevance $\mathcal{S}_h$ in Lines 9–11, followed by computing the tail relevance $\mathcal{S}_t$ in Lines 12–14 when the head relevance is greater than 0. If the perturbation is relevant, i.e., $\mathcal{S}(e) > 0$, we evaluate its strikingness measure in Lines 16–20. We execute a BackwardPass similar to the one in Algorithm 1 to get the perturbed peer entity set V_0'. Since only $e = (u, v)$ is augmented to the KG, we run BackwardPass for $P(\tilde{v}_0, u)$ from level $i < k$ (instead of $P(\tilde{v}_0, c)$ from level k) to get the additional peer entities when considering e. After enumerating the data perturbation space, we normalize the weighted sum of $\mathcal{I}(Q')$ and apply Equation 1 to compute the deviation measure $\Delta(Q)$ in Lines 21–22. As previously, Appendix C provides a time complexity analysis.

Algorithm 2: Exact Algorithm for Data Perturbation

```
   Input: KG $\mathcal{G}$; OF $Q(\mathcal{A}, \mathcal{X}, P(\tilde{v}_0, c))$; matching node sets $\{V_i\}$ for $Q$;
          strikingness score $\mathcal{I}(Q)$
   Output: Deviation measure $\Delta(Q)$
1  $\overline{V_0} \leftarrow \mathcal{V}_{\mathcal{A}} - V_0$;  // $\mathcal{V}_{\mathcal{A}}$ are nodes with attribute $\mathcal{A}$
2  for $i \in \{0, \ldots, k-1\}$ do
3      $\overline{V_{i+1}} \leftarrow \mathcal{N}(\overline{V_i}, E_i) - V_i$;
4      $V_i^* \leftarrow$ SearchSpaceReduction($V_i, V_{i+1}, E_i, \overline{V_i}$);
5  $\mu \leftarrow 0$ , $Z \leftarrow 0$;
6  for $i \in \{0, \ldots, k-1\}$ do
7      foreach $v \in V_{i+1}$ do
8          foreach $u \in V_i^*$ do
9              $e \leftarrow (u, v)$; $\mathcal{S}_h, \mathcal{S}_t \leftarrow 0$;
10             foreach $u' \in V_i \cap \mathcal{N}(v, E_i^{-1})$ do
11                 $\mathcal{S}_h \leftarrow \mathcal{S}_h + \mathcal{S}(u, u')$;
12             if $\mathcal{S}_h > 0$ then
13                 foreach $v' \in \mathcal{N}(u, E_i)$ do
14                     $\mathcal{S}_t \leftarrow \mathcal{S}_t + \mathcal{S}(v, v')$;
15             $\mathcal{S}(e) \leftarrow \mathcal{S}_h \cdot \mathcal{S}_t$;
16             if $\mathcal{S}(e) > 0$ then
17                 $V_0' \leftarrow$ BackwardPass($P(\tilde{v}_0, \tilde{v}_i = u)$);
18                 $\mathcal{I}(Q') \leftarrow$ Strikingness($V_0 \cup V_0', \mathcal{A}, \mathcal{X}$);
19                 $\mu \leftarrow \mu + \mathcal{I}(Q') \cdot \mathcal{S}(e)$;
20                 $Z \leftarrow Z + \mathcal{S}(e)$;
21 $\mathbb{E}_{Q' \sim \mathcal{D}} \leftarrow$ Normalization($\mu, Z$);
22 $\Delta(Q) \leftarrow$ DeviationMeasure($\mathbb{E}_{Q' \sim \mathcal{D}}, \mathcal{I}(Q)$);
23 return $\Delta(Q)$;
```

Sampling-based Approximation. Unlike entity perturbation where the perturbation space is bounded by the number of nodes in the KG in the worst case, the data perturbation space has a quadratic complexity of $O(\sum_i |V_i^*| \cdot |V_{i+1}|)$. Thus, it is often prohibitive to obtain the PRD for data perturbation explicitly. We devise a sampling strategy to estimate the data PRD on the fly. We first randomly sample a level i for the k-hop query P of the OF $Q(\mathcal{A}, \mathcal{X}, P)$. For each sampled level i, we randomly sample u from V_i^* and v from V_{i+1}. Seen as a plausible edge $e = (u, v)$, the sampled node pair has a probability $\mathbb{P}_{\mathcal{U}}(e) = \frac{1}{k \cdot |V_i^*| \cdot |V_{i+1}|}$ under our sampling distribution $\mathcal{U}$ to evaluate the data PRD $\mathcal{D}$. We then execute Lines 7–20 in Algorithm 2 to compute the head-tail relevance $\mathcal{S}(e)$ and the strikingness of the perturbed OF, $\mathcal{I}(Q')$. Since $\mathbb{P}_{\mathcal{D}}(e) \propto \mathcal{S}(e)$ by a normalizing constant $c = \sum_e \mathcal{S}(e)$, we employ importance sampling [42] to estimate $\mathbb{E}_{Q' \sim \mathcal{D}}$ by taking n samples under $\mathcal{U}$ as $\mu_n = \frac{\sum_{j=1}^n \mathcal{I}(Q_j') w(e_j)}{\sum_{j=1}^n w(e_j)}$, where $w(e) = \mathcal{S}(e) / \mathbb{P}_{\mathcal{U}}(e)$. Since all samples are i.i.d., μ_n converges to $\mathbb{E}_{Q' \sim \mathcal{D}}$ with probability 1 [16]; the convergence rate is again $O(1/\sqrt{n})$, as mentioned in Section 4.

6 EXPERIMENTS

In this section, we evaluate the efficiency and effectiveness of our framework. First, we present the experimental setup.

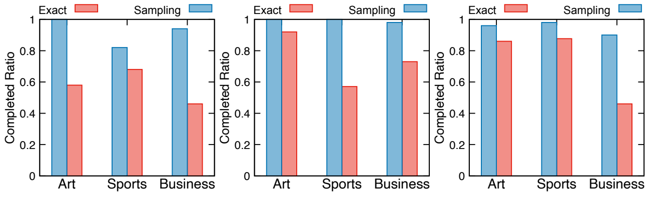
Knowledge Graph (KG) and Outstanding Facts (OFs). We use the Wikidata KG [40] as our data source here, while Appendix G also includes experiments on DBpedia [27]. We extract a subgraph from the Wikidata dump [41] with 17.8M entities and 63.4M edges featuring human-related knowledge. Following existing studies that mine OFs related to celebrities [46, 49], we select 100 celebrities, the top 70 from "The Celebrity 100" [12] and the top 30 from "Forbes Billionaires" [14]. We divide them into three groups, **Art**, **Sports**, and **Business**, based on their profile information. We extract OFs for these celebrities using FMiner [46], the only state-of-the-art method that can scale to large KGs. To extract an OF for a celebrity group, we input a randomly selected celebrity as the "target" into FMiner. We assess the scalability of our perturbation analysis by varying the path query length k from 2 to 4. Note that longer path queries are rare in practice, as they are difficult to interpret in natural language [46]. For each group and path length setting, we mine 50 OFs with a strikingness score of at least 0.8.

Default Sampling Setting. Given that the sizes of the entity and data perturbation spaces can be directly obtained, we set the sample size n for the sampling-based algorithms to $n = rN$, where r is the sampling rate and N is the size of the respective perturbation space. By default, we set $r = 5\%$ for both perturbation regimes.

Hardware Configuration. We conduct all experiments on a Linux server with a 48-core CPU@3.45GHz and 256GB memory. All algorithms are implemented[1] using C++ with -O3 optimization.

6.1 Empirical Study on Entity Perturbation

Efficiency. We execute the exact and sampling algorithms for all OFs under the default setting ($r = 5\%$). Figure 6 presents the ratio of OFs on which the perturbation analysis is completed in 5 minutes. The sampling algorithm processes almost all OFs in 5 minutes. The exact algorithm performs well in certain scenarios, e.g., $k = 3$ for Art, $k = 4$ for Art and Sports. As elaborated in Section 4, a larger perturbation space implies a longer execution time. Table 2 indicates the average size of the entity perturbation space for different celebrity groups and path lengths. Notably, longer paths often lead to a smaller entity perturbation space, as they impose more constraints on the matching instances, rendering OFs with longer paths easier to analyze. For example, both exact and sampling algorithms perform well on the Sports group for $k = 4$, where the size of the perturbation space is only 13% of that for $k = 2$. We also observe that the perturbation space of the Business group is the largest, since the Business OFs contain more "general" edge types (e.g., position held) with numerous edge instances in the associated path queries. As a result, the exact method completes the perturbation analysis for only half of the Business OFs. Nevertheless, even in that challenging setting, the sampling algorithm still analyzes over 90% of the OFs within 5 minutes. Appendix G presents (in Figure 17) the completion ratios for 1-minute and 3-minute time limits, demonstrating similar trends.



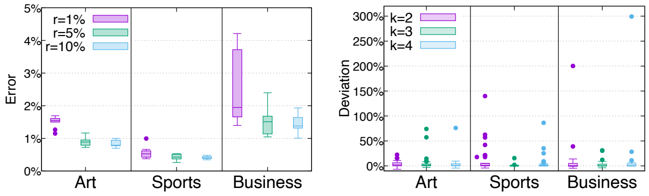
(a) 2-hop Query 5Min **(b) 3-hop Query 5Min** **(c) 4-hop Query 5Min**

Figure 6: Efficiency studies for entity perturbation.

Error Analysis. For a given OF Q, we use $|\Delta^*(Q) - \Delta(Q)|$ to measure the error of the sampling algorithm, where $\Delta^*(Q)$ is the estimation of the true deviation measure $\Delta(Q)$. To evaluate the convergence speed, we vary the sampling rate r, testing three different values: 1%, 5%, and 10%. To measure the error, we must focus on those OFs where the exact algorithm completes the perturbation analysis within 5 minutes. We run the sampling algorithm ten times for each of these OFs and report the mean error per celebrity group and path length. Figure 7a presents the error distribution of 3-hop OFs, while Appendix G provides further results. With increasing sample size (i) the variance of the error decreases, and (ii) the error converges to a small range (2.7% on average) for all celebrity groups. This result follows the convergence analysis in Section 4.

Table 2: Size of entity perturbation space (50 OFs per group).

Group	$k = 2$	$k = 3$	$k = 4$
Art	1.7×10^5	8.3×10^4	1.1×10^5
Sports	4.7×10^5	2.4×10^5	6.5×10^4
Business	1.6×10^6	5.1×10^5	7.8×10^5



(a) Sampling error of 3-hop OFs **(b) OF Deviation**

Figure 7: Effectiveness studies on entity perturbation.

OF Robustness. To scrutinize the problem further, Figure 7b depicts deviation values $\Delta(Q)$ for all mined OFs where the exact algorithm terminates within 5 minutes. Notably, under entity perturbation, the strikingness scores of most OFs deviate marginally from the score expected by the entity PRD, with an average deviation of 6.8%. However, some OFs exhibit high deviation. Recall that FMiner [46] mined the original OFs *orthogonally* to our work. Figure 7b leads to two critical conclusions. First, the low *average* deviation of the OFs suggests the miner's reliability. Meanwhile, outlier OFs with very high deviation evince that OF mining may fumble, thus a rigorous perturbation analysis, as we propose, is vital to assess the robustness of mined OFs before publicizing them. Appendix E contributes a case study on an OF about NBA players.

6.2 Empirical Study on Data Perturbation

Effectiveness of Perturbation Space Reduction. To assess the effectiveness of our space reduction techniques, we report the sizes of the perturbation spaces originally ('Ori'), after admissibility-based reduction ('Adm'), and eventually after relevance-based reduction ('Adm+Rel') in Figure 8. Admissibility-based reduction prunes over 50% of the edges in the original space (note that the intra-bar partitions are in log scale) and relevance-based reduction further prunes over 99% of the surviving edges on average; these results attest the effectiveness of our space reduction approach.

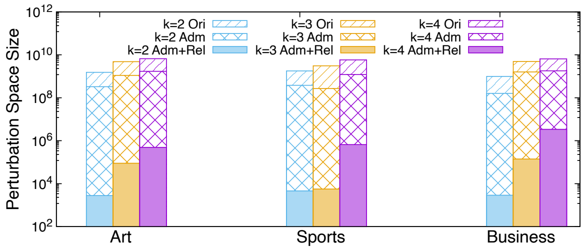
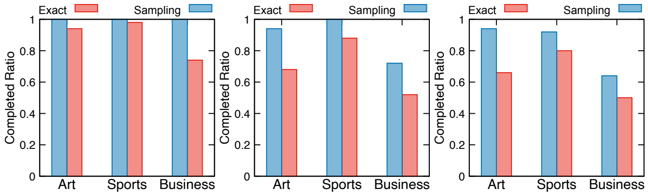


Figure 8: Size of data perturbation space (50 OFs per group).

[1] Code and data are available at https://anonymous.4open.science/r/RobustFacts-0D73/.

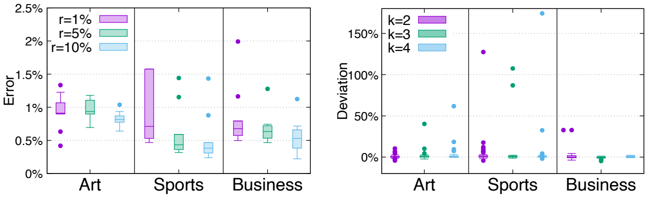
Efficiency. Figure 9 shows the ratio of OFs where data perturbation analysis ends within 5 minutes. The sampling algorithm processes over 90% of the OFs within 5 minutes. Unlike entity perturbation (cf. Figure 6), the completion ratio drops as k grows, because longer paths imply larger search space in data perturbation. In Appendix G, we try 1-minute and 3-minute cutoffs, with aligned findings.



(a) 2-hop Query 5Min **(b) 3-hop Query 5Min** **(c) 4-hop Query 5Min**

Figure 9: Efficiency studies for data perturbation.

Error Analysis. As in Section 6.1, we select OFs where the exact algorithm completes the perturbation analysis in 5 minutes. Figure 10a presents the approximation error for each celebrity group for 3-hop OFs. We defer further results to Appendix G. The estimation converges as we increase the sampling rate. When $r = 10\%$, the error is <1% in most cases. Generally, the error is smaller than entity perturbation because adding an edge to the KG alters the peer entity set much less drastically than matching a different path query. The smaller extent of perturbation naturally implies a smaller error.



(a) Sampling error of 3-hop OFs **(b) OF Deviation**

Figure 10: Effectiveness studies on data perturbation.

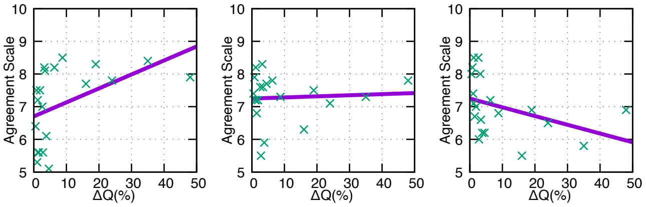
OF Robustness. We present the deviation values for OFs under data perturbation in Figure 10b. The values are generally smaller than entity perturbation because data perturbation is less invasive to the peer entity set, as explained in the previous paragraph. The average deviation under data perturbation is 2.6%. Nonetheless, we also detect OFs that exhibit large deviation and sensitivity to data perturbation, seen as outliers in Figure 10b. To further showcase our method, Appendix F analyzes an OF about US politics.

6.3 Crowdsourced User Study

While existing works on perturbation analysis primarily rely on case studies to demonstrate their efficacy [2, 28, 45], we believe it is valuable to complement case studies with users' perspectives for additional insight. To this end, we recruited 600 crowdsourced participants from the Amazon Mechanical Turk platform to validate the effectiveness of our system for both types of perturbation.

We conduct an ablation analysis for the effectiveness of deviation measures and counterarguments in assessing the robustness of OFs (Appendix E details how deviation measures and counterarguments are used in our case studies). We use 20 OFs with deviation values ranging from 0.4% to 48%. Each participant receives an OF and one of the following: (1) the OF alone; (2) the OF and the corresponding counterargument; or (3) the OF, the counterargument, and the deviation value with an explanation of what the deviation means. Given the information provided, participants are asked to rate their agreement with a generalized statement on a scale from 1 (not likely) to 10 (absolutely likely). Each setting is rated by 10 participants.

Figure 11 shows scatter plots with linear regression lines, which show the average rating by participants of their agreement with the generalized fact under each setting (10 responses for each point). We check statistical significance by the t-test for correlation. The presentation of only an OF to participants (Setting 1) yields a positive correlation between the deviation measure and the average rating (p-value 0.0005), implying a tendency to jump to conclusions from a striking OF. When the OF comes with a counterargument (Setting 2), the regression line slope falls (p-value 0.4643), yet the decline is not statistically significant, indicating that the counterargument alone does not suffice to hinder bias. Finally, when both a counterargument and a deviation measure escort the OF (Setting 3), a negative correlation emerges between rating and deviation measure (p-value 0.0048). The statistically significant shift from positive to negative correlation underscores the effectiveness of our approach in enhancing users' understanding of the context, thereby preventing jumping to conclusions. We defer details to Appendix D.



(a) Setting 1 **(b) Setting 2** **(c) Setting 3**

Figure 11: Crowdsourced user study.

7 RELATED WORK

OF Mining. Early OF mining focused on top-k [10] and skyline queries [5]. With advances in computational journalism and business intelligence, OF mining within data subsets gained traction. Some studies find a local subspace as the context that makes a target data point outstanding [43, 48], while others identify streaks in temporal data [25, 35, 44]. Such approaches work primarily on relational databases, often on a single table, by defining constraints to select context data. OF mining in KGs [49] finds OFs of a target entity based on attribute rarity vs. peer entities defined by a graph pattern. FMiner [46] introduces context entities to extract more relevant OFs. Contrariwise, our work measures the robustness of *the given OFs* mined from KGs.

Perturbation Analysis. Perturbation analysis is commonly used in uncertain data management and mining [1, 24]. Wu et al. [45] applied it to computational fact-checking; they convert facts into parameterized SQL queries and formulate the optimization problem of finding parameter settings that yield the weakest results under parameter sensitivity constraints. Asudeh et al. [2] propose a method that perturbs data points to check whether trend facts in temporal data remain valid. Lin et al. [28] propose perturbation analysis with OLAP operations. Although our perturbation analysis shares the spirit of prior studies, previous methods focus on relational data and are inapplicable to our setting. To our knowledge, this is the first attempt at perturbation analysis on KGs.

8 CONCLUSION

We crafted a methodology to measure the robustness of outstanding facts mined from knowledge graphs by perturbing entities and data. We designed exact and approximate algorithms for both types of perturbation, along with a novel strategy to effectively reduce the data perturbation space and hence the computational cost. Experiments on the Wikidata KG validate the efficiency of our proposals and their effectiveness in mitigating the *jumping conclusion bias.*

REFERENCES

[1] Charu Aggarwal. 2009. *Managing and Mining Uncertain Data.* Springer, New York, NY, USA.

[2] Abolfazl Asudeh, H. V. Jagadish, You Wu, and Cong Yu. 2020. On Detecting Cherry-picked Trendlines. *Proc. VLDB Endow.* 13, 6 (2020), 939–952.

[3] Russa Biswas. 2020. Embedding Based Link Prediction for Knowledge Graph Completion. In *CIKM.* 3221–3224.

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *SIGMOD.* 1247–1250.

[5] Stephan Börzsönyi, Donald Kossmann, and Konrad Stocker. 2001. The Skyline Operator. In *ICDE.* 421–430.

[6] CBINSIGHTS. 2023. Reddit. https://www.cbinsights.com/company/reddit/financials

[7] Daniel Daza, Michael Cochez, and Paul Groth. 2021. Inductive Entity Representations from Text via Link Prediction. In *WWW.* 798–808.

[8] Yanlei Diao, Paweł Guzewicz, Ioana Manolescu, and Mirjana Mazuran. 2021. Efficient Exploration of Interesting Aggregates in RDF Graphs. In *SIGMOD.* 392–404.

[9] Robert M. Entman. 1989. How the Media Affect What People Think: An Information Processing Approach. *The Journal of Politics* 51, 2 (1989), 347–370.

[10] Ronald Fagin, Amnon Lotem, and Moni Naor. 2001. Optimal Aggregation Algorithms for Middleware. In *PODS.* 102–113.

[11] James Field. 2023. Fintech investment advisor Titan slapped with 1M USD fine over misleading ads. https://coingeek.com/fintech-investment-advisor-titan-slapped-with-1m-fine-over-misleading-ads/

[12] Forbes. 2020. The Celebrity 100: The World's Highest-Paid Celebrities 2020. https://www.forbes.com/celebrities/

[13] Forbes. 2021. Reddit IPO: What You Need To Know. https://www.forbes.com/advisor/investing/reddit-ipo

[14] Forbes. 2022. Forbes Billionaires 2022. https://www.forbes.com/billionaires/

[15] Forbes. 2022. NBA Team Value 2022. https://www.forbes.com/sites/mikeozanian/2022/10/27/nba-team-values-2022-for-the-first-time-in-two-decades-the-top-spot-goes-to-a-franchise-thats-not-the-knicks-or-lakers/?sh=4e0367e81cce

[16] John Geweke. 1989. Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica* 57, 6 (1989), 1317–1339.

[17] Roberto Grossi. 2016. Enumeration of Paths, Cycles, and Spanning Trees. In *Encyclopedia of Algorithms.* 640–645.

[18] Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguistics* 10 (2022), 178–206.

[19] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster. In *KDD.* 1803–1812.

[20] Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting Check-Worthy Factual Claims in Presidential Debates. In *CIKM.* 1835–1838.

[21] Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. 2014. Data In, Fact Out: Automated Monitoring of Facts by FactWatcher. *Proc. VLDB Endow.* 7, 13 (2014), 1557–1560.

[22] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. *Proc. VLDB Endow.* 10, 12 (2017), 1945–1948.

[23] M. Shahriar Hossain, Patrick Butler, Arnold P. Boedihardjo, and Naren Ramakrishnan. 2012. Storytelling in Entity Networks to Support Intelligence Analysts. In *KDD.* 1375–1383.

[24] Ravi Jampani, Fei Xu, Mingxi Wu, Luis Perez, Chris Jermaine, and Peter J. Haas. 2011. The Monte Carlo Database System: Stochastic Analysis Close to the Data. *ACM Trans. Database Syst.* 36, 3 (2011), 18:1–18:41.

[25] Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, and Yong Yu. 2011. Prominent Streak Discovery in Sequence Data. In *KDD.* 1280–1288.

[26] Kristy M. Johnstone, Junwen Chen, and Ryan P. Balzan. 2017. An investigation into the jumping-to-conclusions bias in social anxiety. *Conscious. Cogn.* 48 (2017), 55–65.

[27] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

[28] Yin Lin, Brit Youngmann, Yuval Moskovitch, H. V. Jagadish, and Tova Milo. 2021. On Detecting Cherry-picked Generalizations. *Proc. VLDB Endow.* 15, 1 (2021), 59–71.

[29] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized News Recommendation Based on Click Behavior. In *IUI.* 31–40.

[30] Anaïs Llorens, Athina Tzovara, Ludovic Bellier, Ilina Bhaya-Grossman, Aurélie Bidet-Caulet, William K Chang, Zachariah R Cross, Rosa Dominguez-Faus, Adeen Flinker, Yvonne Fonken, et al. 2021. Gender bias in academia: A lifetime problem that needs solutions. *Neuron* 109, 13 (2021), 2047–2074.

[31] Chien-Chun Ni, Kin Sum Liu, and Nicolas Torzec. 2020. Layered Graph Embedding for Entity Recommendation Using Wikipedia in the Yahoo! Knowledge Graph. In *WWW.* 811–818.

[32] Reddit. 2022. All else equal, do you think small markets teams would prefer international players as their stars? https://www.reddit.com/r/nba/comments/vz0kgp/all_else_equal_do_you_think_small_markets_teams/

[33] Reddit. 2023. A little help here... Title: Research Finds No Gender Bias in Academic Science. https://www.reddit.com/r/MensRights/comments/1352zux/a_little_help_here_title_research_finds_no_gender/

[34] Susan Singleton. 2008. The Consumer Protection from Unfair Trading Regulations 2008 and IT/Internet-viral and buzz marketing issues. *Communications law (Haywards Heath)* 13, 4 (2008), 117–119.

[35] Afroza Sultana, Naeemul Hassan, Chengkai Li, Jun Yang, and Cong Yu. 2014. Incremental Discovery of Prominent Situational Facts. In *ICDE.* 112–123.

[36] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *IJCAI.* 4396–4402.

[37] New York Times. 2020. Kamala Harris Makes History as First Woman and Woman of Color as Vice President. https://www.nytimes.com/2020/11/07/us/politics/kamala-harris.html

[38] Hanghang Tong, Christos Faloutsos, Brian Gallagher, and Tina Eliassi-Rad. 2007. Fast Best-Effort Pattern Matching in Large Attributed Graphs. In *KDD.* 737–746.

[39] Joris van Hoboken and Ronan Ó Fathaigh. 2021. Regulating Disinformation in Europe: Implications for Speech and Privacy. *UC Irvine J. Int'l Transnat'l & Comp. L.* 6 (2021), 9.

[40] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[41] Wikidata. 2023. Wikidata Dumps. https://dumps.wikimedia.org/wikidatawiki/entities/

[42] Wikipedia. 2023. Importance Sampling. https://en.wikipedia.org/wiki/Importance_sampling

[43] Tianyi Wu, Dong Xin, Qiaozhu Mei, and Jiawei Han. 2009. Promotion Analysis in Multi-Dimensional Space. *Proc. VLDB Endow.* 2, 1 (2009), 109–120.

[44] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2012. On "One of the Few" Objects. In *KDD.* 1487–1495.

[45] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward Computational Fact-Checking. *Proc. VLDB Endow.* 7, 7 (2014), 589–600.

[46] Yueji Yang, Yuchen Li, Panagiotis Karras, and Anthony K. H. Tung. 2021. Context-Aware Outstanding Fact Mining from Knowledge Graphs. In *KDD.* 2006–2016.

[47] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for Knowledge Graph Completion. arXiv:1909.03193 [cs.CL]

[48] Gensheng Zhang, Xiao Jiang, Ping Luo, Min Wang, and Chengkai Li. 2014. Discovering General Prominent Streaks in Sequence Data. *ACM Trans. Knowl. Discov. Data* 8, 2 (2014), 9:1–9:37.

[49] Gensheng Zhang, Damian Jimenez, and Chengkai Li. 2018. Maverick: Discovering Exceptional Facts from Knowledge Graphs. In *SIGMOD.* 1317–1332.

[50] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1 (2019), 5:1–5:38.

A HEAD AND TAIL RELEVANCE RATIONALE

In Figure 12, edge e = (‘Alibaba Group’, ‘Reddit’) may be a probable data perturbation for the Reddit OF based on head relevance; both Alibaba Group and Tencent are Chinese Internet conglomerates and share common information in the KG, such as country, stock exchange, and invested companies. However, the retail companies invested by Alibaba, such as Farfetch, Sun Art Retail, and Fanatics, are less relevant to Reddit, suggesting that Alibaba is unlikely to invest in Reddit. Tail relevance addresses the discrepancy by downgrading the relevance score of e because of the relatively lower relevance of the companies invested by Alibaba.

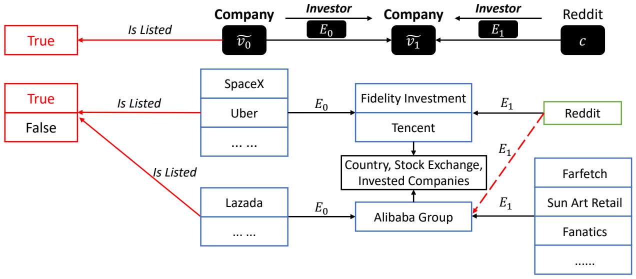


Figure 12: An example of head relevance only.

In Figure 13, taking tail relevance into account, e' = (‘Sequoia Capital’, ‘Reddit’) emerges as a more suitable data perturbation. Among Sequoia Capital’s backed companies in the KG, there are several US social networking platforms like Instagram and LinkedIn, which share common services with Reddit. Additionally, both Reddit and Google have partnerships with the Entertainment Consumers Association and membership in the Internet Association. This information strongly suggests Sequoia Capital’s investment in Reddit. Introducing e' to the KG incorporates businesses that have received investments from Sequoia Capital into the peer entity set, contradicting the original OF. For instance, Nvidia is publicly listed.

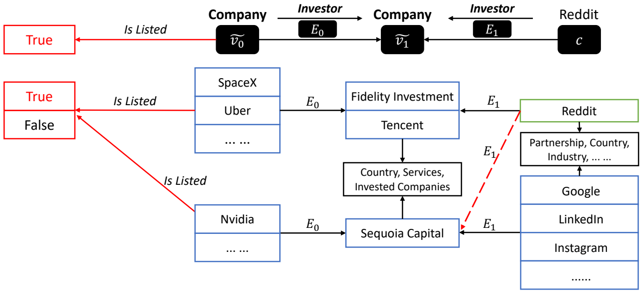


Figure 13: An example of head-tail relevance.

B SPACE REDUCTION ALGORITHM

Algorithm 3 illustrates the combined process for admissibility-based and relevance-based reductions.

C COMPLEXITY ANALYSIS

Exact Method for Entity Perturbation. The time complexity of the ForwardPass stage is upper bounded by $O(\sum_i |E_i|)$, where $|E_i|$ is the number of edges of type E_i in the path query. Thus, the complexity of the exact algorithm is dominated by the BackwardPass stage, i.e., upper bounded by $O(|Q| \cdot \sum_i |E_i| \cdot J)$, where $|Q|$ is the size of the entity perturbation space and J is the average time to evaluate $\mathcal{S}(c, c')$. In practice, the cost of J is low using efficient set intersections between the neighbor lists of c and c'.

Algorithm 3: Data Perturbation Space Reduction

```
   Input: $V_i, V_{i+1}, E_i, \overline{V_i}$
   Output: $V_i^*$
1  $\widetilde{V_i}, \widehat{V_{i+1}}, V_i^* \leftarrow \emptyset$;
2  foreach $v \in \overline{V_i}$ do
3      foreach $n \in \mathcal{N}(v)$ do
4          if $n \in \mathcal{N}(V_i)$ then
5              $\widetilde{V_i} \leftarrow \widetilde{V_i} \cup \{v\}$; break;
6  foreach $v \in \mathcal{N}(\widetilde{V_i}, E_i)$ do
7      foreach $n \in \mathcal{N}(v)$ do
8          if $n \in \mathcal{N}(V_{i+1})$ then
9              $\widehat{V_{i+1}} \leftarrow \widehat{V_{i+1}} \cup \{v\}$; break;
10 foreach $v \in \widetilde{V_i}$ do
11     foreach $n \in \mathcal{N}(v, E_i)$ do
12         if $n \in \widehat{V_{i+1}}$ then
13             $V_i^* \leftarrow V_i^* \cup \{v\}$; break;
14 return $V_i^*$;
```

Data Perturbation Space Reduction. In the worst case, producing V_i^* only requires a *constant* number of linear scans on the neighbors of the matching set V_i and the partial matching set $\overline{V_i}$ in each level i. Let $\mathcal{E}(V)$ denote the set of edges that has one end in a node set V. The complexity of admissibility-based reduction is bounded by $O\left(\sum_i |\mathcal{E}(V_i)|\right)$, since we can incrementally compute $\overline{V_{i+1}}$ from $\overline{V_i}$. To support efficient edge look-ups in Lines 4 and 8, we build a hash table on $\mathcal{N}(V_i)$ for each $i = \{0, \ldots, k\}$ in $O(\sum_i |\mathcal{E}(V_i)|)$ time. Similarity, efficient look-ups in Line 12 rely on a hash table on $\widehat{V_{i+1}}$, built in $O\left(\sum_i |\overline{V_i}|\right)$ time as $\widehat{V_{i+1}} \subseteq \overline{V_{i+1}}$. At each level i, computing $\widetilde{V_i}$, $\widehat{V_{i+1}}$ and V_i^* requires scanning the neighbors of the nodes in $\overline{V_i}$ and $\overline{V_{i+1}}$ in the worst case. Thus, the complexity of the relevance-based reduction is bounded by $O\left(\sum_i |\mathcal{E}(\overline{V_i})|\right)$. Combining admissibility-based and relevance-based reductions yields a worst-case time complexity of $O\left(\sum_i (|\mathcal{E}(V_i)| + |\mathcal{E}(\overline{V_i})|)\right)$.

Exact Method for Data Perturbation. The complexity of the first stage is $O(\sum_i (|\mathcal{E}(V_i)| + |\mathcal{E}(\overline{V_i})|))$. For the second stage, at each level i, the number of potentially relevant data perturbations is $|V_i^*| \cdot |V_{i+1}|$. For each such perturbation, the time to compute the head and tail relevance is bounded by $O(|E_i| \cdot J)$, where J is the average time to evaluate $\mathcal{S}(c, c')$. The BackwardPass and strikingness evaluation take $O(\sum_{j<i} |E_j|)$ time. To perform this process at all levels, the exact algorithm runs in $O(\sum_i (|V_i^*| \cdot |V_{i+1}|) \cdot \sum_{j<i} |E_j| \cdot |E_i| \cdot J)$ time.

D STATISTICAL SIGNIFICANCE TESTING

Hypothesis Testing for Each OF. To validate the effectiveness of our user study, we conduct hypothesis testing using a T-test for all OFs under Setting 1 (Participants receive the OF only) and under Setting 3, which represents the complete support our framework renders (Participants receive the OF, the counterargument and the deviation score with an explanation of what the score means). Here are the settings for hypothesis testing:

- X denotes the mean scale of the agreement from participants with a generalized statement under Setting 1;

- Y denotes the mean scale of the agreement from participants with a generalized statement under Setting 3;
- Number of samples $n = 10$ for each setting, since each setting is rated by 10 participants;
- Null Hypothesis (H_0): diff = $Y - X = 0$;
- Alternative Hypothesis 1 (H_1): diff < 0;
- Alternative Hypothesis 2 (H_2): diff > 0.

Table 3: T-test results for OFs.

$\Delta(Q)$	X: Mean/Stddev	Y: Mean/Stddev	p-value of H_1	p-value of H_2
48%	7.9/1.371	6.9/1.197	**0.0498**	0.9502
35%	8.2/1.475	5.8/2.251	**0.0063**	0.9937
24%	7.8/1.032	6.5/1.779	**0.0325**	0.9675
19%	8.3/1.766	6.6/1.265	**0.0124**	0.9876
16%	7.7/1.888	5.5/1.271	**0.0039**	0.9961
8.8%	8.5/1.509	6.8/1.932	**0.0213**	0.9787
6.3%	8.1/1.197	7.1/1.101	**0.0339**	0.9661
4.5%	8.1/1.663	6.7/1.449	**0.0052**	0.9948
3.8%	6.1/1.101	6.2/1.549	0.5651	0.4349
3.4%	8.0/1.333	6.7/1.494	**0.0276**	0.9724
3.2%	8.2/1.135	7.8/1.398	0.2459	0.7541
2.8%	5.6/1.349	6.0/1.414	0.7371	0.2629
2.6%	8.5/1.433	7.0/1.699	**0.0237**	0.9763
1.9%	7.5/1.354	7.0/1.491	0.2213	0.7787
1.5%	7.2/1.751	7.2/1.229	0.5	0.5
1.3%	5.6/1.897	8.5/1.354	0.9994	**0.0006**
1.2%	5.6/1.712	6.8/1.135	0.9581	**0.0419**
1.0%	5.3/2.002	7.4/1.349	0.9928	**0.0072**
0.8%	7.3/1.159	8.3/1.251	0.9598	**0.0402**
0.5%	6.4/1.506	8.0/1.491	0.9859	**0.0141**

According to the T-test results in Table 3, for all OFs with deviation greater than 4.5% (recall that OFs with high deviation are not robust), the observed decrease in agreement levels between Settings 1 and 3 is statistically significant since the p-values of H_1 are smaller than 0.05; this indicates that our methodology can effectively prevent users from jumping to conclusions under unrobust OFs, which is the main objective of our work. Furthermore, for OFs with deviation smaller than 1.5% (recall that OFs with low scores are robust), we observe that the average scale of the agreement of the participants increases and that increase is statistically significant because the p-values of H_2 are smaller than 0.05; this suggests that our approach improves users' understanding of the context of OFs, making them more confident about robust OFs. In general, the results of the T-test validate the effectiveness of our approach.

Hypothesis Testing for the Slope of the Regression Line. We run regression analysis on the **200 responses** for each setting (each response is a data point) and conduct hypothesis tests for the regression slope under the 3 settings. Here are the settings for hypothesis testing:

- Null Hypothesis (H_0): slope = 0;
- Alternative Hypothesis 1 (H_1): slope < 0;
- Alternative Hypothesis 2 (H_2): slope > 0.

Table 4: T-test results for slopes.

Setting	p-value of H_1	p-value of H_2
1	**0.0005**	0.9995
2	0.4643	0.5357
3	0.9952	**0.0048**

Table 4 presents the p-values of all settings. The positive slope for Setting 1 is statistically significant since its p-value of H_1 is smaller than 0.05. Also, the negative slope for Setting 3 is statistically significant since its p-value of H_2 is smaller than 0.05. These results corroborate our findings in Figure 11 that "*The shift from positive to negative correlation is statistically significant and underscores the effectiveness of our approach in enhancing users' understanding of the context, thereby preventing jumping to conclusions.*"

E CASE STUDY FOR ENTITY PERTURBATION

OFs are frequently used as evidence to bolster generalizations in social platform discussions. An intriguing subject that arises on Reddit is whether small-market teams demonstrate a preference to draft international players [32]. One could argue that no such preference exists, based on the generalization of the following OF: Memphis Grizzlies, with the 29th market size among a total of 30 NBA teams [15], have only drafted 6 foreign players, with a strikingness of $I(Q) = 0.91$. Figure 14 shows the associated path query. Through our entity perturbation analysis, we derive an expected strikingness score of $\mathbb{E}_{Q' \sim \mathcal{D}} = 0.78$, which is substantially lower than 0.91, with deviation measured at $\Delta(Q) = 16\%$. Furthermore, we return a counterargument to the OF by identifying the perturbation Q' that maximizes $(I(Q) - I(Q')) \cdot \mathbb{P}_{\mathcal{D}}(Q')$, where $\mathcal{D}$ is the PRD for entity perturbation. The retrieved counterargument is a relevant but less striking perturbation than the original OF. Specifically, it suggests San Antonio as the context entity to replace Memphis. San Antonio is related to Memphis since both belong to the southwest NBA division and both are small-market teams (San Antonio ranking 20th among NBA teams). However, the San Antonio Spurs have drafted 25 foreign players, a fact that brings the strikingness down to 0.72, revealing that the original OF is not robust under entity perturbation.

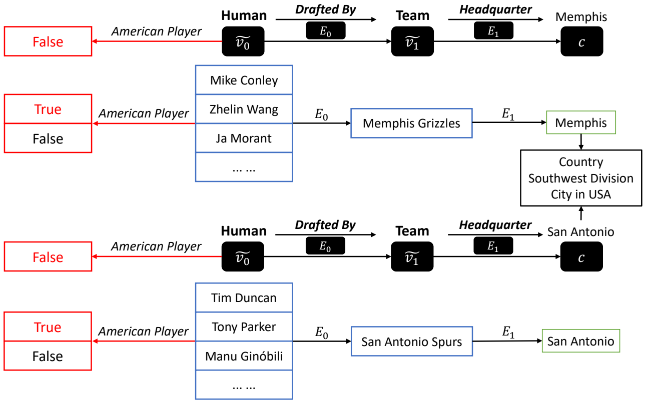


Figure 14: Case study for entity perturbation.

F CASE STUDY FOR DATA PERTURBATION

Gender parity is always a controversial topic. We consider an OF regarding Janet Yellen, the current United States Secretary of the Treasury, who also held the Chair of the Federal Reserve from 2014 to 2018. According to the Wikidata KG, among all politicians who have held any of Yellen's positions, Yellen is the **only** woman vis-à-vis 91 men. The strikingness score of this OF is $I(Q) = 0.99$. Figure 15 depicts the associated path query. This uniqueness could spawn a radical generalization that the gender gap in US politics is vast. However, based on data perturbation, the expected strikingness is only $\mathbb{E}_{Q'\sim\mathcal{D}} = 0.91$ (down from the $I(Q)$ of 0.99). Also, the associated deviation is $\Delta(Q) = 8.8\%$, which is large according to the data perturbation standards. Indeed, our methodology suggests a data perturbation (edge addition) which turns out to be valid in reality but missing from the KG, and which refutes the original OF. Specifically, Yellen also held the Chair of the Council of Economic Advisers, which adds three more female politicians to her peer entity set, namely, Laura Tyson, Christina Romer, and Cecilia Rouse. The addition brings the strikingness of the OF down to 0.95. Looking at the inner workings of our methodology, this data perturbation was identified as highly relevant to the original OF for a couple of reasons. First, both the US Secretary of the Treasury and the Chair of the Council of Economic Advisers are US Cabinet positions. Second, high-profile politicians like Ben Bernanke and Alan Greenspan, as well as renowned economists such as Joseph E. Stiglitz, have held both of these positions. Interestingly, Yellen was a student of Stiglitz.

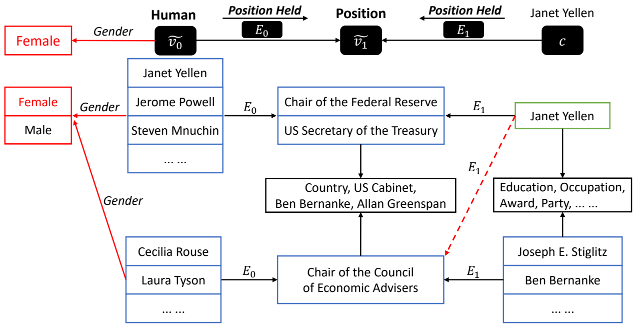


Figure 15: Case study for data perturbation.

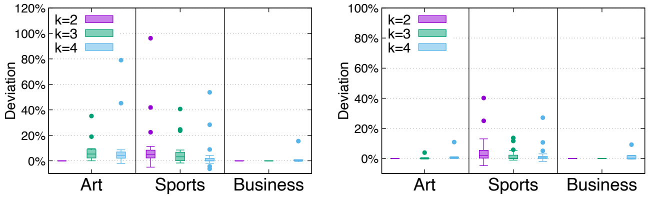
G SUPPLEMENTARY EXPERIMENTS

Experiments on DBpedia. Our method is generic and can be applied to other knowledge bases. To support our claim, we collect a knowledge graph with 5.2M entities and 26M edges from the DBpedia [27] knowledge base. We first feed the celebrities into FMiner [46] for OF mining. However, there is much less human-related information in this dump as well as missing attributes compared to Wikidata [40] (17.8M entities and 63.4M edges). Because of this, FMiner cannot identify 50 valid OFs (strikingness score > 0.8) for each setting we consider, so we are bound to use fewer. Table 5 presents the exact number of valid OFs we use per setting.

Table 5: Valid OFs of DBpedia.

Setting	Number of OFs
Art, k = 3	20
Art, k = 4	30
Sports, k = 2	50
Sports, k = 3	50
Sports, k = 4	50
Business, k = 4	10

In Figure 16, we present the distribution of OF deviation values per group, under entity perturbation and data perturbation. The results are similar to Figure 7b and Figure 10b in the main paper. Most OFs marginally deviate from their expected score and there are some outliers in each group. Also, the deviation values under data perturbation are generally smaller than entity perturbation since data perturbation is less invasive to the peer entity set. This is consistent with our findings on Wikidata.



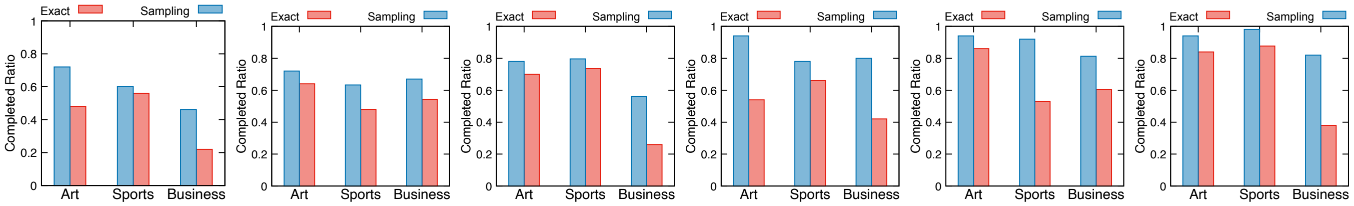
(a) Entity Perturbation

(b) Data Perturbation

Figure 16: OF Deviation for DBpedia OFs.

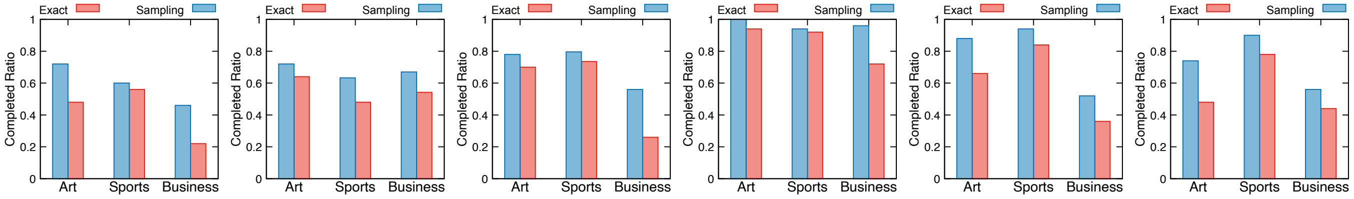
Runtime Analysis for Wikidata OFs. Figures 17 and 18 show the ratio of OFs where the entity and data perturbation analysis completes within 1 and 3 minutes, respectively. Similarly, we conclude that the completion ratio decreases for a larger k in data perturbation, which is consistent with the analysis in Section 6.

Error Analysis for Wikidata OFs. Figures 19 and 20 present the error distribution for each celebrity group under different path length settings for entity and data perturbation, respectively. In line with the experiments in Section 6, the error variance decreases as the sample size increases and the error converges for all celebrity groups.



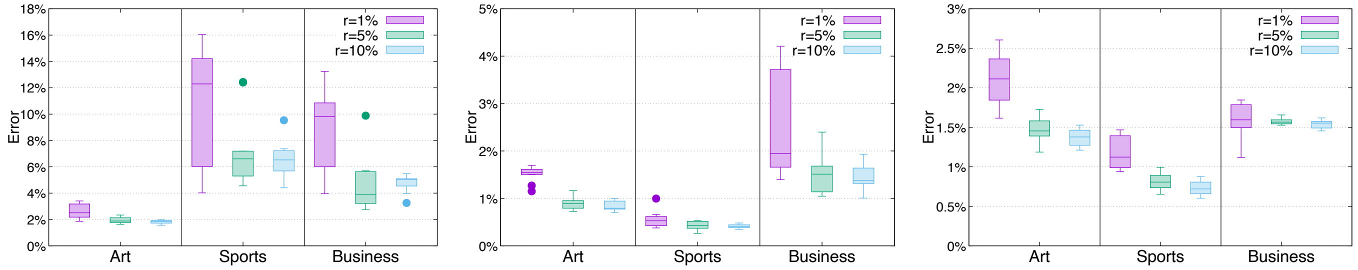
(a) 2-hop Query 1Min (b) 3-hop Query 1Min (c) 4-hop Query 1Min (d) 2-hop Query 3Min (e) 3-hop Query 3Min (f) 4-hop Query 3Min

Figure 17: Efficiency studies for entity perturbation within 1 and 3 minutes.



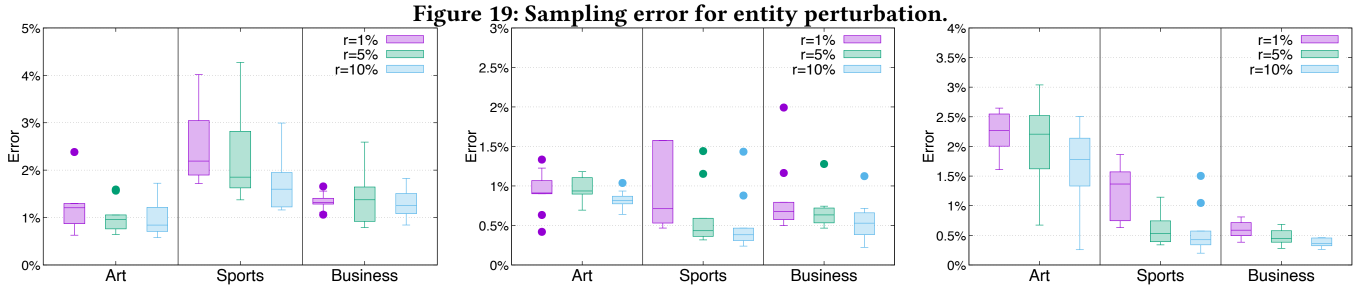
(a) 2-hop Query 1Min (b) 3-hop Query 1Min (c) 4-hop Query 1Min (d) 2-hop Query 3Min (e) 3-hop Query 3Min (f) 4-hop Query 3Min

Figure 18: Efficiency studies for data perturbation within 1 and 3 minutes



(a) 2-hop Query (b) 3-hop Query (c) 4-hop Query

Figure 19: Sampling error for entity perturbation.



(a) 2-hop Query (b) 3-hop Query (c) 4-hop Query

Figure 20: Sampling error for data perturbation.