Assessing the Stability of Rankings in Knowledge Graphs Against Perturbations

Hassan Abdallah, Béatrice Markhoff, Louise Parkin, Arnaud Soulet

Assessing the Stability of Rankings in Knowledge Graphs Against Perturbations

Hassan Abdallah[1][0009−0000−5119−6523], Béatrice Markhoff[2][0000−0002−5171−8499], Louise Parkin[1][0000−0002−6522−3526], and Arnaud Soulet[1][0000−0001−8335−6069]

[1] LIFAT, University of Tours, 3 Pl. Jean Jaurès, 41000 BLOIS, France
{hassan.abdallah,louise.parkin,arnaud.soulet}@univ-tours.fr
[2] UMR 7324 CITERES, CNRS and University of Tours
beatrice.markhoff@univ-tours.fr

Abstract. Knowledge graphs such as Wikidata serve as valuable resources for structuring and analyzing information across various domains. However, their crowdsourced nature makes them vulnerable to perturbations, including both intentional vandalism and unintentional errors, which can significantly impact rankings derived from these graphs. While previous studies have primarily focused on detecting and preventing entity-level perturbations, this paper investigates the potential impact of such perturbations on the stability of rankings at the structural level, specifically targeting relationships. We formalize the problem of ranking stability under perturbations, and we propose a probabilistic model to assess the likelihood of modifications to knowledge graph relationships changing the ranks of entities. We leverage complex network analysis, in order to evaluate ranking vulnerabilities. Our experimental study demonstrates varying levels of resilience in rankings depending on entity degree distributions and the nature of perturbations.

1 Introduction

Large crowdsourced Knowledge Graphs (KGs) [4,14,16] available on the Web are important resources for knowledge workers, especially when they are shared within communities and managed as commons. To take advantage of this vast amount of knowledge, multiple tools exist to support analysis. For instance, ranking indicators help in understanding and evaluating the significance of entities in various domains, from scientific research to cultural impact as they provide a structured method to order entities based on defined criteria, supporting informed decision-making [6]. In the era of big data, knowledge graphs serve as an invaluable resource for building and refining these rankings [1].

Analytics based on crowdsourced knowledge graphs such as Wikidata [16] draw their strength from the richness of facts contributed by a global community of editors. Yet this crowdsourced foundation also raises a concern: How correct and complete is the data used to build the analyses? The ease of contributing to a crowdsourced KG increases the risk of introducing perturbations – either through deliberate vandalism (the insertion of false statements, or deletion of

Relationship	Nb. of perturbations
instance of	8,872
sex or gender	2,409
occupation	2,013
country	1,527
contains the administrative territorial entity	1,102
country of citizenship	1,001
given name	732
member of sports team	635
place of birth	602
located in the administrative territorial entity	487

Table 1: The 10 most perturbed relationships in Wikidata (2012-2016)

truthful ones) or through unintentional errors arising during routine editing. In the context of rankings, vandalism may target specific entities with the intent to manipulate their perceived importance or disrupt the ranking at a global level. Similarly, unintentional errors, particularly those arising from batch updates using scripts, can introduce biases or inaccuracies that significantly impact the rankings. Based on the corpus of Wikidata edits flagging vandalism and errors [9], Table 1 shows the top ten most perturbed relationships in Wikidata from 2012 to 2016, and reveals that certain relationships are more frequently targeted by perturbations than others. For example, many of the modifications involving the *place of birth* relationship are likely mistakes, whereas the *member of sports team* relation has been affected by edit wars surrounding football and rugby. From May 2013 to June 2015 ten updates falsely linked players such as *Davide Moscardelli* or *Paul Pogba* to *Barcelona*, while thirty-five scattered edits in the same period reassigned high-profile names including *Lionel Messi*, *Neymar*, and *Marco Reus* to *Real Madrid CF*. Although these updates are spread over two years, their asymmetric counts underscore a persistent attempt to elevate one club's prominence. By contrast, rugby vandalism appears in a tight burst: on 17 February 2015 eight new items were created that miscategorised historical players as members of the *England national rugby union team*, and less than a month later (11 March 2015) eight legitimate links to the *Scotland national rugby union team* were removed. The near-simultaneous timestamps and equal magnitudes suggest a coordinated campaign to inflate England's importance at Scotland's expense. Such vulnerabilities highlight a critical research problem: the resilience of knowledge graphs in maintaining stable rankings under conditions of vandalism and accidental errors. Thus, this work aims to answer the following research question: *Do certain knowledge graph relationships possess inherent robustness that ensures the integrity of its induced ranking against vandalism and errors?*

Existing literature on vandalism and errors primarily focuses on entry-point control, i.e. detecting and reverting bad edits as they occur [10,11,12,15]. In this paper, we adopt a complementary perspective, examining how malicious or erroneous edits impact derived structures, in particular ranked lists of entities.

The question becomes not just whether an edit is a perturbation, but to what extent it can distort the perceived importance of entities if left uncorrected. More precisely, this paper proposes a model to study the link between the number of facts of an entity for a relationship and its stability in the induced ranking against perturbations. Interestingly, this model thus makes it possible to define a threshold of changes beyond which vandalism is highly probable. Inspired by the work of Albert et al. [3], which examined the robustness of complex networks under various conditions, our approach integrates insights from graph analysis, network science, ranking indicators, and probabilistic modeling. We address the two threats of vandalism and errors with the following contributions:

- **Formalization of ranking stability issues:** We present a formalization to examine how vandalism and errors can alter rankings, providing a theoretical foundation to study these phenomena systematically.
- **Probabilistic attack model:** A theoretical model is proposed to quantify the likelihood that attacks or errors will successfully modify an entity's rank, allowing for predictive assessments of ranking vulnerabilities.
- **Empirical evaluation:** We conduct an experimental study using open crowdsourced knowledge graphs to evaluate the stability of rankings. This study not only validates the theoretical model but also sheds light on the practical implications of ranking stability.

In Section 2, we discuss related work. Section 3 defines key preliminary notations, while Section 4 formalizes the problem. In Section 5, we introduce perturbation scenarios used as models of perturbations. Section 6 presents the theoretical results of a probabilistic model. Section 7 details the experimental study, and finally, Section 8 concludes this work.

2 Related Work

In recent years, numerous studies have focused on analyzing the stability of crowdsourced data, which often sparks debate due to its nature of being generated by individuals dispersed globally, many of whom do not know each other. Halpin et al. [7] examined the stability of distributions in crowdsourced tagging systems, discovering that the frequency distribution of tags follows a power-law pattern. They further analyzed the underlying dynamics that lead to this stable distribution. Given that knowledge graphs (KGs) act as a mirror of the real world, which is dynamic and continuously evolving, Shrinivasan and Razniewski [13] highlighted the inherent instability of KGs. Their work investigated the concept of knowledge base stability, specifically examining how KGs are affected by real-world changes. On the other hand, detecting and mitigating vandalism in knowledge graphs, particularly in Wikidata, has garnered significant attention in recent years. A predominant approach involves supervised learning techniques that leverage labeled datasets to train detection models. Heindorf et al. [10] introduced a machine learning-based method for identifying vandalism in Wikidata, proposing 47 features that exploit the crowdsourced nature of KGs to enhance

detection accuracy. Later, Heindorf et al. [11] addressed biases in vandalism detection models, analyzing their sources and developing a model that reduces bias while maintaining fine predictive performance. Sarabadani et al. [12] extended this research by building automated tools tailored for vandalism detection in Wikidata. Their method, while effective, relies on contextual features of user edits and heavily on specific user behaviors, which introduces potential biases. More recently, Trokhymovych and Saez-Trumper [15] proposed a system that aids the Wikidata community in detecting vandalism using advanced feature engineering techniques. Their work underscores the importance of structural modifications in knowledge triples and their potential implications. Despite these advancements, a significant gap remains in the literature regarding the probabilistic analysis of structural manipulations in KG relationships. Such manipulations, whether through errors or targeted attacks, can subtly alter relationships to influence entity rankings, artificially inflating or deflating an entity's perceived importance. While existing studies primarily focus on detecting and preventing overt acts of vandalism or evaluating stability of distribution generally, this paper examines the specific effect of perturbations on rankings induced by ranking indicators [1]. Addressing this challenge is particularly complex in scenarios where ground truth data is unavailable, necessitating novel methodologies that leverage the inherent patterns of data distributions to help safeguarding against structural manipulations, without relying solely on supervised learning. To support research in this domain, Heindorf et al. [9] developed the Wikidata Vandalism Corpus (WDVC-2015), the first corpus providing ground truth for vandalism in Wikidata. This resource has been instrumental in training detection models. Subsequently, the Wikidata Vandalism Corpus 2016 (WDVC-2016) was introduced [8], an updated version of WDVC-2015. In our work, we kept modifications targeting relationships in WDVC-2016 (see Section 7) and extracted several cases of attacks resulting in altering entities' ranks.

3 Preliminaries

Knowledge graph Considering distinct infinite sets I and L (IRIs [5] and literals, respectively), a knowledge graph $\mathcal{K} \subseteq I \times I \times (I \cup L)$ is a set of facts. Each fact is a triple $\langle s, p, o \rangle \in \mathcal{K}$, where s, p and o denote respectively the *subject*, the *predicate* (or relationship) and the *object*. An example of fact is: $\langle$`Kylian Mbappe`, `member of sports team`, `Real Madrid CF`$\rangle$.

Given a relationship $p \in \mathcal{K}$, $\mathcal{K}_p$ is the set of facts in $\mathcal{K}$ having p as relationship: $\mathcal{K}_p = \{\langle s, p', o \rangle \in \mathcal{K} \ : \ p' = p\}$. Thereafter, we work mostly on a single relationship p at a time (e.g., $\mathcal{K}_{\texttt{member of sports team}}$ selects all the facts about members of sports teams). Given a relationship p, its number of facts in $\mathcal{K}_p$ is denoted by n, and its number of subjects (resp. objects) in $\mathcal{K}_p$ is denoted by n_s (resp. n_o). For formulas that work with both subjects and objects, we use n_e to denote the number of entities. Finally, we denote $p_s = n_s/n$ and $p_o = n_o/n$.

Ranking induced by a relationship Any relationship p induces a ranking over a set of entities $E = \{e_1, e_2, \ldots, e, \ldots, e_N\}$. The rank of an entity e in such a ranking

is determined by its degree k_e defined as the number of facts in $\mathcal{K}_p$ associated with e (as subject or object). Formally, the ranking is represented as an ordering $\sigma_p : E \to \{1, \ldots, N\}$, where $\sigma_p(e)$ denotes the rank of e based on k_e such that $k_{e_1} > k_{e_2} \implies \sigma_p(e_1) < \sigma_p(e_2)$. As shown in Figure 1, an entity with more facts for a given relationship p is ranked higher. In the following (and in Figure 1), we omit references to the relationship p.

Rank change Let Δk_e denote the modification in the degree of entity e for relationship p resulting from a perturbation, such that $k'_e = k_e + \Delta k_e$, where k'_e represents the updated degree after the perturbation. The rank change $\Delta\sigma(e)$ can be expressed as: $\Delta\sigma(e) = \sigma'(e) - \sigma(e)$ where $\sigma'(e)$ is the updated rank of e in the perturbed distribution. Figure 1 shows rank changes after perturbations.

Probability Mass Function (PMF) A fundamental concept is the Probability Mass Function (PMF), which describes the probability distribution of entity degrees with respect to a specific relationship p. The PMF captures the likelihood that an entity in the knowledge graph has a particular degree, enabling analysis of rankings. The Probability Mass Function denoted as $P(k)$ is a function that assigns to each degree k the probability that a randomly selected entity e has exactly k facts with the relationship p: $P(k) = \Pr(k_e = k)$.

4 Problem Formulation

Our objective is to quantify the robustness of a ranking under perturbations caused by either the addition or the removal of Δk_e facts in $\mathcal{K}_p$. Specifically, we aim to:

1. Analyze the impact of such perturbations on both individual entities and the overall ranking, providing insights into the stability and vulnerability of the ranking.
2. Model the probability that an entity e, initially ranked at position $\sigma(e)$, with a degree k_e will experience a rank change $\Delta\sigma(e)$ due to a perturbation of size Δk_e targeting that entity: $\Pr(\Delta\sigma(e) \neq 0 \mid \Delta k_e, k_e)$.

Thus our goal is twofold. Firstly, characterize the processes of perturbations due to vandalism or errors. This enables the empirical analysis of these processes, the extraction of statistics and insights about the inherent robustness of the KGs against such processes, and assessing the impact of their perturbations. Secondly, model the probability that reflects the likelihood of a rank change for an entity e, with a degree before perturbation k_e, given the magnitude Δk_e of a perturbation targeting it. This probability depends on several factors, including: the initial degree k_e of the entity, the magnitude and nature of the perturbation Δk_e, and finally, the global distribution of degrees $\{k_{e_1}, k_{e_2}, \ldots, k_{e_N}\}$ across all entities which, as we will see later, will play an essential role in determining the probability.

Challenges Key difficulties include the heterogeneous nature of rankings in knowledge graphs and the complex interplay between local (entity-specific) and global

(distribution-wide) effects of perturbations. A solution should involve an extraction of statistical insights taking into consideration the entire distribution, which contributes to determining the likelihood that an entity changes its rank after a perturbation.

5 Perturbation Scenarios

In this section, we introduce perturbation models designed to assess the impact of intentional attacks and unintentional errors on rankings. These models aim to represent various scenarios in which perturbations influence entity rankings, either by altering the overall structure of the distribution (global perturbation in Section 5.1) or by targeting individual entities (local perturbation in Section 5.2).

5.1 Global Perturbation: Altering the Entire Distribution

This *global perturbation* scenario represents cases where a ranking is subjected to perturbations scattered across many entities, which possibly disrupt its overall structure. This scenario fits unintentional errors, which could arise from multiple erroneous edits and potential side-effects of vandalism. It involves adding some facts to (or removing facts from) randomly picked entities, thereby reshaping the entire distribution and disturbing the overall ranking. This scenario aims to assess the vulnerabilities of distributions to small and large-scale disruptions. For instance, Figure 1 illustrates an example of this type of perturbation as the *Global perturbation* scenario. Starting from an example distribution, where entities are ranked by their number of incoming facts, 4 new facts (in red) are added randomly. This results in a change in the ranking distribution, for entities e_4 and e_3, whose order is reversed.

5.2 Local Perturbation: Altering a Single Entity

In this *local perturbation* scenario, the perturbation is designed to target a specific entity with the objective of altering its rank within the distribution. This is achieved by adding or removing a certain number of facts associated with the entity, with the sole aim of modifying its position in the ranking. Such targeted perturbations provide insight into the proportion of the vulnerable individual entities w.r.t the perturbation. This scenario fits cases of vandalism targeting specific entities, but can also model errors from a faulty script making multiple changes involving a single entity. For instance, Figure 1 illustrates an example of this type of perturbation as the *Local perturbation* scenario. Unlike with the global perturbation, this scenario specifically targets the entity e_3 to improve its rank relative to the entire distribution by focusing all the fact additions on this entity. As a result, e_3 moves to rank 2, replacing the entity e_2.

Section 7.1 provides a quantitative analysis of the global effects of these two different types of perturbations, using precise measures to assess the inherent stability of rankings in knowledge graphs under such conditions of diverse perturbation scenarios and structural manipulations.

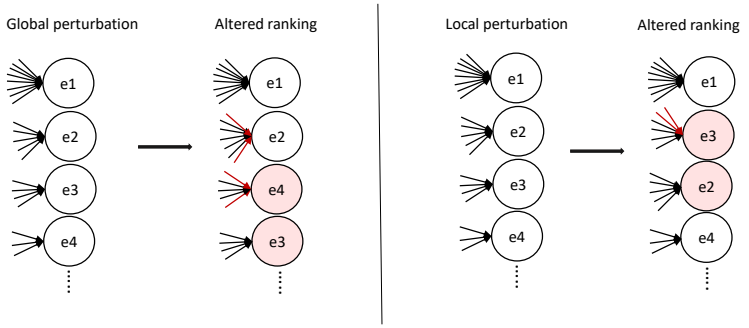


Fig. 1: An illustration of the global and local perturbation scenarios and their impact on entity rankings

6 Theoretical Model for Ranking Stability

The stability of rankings is linked to the underlying distribution of entity degrees within a knowledge graph. To quantify the impact of perturbations on ranks, it is essential to first identify the probability distribution model that governs the degree distribution of entities (see Section 6.1). This identification provides the foundation for modeling the likelihood of rank changes under various scenarios. Indeed, we use the PMF which plays a crucial role in understanding the structural properties of knowledge graphs. It does not only characterize the degree distribution, but it also provides a probabilistic basis for evaluating the effects of perturbations, such as the addition or removal of facts, on rankings. Building upon this probabilistic foundation and by leveraging the PMF, we propose in Section 6.2 a theoretical model for evaluating the robustness of rankings under perturbations.

6.1 Knowledge RELationship Model (KRELM)

The structural properties and dynamics of knowledge graphs relationships have been widely studied to better understand their evolution. A recent work by Abdallah et al. [2] introduces a novel model named KRELM (Knowledge RELationship Model), which provides insights into the distribution patterns of relationships in knowledge graphs. This model is especially relevant to our work as it offers a theoretical foundation to analyze and simulate the behavior of facts accumulation within KGs.

The KRELM model leverages a bipartite graph representation for relationships, where subjects and objects are treated as distinct vertex sets connected by edges representing facts. By focusing on the decentralized and crowdsourced nature of KG construction, KRELM captures the dual processes of continuous growth (addition of new entities) and asymmetric attachment (different behaviors for subjects and objects). The authors demonstrate that this model suc-

cessfully reproduces key structural patterns observed in real-world KGs, such as Wikidata, DBpedia, and YAGO.

Objects KRELM predicts that the PMF of the degree distribution for objects follows a power law with exponent $\gamma = 1 + \frac{1}{1-p_o}$, as expressed by:

$$P(k_o) = \frac{1}{1-p_o} \times k_o^{-\left(1+\frac{1}{1-p_o}\right)}, \tag{1}$$

where p_o denotes the probability of introducing a new object entity. The power-law behavior reflects the preferential attachment mechanism. Consider for example the *place of birth* relation which has a person as subject and a place as object. When a new fact with the relationship *place of birth* is added, it is very likely that the object is a large city where many people were born. This results in popular objects accumulating a disproportionate number of facts over time. This manifests in a degree distribution across objects where a lot of small cities have very few facts (few people were born there) and relatively few large cities concentrate much larger amounts of facts.

Subjects In contrast, in KRELM the PMF of the degree distribution for subjects follows an exponential law parameterized by $\beta = \frac{1}{1-p_s} - 1$, described as:

$$P(k_s) = \frac{p_s}{1-p_s} \times \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right), \tag{2}$$

where p_s represents the probability of adding a new subject entity. This result aligns with the intuition that subjects exhibit a relatively uniform attachment mechanism and facts with a given relation are spread evenly across subjects. Consider for example, the *cast member* relation which has a movie as subject and an actor as object. While the number of films associated with an actor may vary a lot (the preferential attachment of objects), most movies have a similar number of actors, and no movie concentrates all the actors.

These theoretical results have been validated through extensive experiments across multiple KGs including Wikidata. The findings highlight that the asymmetric attachment mechanisms and growth dynamics of KRELM closely replicate the observed real-world distributions of facts for both subjects and objects.

Based on these degree distributions derived from KRELM, we derive a probabilistic model to assess the likelihood that an entity changes its rank after a given perturbation.

6.2 Rank Change Likelihoods

In this section, we focus on the local perturbation scenario with a targeted entity. Specifically, we aim to calculate the probability that the rank of e will be altered due to the addition or removal of Δk_e facts related to that entity. The results presented here are obtained by leveraging the degree distributions given by KRELM, which characterize objects with a power law, and subjects with an exponential law.

Building upon the initial formulation of our second goal in Section 4, we now reformulate the task to focus on how to determine the probability $\Pr(\Delta\sigma(e) \neq 0 \mid \Delta k_e, k_e)$. Given an entity e with a degree k_e under a relationship p, for a perturbation which adds (resp. removes) Δk_e facts to e, the probability that the rank will be changed is the probability that there exists another entity e' with a degree less than Δk_e over (resp. under) k_e. Therefore, the objective is to compute the probability: $\Pr(\exists e', k_e \leq k_{e'} \leq k_e + \Delta k_e)$ (resp. $\Pr(\exists e', k_e - \Delta k_e \leq k_{e'} \leq k_e)$).

The probability that another entity e' has a degree $k_{e'}$ between k_e and $k_e + \Delta k_e$ (corresponding to the probability of a rank gain for entity e from a perturbation adding Δk_e facts to e) is given by:

$$\Pr(\exists e', k_e \leq k_{e'} \leq k_e + \Delta k_e) = \int_{k_e}^{k_e + \Delta k_e} P(k)\, dk.$$

For a perturbation removing Δk_e facts from e, the probability of a rank loss for e, which is the probability that another entity e' has a degree $k_{e'}$ between $k_e - \Delta k_e$ and k_e is given by:

$$\Pr(\exists e', k_e - \Delta k_e \leq k_{e'} \leq k_e) = \int_{k_e - \Delta k_e}^{k_e} P(k)\, dk.$$

By injecting Equation 1 in these integrals, we obtain the following theorem:

Theorem 1 (Rank change likelihood for objects). *The rank change likelihood of objects with degree k_o for Δk_o added facts is given by:*

$$\Pr(\exists o', k_o \leq k_{o'} \leq k_o + \Delta k_o) = k_o^{1-\gamma} - (k_o + \Delta k_o)^{1-\gamma}$$

Likewise, for Δk_o removed facts, the rank change likelihood of objects with degree $k_o \geq \Delta k_o$ is given by:

$$\Pr(\exists o', k_o - \Delta k_o \leq k_{o'} \leq k_o) = (k_o - \Delta k_o)^{1-\gamma} - k_o^{1-\gamma}$$

Considering a perturbation that may either add Δk_o facts or remove Δk_o facts, the rank change likelihood of objects with degree $k_o \geq \Delta k_o$ is given by:

$$\Pr(\Delta\sigma(o) \neq 0 \mid \Delta k_o, k_o) = (k_o - \Delta k_o)^{1-\gamma} - (k_o + \Delta k_o)^{1-\gamma}$$

Now, by injecting Equation 2 in the above integrals, it is possible to get a similar theorem for the out-degree of subjects:

Theorem 2 (Rank change likelihood for subjects). *The rank change likelihood of subjects with degree k_s for Δk_s added facts is given by:*

$$\Pr(\exists s', k_s \leq k_{s'} \leq k_s + \Delta k_s) = \left(1 - \mathrm{e}^{-\beta \Delta k_s}\right) \mathrm{e}^{-\beta(k_s - 1)}$$

Likewise, for Δk_s removed facts, the rank change likelihood of subjects with degree $k_s \geq \Delta k_s$ is given by:

$$\Pr(\exists s', k_s - \Delta k_s \leq k_{s'} \leq k_s) = \left(\mathrm{e}^{\beta \Delta k_s} - 1\right) \mathrm{e}^{-\beta(k_s - 1)}$$

Considering a perturbation that may either add Δk_s facts or remove Δk_s facts, the rank change likelihood of subjects with degree $k_s \geq \Delta k_s$ is given by:

$$\Pr(\Delta\sigma(s) \neq 0 \mid \Delta k_s, k_s) = \left(\mathrm{e}^{\beta\Delta k_s} - \mathrm{e}^{-\beta\Delta k_s}\right) \mathrm{e}^{-\beta(k_s-1)}$$

For an entity with k_e facts, Theorems 1 and 2 indicate the likelihood of changing its rank by allowing addition only, removal only, or both addition and removal depending on the entities. A first corollary of this theorem is that the rank change likelihood (by adding or removing facts) decreases with an increase of the degree k_e of the entity. This shows the resilience of highly connected objects to changes in their degree. It also follows from the theorem that, for a fixed perturbation size Δk_e, the likelihood of a rank change is higher when facts are removed than when they are added, regardless of an entity's degree. In short, for an attacker, vandalism is more effective at lowering an entity's rank than at boosting it. Unsurprisingly, as the Δk_e increases, the impact of the perturbation is more likely to change the ranking. Obviously, in the case of vandalism, the greater the perturbation, the greater the chance of it being detected (as it exceeds the crowdsourcing system's alert threshold). Interestingly, with our model, it is possible to define an alert threshold (a maximum authorized perturbation size) specific to the parameters of the relationship to be monitored.

Figure 2 illustrates Theorems 1 and 2 on the cast member relationship in Wikidata (`P161`) having $\beta = 0.181$ and $\gamma = 2.179$ as parameters. More precisely, we plot the theoretical perturbation likelihoods with the degree k ranging from 3 to 1,000 and a perturbation size $\Delta k_e = 2$. The solid and dashed lines correspond to objects and subjects respectively, with gain (addition of facts), loss (removal of facts) and total (addition or removal according to entities). Firstly, we observe that all three curves decline as the degree increases, in accordance with the theorems. Secondly, the perturbation likelihood is clearly higher for fact removal (blue and orange curves) than for fact addition (red and green curves), especially for low degrees. The difference fades for large degrees, where attacks become harder, and the two probabilities converge. Changing a ranking by performing both additions and removals according to the considered entity remains the easiest strategy, as indicated by the purple and brown curves. Finally, for small degree, the probability of a rank change is higher for subject rankings than for object rankings (here, the subjects and the objects are involved in the same relationship, but this observation can be generalized to all relationships). Once the degree becomes large, the subject probability drops toward zero, while the object probability continues to decrease normally. This crossover occurs because only a few subject entities have very high degrees, whereas most have low degrees.

Theorems 1 and 2 are fundamental. They allow us not only to confirm intuitions, but also to quantify them precisely. In particular, it is possible to identify degrees and entities in the knowledge graph that are most vulnerable to perturbations given a relationship. It also helps in identifying thresholds for perturbation sizes beyond which rank stability is significantly compromised. These results for the local perturbation scenario are also interesting for the global perturbation scenario. Indeed, when modifications are uniformly distributed over

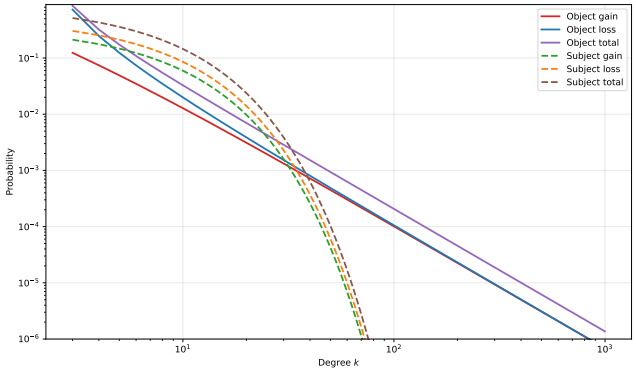


Fig. 2: Theoretical rank change likelihoods for the cast member relationship in Wikidata

all entities, this is like considering local perturbations with a low number of fact modifications. In this case, our model predicts better ranking stability, especially for entities with a high degree. Next section shows that these theoretical rank chance likelihoods for the local perturbation and its consequences for the global perturbation are accurately verified in practice, demonstrating the scope of our modeling.

7 Experimental Study

In this section, we aim to quantitatively evaluate the robustness of rankings gathered from several crowdsourced KGs against errors and vandalism. Section 7.1 provides an empirical analysis of ranking stability under the global and local perturbation scenario. Then, Section 7.2 focuses on assessing our theoretical results of Section 6 that compute the likelihood of a rank change given a perturbation.

Note that source code and experimental results are available in the Git repository: `https://scm.univ-tours.fr/habdallah/rankingsperturbations`

7.1 Empirical Analysis of Ranking Stability

In Section 5, we introduced two distinct perturbation scenarios. Here, we conduct an empirical analysis to evaluate the impact of these scenarios on ranking induced by relationships in Wikidata.

Protocol To apply the proposed scenarios, we utilized the top 500 ranking indicators identified in [1] to get their distributions from Wikidata. For each relationship, we selected the top 5,000 ranked entities and applied the two scenarios to measure the impact of perturbations. Additionally, each distribution was divided into two segments: the first from the top to the middle of the ranking (*FirstDistribution*), and the second from the middle to the end (*SecondDistribution*).

This segmentation enabled us to measure the impact on different parts of the distribution while also evaluating the overall effect on the global distribution.

To apply the global perturbation scenario, we randomly added Δk new facts targeting entities within the distribution. The entities receiving the new facts were selected randomly, and Δk was varied from 1 to 50,000, increasing by 100 at each step. We then measured the Kendall's tau divergence between the resulting distribution and the initial one to quantify the perturbation's impact. The experiment is repeated 5 times, then we take the average value of Kendall's tau for each value of Δk. Kendall's tau, denoted by τ (also known as Kendall's rank correlation coefficient) quantifies the resemblance between two rankings by assessing the proportion of pairwise agreements and disagreements between the ranks. It is a non-parametric measure, meaning it does not assume any specific distribution of the data, and it is suitable for ordinal data. It can be computed using the following formula: $\tau = \frac{\mathrm{C-D}}{\mathrm{C+D}}$ where C is the number of concordant pairs, and D is the number of discordant pairs. Kendall's tau ranges from -1 (i.e., perfect disagreement) to 1 (i.e., perfect agreement).

For the local perturbation scenario, we considered both fact addition and removal for evaluating Theorems 1 and 2. More precisely, we observed for each entity of degree k whether an addition of Δk facts or a removal of Δk facts modified its ranking. This enabled us to estimate for each degree k the proportion of entities whose ranks changed as a result (either when facts are added or removed). Here, Δk was varied from 1 to 100, increasing by 1 at each step. This allowed us to assess the sensitivity of individual entities to these perturbations and their influence on the overall ranking stability. This protocol ensures a comprehensive evaluation of the two perturbation models across different segments of the rankings and under varying levels of perturbation intensity.

Results Figure 3 presents the variation of Kendall's tau after applying the experiment outlined in the global perturbation scenario on the ranking indicator that ranks *film actors* according to the number of films they were *cast members* in. The x-axis represents the number of added facts, while the y-axis shows the corresponding Kendall's tau values. The figure illustrates a sharp decrease in Kendall's tau when the number of added facts increases from 0 to 10,000. Beyond this point, the curves exhibit a gradual stabilization, although Kendall's tau continues to decrease strongly. As expected, the first half of the distribution demonstrates greater stability and robustness compared to the second half. Additionally, the global distribution is more resilient than either of the two halves individually. This behavior indicates that changing the ranking of entities in the first half is more challenging due to the significant inequality in the number of facts associated with these entities. Similarly, changing the ranking of the global distribution is more difficult because it encompasses the inherent stability of both halves combined. These results highlight the varying degrees of robustness across different segments of the rankings.

Considering the local perturbation scenario, Figure 4 illustrates the proportion of entities that change their rank after adding or removing a specified number of facts associated with them in the original distribution (using the same

ranking indicator as the first scenario). As expected, the proportion consistently increases as the number of modified facts grows. For the second half of the distribution, all entities experience a rank change as soon as a single fact is modified. In contrast, for the first half, the proportion remains below 1 when fewer than 19 facts are modified. This disparity highlights the greater stability of entities in the first half, which is likely due to the larger number of facts associated with these entities, making them more resistant to perturbations.

The results detailed here for the cast member relationship are similar for the other relationships in Wikidata. To summarize this general behavior, Table 2 presents the results of computing the average Kendall's tau and proportions across various threshold values for the top 500 ranking indicators. For the global perturbation scenario, the results highlight the greater stability and robustness of the first half of the distribution compared to the second half. The global distribution on average is slightly less stable than the first half of the ranking. These results underscore the inherent robustness of the rankings in Wikidata for top-ranked entities (the first half), providing strong evidence of their resilience under errors. For the local perturbation, we observe that most relationships are highly sensitive to perturbation. A single perturbation has a high impact, with a probability greater than 97% (on average) of changing an entity's rank.

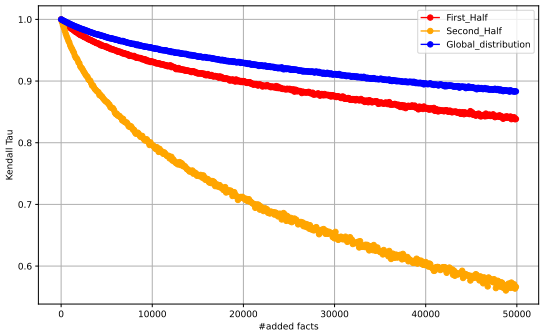


Fig. 3: Kendall's tau with Δk facts added to the entire distribution (global perturbation)

7.2 Evaluation of Theoretical Likelihoods

In this section, we evaluate the effectiveness of the theoretical rank change likelihoods computed in Section 6.2 that quantifies the probability of rank changes in a distribution for the local perturbations. This metric directly translates to the likelihood that an attack succeeds in altering the rank of an entity, particularly in cases of vandalism, given the degree of the targeted entity.

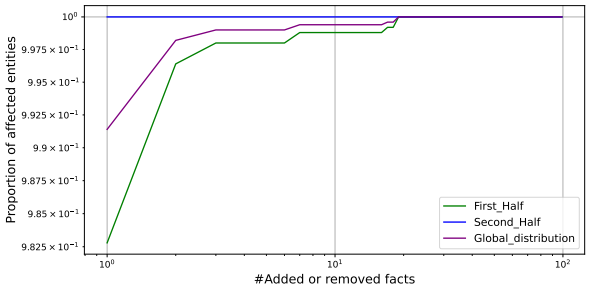


Fig. 4: Proportion of entities who change rank with Δk facts changed for a single entity (local perturbation)

	Global perturbation Avg. Kendall's tau			Local perturbation Avg. proportion		
Thresholds	10,000	30,000	50,000	1	10	100
First Half	0.592	0.472	0.420	0.975	0.997	0.999
Second Half	0.183	0.116	0.092	0.999	1	1
Global Distribution	0.571	0.451	0.399	0.987	0.998	0.999

Table 2: Impact of Δk facts in global and local perturbation scenarios for the top 500 relationships in Wikidata

Protocol To evaluate the performance of our theoretical results, we employed the real degree distributions of relationships of three major KGs: the Wikidata 2022 snapshot, DBpedia, and YAGO. The evaluation follows these steps:

1. For each relationship, we computed the probability distributions for subjects and objects entities, induced by the probability mass function (PMF). Where each probability point, considered for a degree interval $[k - \Delta k, k + \Delta k]$, and iteratively shifted k by $\Delta k + 1$. This approach enables us to calculate $\Pr(\exists e', k - \Delta k \leq k_{e'} \leq k + \Delta k)$ for each degree k and construct a real-world probability distribution.
2. Using the theoretical formulas derived in Section 6.2, we similarly computed a theoretical probability distribution for comparison.

To assess the efficiency and performance of the probabilistic model, we utilized the following metrics:

Jensen-Shannon Divergence (D_{JS}): To compare the theoretically generated probability distributions with the real ones for the three knowledge graphs, we employed the Jensen-Shannon divergence (D_{JS}), a widely used metric that quantifies the similarity between two probability distributions. The D_{JS} is defined as:

$$D_{JS}(P||Q) = \frac{1}{2}\left(D_{KL}(P||M) + D_{KL}(Q||M)\right),$$

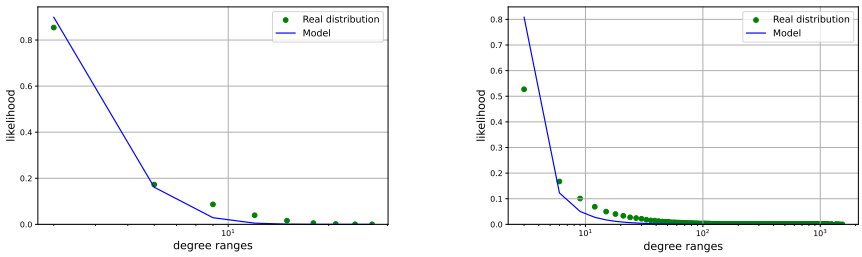
where P and Q are the two distributions being compared, and $M = \frac{1}{2}(P + Q)$ is the average of the distributions, and the $D_{KL}(P||Q)$ represents the Kullback-Leibler Divergence between distributions P and Q: $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$. This symmetric metric provides a measure of how closely the theoretical distribution approximates the empirical distribution. Its range is [-1,1] with smaller values indicating higher similarity.

***Coverage* Measure:** In addition to Jensen-Shannon divergence, we used the *coverage* measure to further evaluate the model's performance. Coverage is defined as the proportion of properties for which the model successfully reproduces reality. It is calculated as:

$$Coverage = \frac{|\{p \in \mathcal{K} : D_{JS}(T_{\tilde{p}} \| R_p) \leq 0.2\}|}{|\{p \in \mathcal{K}\}|}$$

where R_p (resp. $T_{\tilde{p}}$) represent the real (resp. theoretically generated) probability distribution. This measure allows us to assess the model's effectiveness in replicating real-world behavior across a range of properties.

To determine success for a distribution replication, we set a threshold of 0.2 for D_{JS}, a value that analytically and visually indicates strong model performance. In our experiments, we evaluated the model on 3,228, 78, and 971 relationships for DBpedia, YAGO, and Wikidata, respectively. As well, we set $k = 3$ as the starting degree and incremented it by 3 in each step, with $C = 2$.



(a) Subjects (b) Objects

Fig. 5: Comparison of likelihood of a rank change for the `member of sports team` relationship in Wikidata between the model and the real distribution

Results Figure 5 illustrates the real and theoretical probability distributions for the `member of sport team` relationship (P54) in Wikidata. The y-axis represents the rank change likelihood, while the x-axis denotes the degree k. The real-world estimated probabilities are visualized as green points, and the model's theoretical probabilities are depicted as a continuous blue line. The figure demonstrates that the model closely approximates the real distribution for both subjects and objects. Specifically, the theoretical likelihoods follow an exponential law for

subjects and a power law for objects, aligning with the expected behavior of the degree distributions. This close fit is further corroborated by the Jensen-Shannon divergence which is 0.019 for subjects and 0.090 for objects in average, highlighting the strong similarity between the model and the real data. Table 3 provides a comprehensive comparison of the real degree distributions and the theoretically generated ones for the three knowledge graphs (Wikidata, DBpedia, and YAGO). The results indicate that the model achieves an extensive coverage of the successfully approximated probability distributions across KGs. Coverage values consistently exceed 0.94 for both subjects and objects, with an average of 0.994 for subjects and 0.960 for objects across the three KGs. Similarly, the average D_{JS} values across the three KGs are 0.008 for subjects and 0.038 for objects, underscoring the model's ability to accurately replicate the real distributions. These results demonstrate the model's robustness in predicting the likelihood of rank changes resulting from local perturbations. The consistent performance across multiple KGs further validates the generalizability and effectiveness of the proposed probabilistic model in capturing the dynamics of rank stability in crowdsourced knowledge graphs.

	Coverage		Average D_{JS}	
KG	subjects	objects	subjects	objects
DBpedia	0.999	0.989	0.007	0.022
Wikidata	0.996	0.941	0.009	0.044
YAGO4	0.987	0.949	0.007	0.048
Average	**0.994**	**0.960**	**0.008**	**0.038**

Table 3: Coverage and average JSD for the relationships of three KGs

Parameter analysis Despite the strong results already presented, the probabilistic model occasionally fails to approximate the real probability distributions for certain relationships. This prompted a deeper investigation into the underlying causes of these failures. Figure 6 provides an analysis of these failures. The x-axis represents the D_{JS} values obtained using the presented probabilistic model, while the y-axis shows the D_{JS} values derived from the KRELM model for the same relationships. The points represent subjects (in red) and objects (in blue), with regression lines fitted to the distributions for each group. Interestingly, we observe a strong linear correlation between the performance of the probabilistic model and the accuracy of KRELM, with the slopes and R values of 0.4535 and 0.8232 for subjects, and 0.4544 and 0.8590 for objects, respectively. Figure 6 also reveals that when the probabilistic model struggles to approximate the real distribution (right part of the figure), it is often due to inaccuracies in the underlying KRELM model (upper part of the figure). This shows the soundness of the method for building a ranking stability model from a degree distribution model. Proposed improvements to KRELM to overcome its limits would therefore also improve the performance of the probabilistic model.

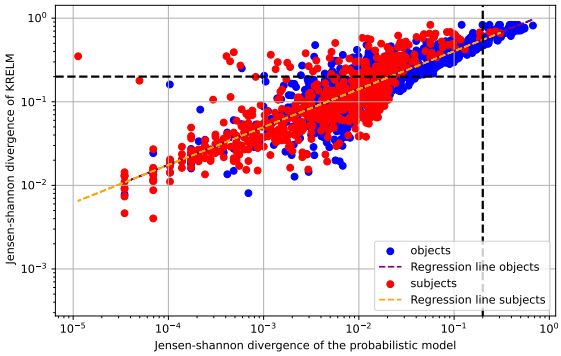


Fig. 6: Comparison of the Jensen-Shannon divergence of the probabilistic model and KRELM

8 Conclusion

In this paper, we investigated the stability of rankings in KGs, under the influence of vandalism and error perturbations. Our work shifts the focus to the structural aspects of knowledge graphs, particularly the impact of perturbations on relationships and their induced rankings. We introduced a formalization of ranking stability issues and proposed a probabilistic model to assess the likelihood of rank changes due to local perturbations. Our probabilistic predictive model, grounded on complex network analysis, provides a framework for evaluating ranking vulnerabilities. The experimental study confirmed the effectiveness of the proposed model, and that the rankings exhibit varying levels of resilience depending on the degree distributions of entities and the nature of the perturbations. We found that removing facts is generally a stronger attack than adding facts. A combined strategy of additions and removals is the most disruptive, and subjects and objects react differently to perturbations depending on their degree. Specifically, we showed that high-degree entities tend to be more resistant to perturbations, while low-degree entities are more susceptible to rank changes. In practice, entities in the lower half of rankings are far more fragile, while the top-ranked entities show greater stability. Our findings highlight the need for robust mechanisms to safeguard knowledge graphs, prioritizing defenses against fact removals with the possibility of setting relationship-specific alert thresholds.

Future research directions include improving the accuracy of the model proposed in [2], and enabling the automatic computation of the proportion and Kendall's tau given basic statistics about the ranking. Moreover, the current model is for a perturbation on one entity. We therefore want to further extend it to also handle more specifically perturbations such as the one illustrated on the left in Figure 1, the global scenario, where no modification has been made to the entity e_3 even though it has moved down the ranking. By addressing the

dual threats of vandalism and errors, this work contributes to the broader understanding of ranking stability in crowdsourced knowledge graphs, provides an initial probabilistic vulnerability prediction model, and paves the way for more resilient knowledge graph management strategies.

References

1. Abdallah, H., Markhoff, B., Soulet, A.: Ranking indicator discovery from very large knowledge graphs. Proceedings of the VLDB Endowment **18**(4), 1183–1195 (2024)
2. Abdallah, H., Markhoff, B., Soulet, A.: A complex network model for knowledge graphs' relationships. Semantic Web **16**(5), 1–20 (2025)
3. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. nature **406**(6794), 378–382 (2000)
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: international semantic web conference. pp. 722–735. Springer (2007)
5. Dürst, M., Suignard, M.: Internationalized resource identifiers (iris). Tech. rep. (2005)
6. Gale, A., Marian, A.: Explaining monotonic ranking functions. Proceedings of the VLDB Endowment **14**(4), 640–652 (2020)
7. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: Proceedings of the 16th international conference on World Wide Web. pp. 211–220 (2007)
8. Heindorf, S., Potthast, M., Bast, H., Buchhold, B., Haussmann, E.: Wsdm cup 2017: Vandalism detection and triple scoring. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 827–828 (2017)
9. Heindorf, S., Potthast, M., Stein, B., Engels, G.: Towards vandalism detection in knowledge bases: Corpus construction and analysis. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 831–834 (2015)
10. Heindorf, S., Potthast, M., Stein, B., Engels, G.: Vandalism detection in wikidata. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 327–336 (2016)
11. Heindorf, S., Scholten, Y., Engels, G., Potthast, M.: Debiasing vandalism detection models at wikidata. In: The World Wide Web Conference. pp. 670–680 (2019)
12. Sarabadani, A., Halfaker, A., Taraborelli, D.: Building automated vandalism detection tools for wikidata. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 1647–1654 (2017)
13. Shrinivasan, S., Razniewski, S.: How stable is knowledge base knowledge? arXiv preprint arXiv:2211.00989 (2022)
14. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: Proceedings of the 16th international conference on World Wide Web. pp. 697–706 (2007)
15. Trokhymovych, M., Saez-Trumper, D.: Wikidata vandalism detection with graph-linguistic fusion. In: Proceedings of the 11th Wiki Workshop (2024)
16. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)