Semantic Web 1 (0) 1–4
IOS Press


# A Complex Network Model for Knowledge Graphs' Relationships

Hassan Abdallah [a], Béatrice Markhoff [b] and Arnaud Soulet [a]
[a] *LIFAT, University of Tours, 3 place Jean Jaurès, 41000 Blois, France*
*E-mails: hassan.abdallah@univ-tours.fr, arnaud.soulet@univ-tours.fr*
[b] *CITERES, University of Tours, 3 place Jean Jaurès, 41000 Blois, France*
*E-mail: beatrice.markhoff@univ-tours.fr*

**Editors:** First Editor, University or Company name, Country; Second Editor, University or Company name, Country
**Solicited reviews:** First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country
**Open reviews:** First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

**Abstract.** When dealing with the structure, content, and quality of Knowledge Graphs (KGs), most analyses focus on entities, overlooking the significance of relationships and their evolution. In this paper, we introduce KRELM, a novel and efficient graph model that mimics the behavior of facts accumulation in crowdsourced KGs and accurately simulates the evolution of their structure. By modeling the decentralized process of crowdsourcing, KRELM reproduces key distribution patterns found in relationships, demonstrating that the facts in a KG can be generated incrementally, either by adding new entities or by further describing existing ones. Our theoretical analysis of KRELM reveals that the distribution of facts for relationships follows an exponential law for subjects and a power law for objects, enabling a deeper understanding of knowledge graph dynamics. Experimental validation on major KGs shows that KRELM successfully captures a large part of the structure of real-world relationships, and a longitudinal study of Wikidata confirms its effectiveness in predicting relationship evolution. This work opens new avenues for analyzing and benchmarking KGs.

Keywords: knowledge graph, complex network model, growth, preferential attachment, crowdsourcing, Wikidata

## 1. Introduction

Crowdsourcing techniques are essential to combine the collective intelligence of contributors with various expertise and opinions [50]. Some knowledge graphs (KGs) are built directly (e.g., Wikidata [56]) or indirectly (e.g., DBnary [51], DBpedia [5], or YAGO4 [44]), using crowdsourcing resulting in KGs that reconcile high quality and large size [37, 42, 60]. Social editing dynamics for semantic web received a lot of attention. Simperl and Luczak-Rösch [53] describe how community-driven tools such as wikis have led to a consensus-building process. Furthermore, Piscopo and Simperl [46] categorize editor tasks and roles to examine the links between task categories, user roles, and KG quality. To understand its social editing dynamics, Piao and Huang [45] propose a method to predict whether a user will remain active on the platform. All these works are focused on social aspects, not on data (even if it has also been noticed that the editor community may be unbalanced, introducing biases into the produced data, leading to cultural or social biases [23]). Obviously, schemas guide (or constrain) the editors for inserting new entities and new facts in each relationship (even if the schema itself is sometimes crowdsourced [57]). Moreover, schema

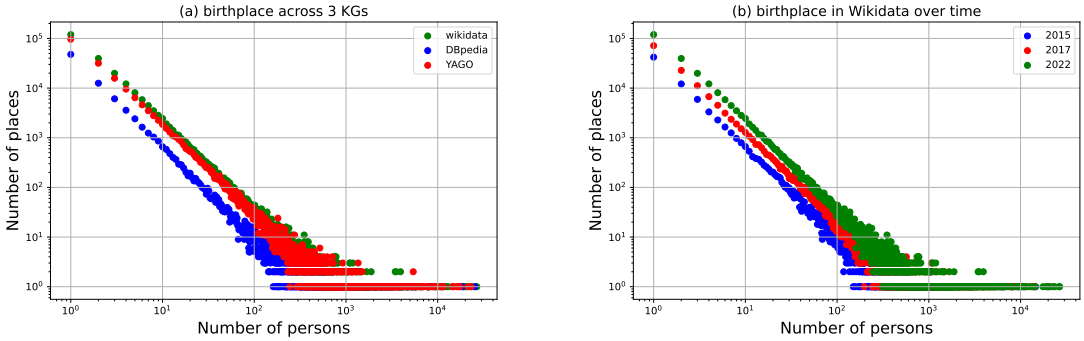


Fig. 1. Place of birth relationship.

formalization and modeling have already been studied in depth in knowledge representation, with various works on Description Logics and Ontologies [55]. However this line of research does not explain how factual knowledge accumulates to describe entities.

To better grasp the content of KGs, many works aim at extracting statistical information [13], summaries [18, 28], schemas [34], constraints [48], and so on. For instance, Figure 1 (a) shows the distribution for `place of birth` (denoted by `wdt:P19` in Wikidata) present in DBpedia [5], YAGO4 [44] and Wikidata [56]. More precisely it represents the number of places as a function of the number of births, for these three graphs. It shows that in all three resources, there are on the order of $10^5$ places in which only one person is declared to have been born (top left), while there is a concentration of birth declarations in a few places (bottom right, between 100,000 and several tens of millions for very few cities). These statistics show that the three distributions appear to follow exactly the same structure, resembling a power law. We also observe that this structure remains the same in Wikidata over the years in Figure 1 (b), again suggesting a form of stability.

More generally, studying statistics over time with numerous metrics is presented in [47] as an important issue for better understanding the evolution of knowledge graphs. In this paper, we propose to go beyond the accumulation of statistical observations to gain deeper insights of what they can provide. We want to reproduce the mechanisms underlying the construction of relationships in crowdsourced knowledge graphs to reproduce the distribution of facts. Indeed, understanding the distribution of facts about entities enables us to better understand the allocation of knowledge (what is the proportion of poorly described entities?), its stability (will the most popular entities remain so?) and its evolution (how will the relationships evolve with the future addition of facts?). Knowledge of these fundamental mechanisms would be invaluable for optimizations, making benchmarks more realistic, supporting new data mining and analyses aiding the development of applications on top of KGs, and so on. In summary, we aim to answer these two research questions:

***RQ1****: Do diverse and distributed contributions tend to produce a stable structure or an ever-changing chaotic structure?*
***RQ2****: How to model the process of knowledge accumulation itself, for explaining the emergence of the structure we observe in the distribution of facts?*

To address these two research questions, this paper explores a novel technique that has not yet been applied to knowledge graphs (KGs), namely the modeling of complex network structures and dynamics [8]. This approach has gained momentum in fields as diverse as computer networks [8] and biology [10]. In particular, it has contributed a significant body of knowledge on the emergence of structures and their dynamics in Web artifacts: HTML page networks and social networks [8], or folksonomies resulting of distributed collaborative annotation systems [31]. Unfortunately, none of the stochastic models for complex networks can be applied to knowledge graphs, because they are a superposition of many bipartite graphs (for a definition of a bipartite graph see Section 3.1 and Fig. 2). Indeed, each relationship is represented by a bipartite graph, whose every edge connects a vertex from subjects and objects. To the best of our knowledge, this particular structure has never been dealt with in the literature. In

addition to this bipartite structure, the semantics behind each relationship has an impact on the concentration of links, which is a challenge to capture. We will see in Section 2 that such stochastic models are preferable to deep learning methods for generating graphs because the latter are black boxes (in addition to having other limits with respect to KG topology). Similarly, Section 2 presents heuristic methods for reproducing a target graph, but they require a lot of input parameters, including the degree distributions, whereas we would like our model to explain these distributions by construction.

Studying the structure of KGs using existing stochastic models from network science is challenging because KGs incorporate semantics, thus they contain relationships across a wide variety of domains which are structured with respect to a logic. To meet this challenge, our contributions are as follows:

- We present the first complex network model, named KRELM, for generating a bipartite graph representing a relationship between subjects and objects, which mimics editors' contributions, dealing with RQ2. This stochastic process relies on the continuous arrival of new entities and on an asymmetric attachment (i.e., uniform for subjects and preferential for objects).
- By analyzing this model, we prove that the degree distribution tends towards an exponential law for subjects and a power law for objects. This important theoretical result answers RQ1 by demonstrating the emergence of a stable structure in knowledge graphs.
- We show experimentally on four real-world KGs that our model succeeds in reproducing most relationships and outperforms baselines. A longitudinal study shows that our model is able to simulate the growth of Wikidata over the years.

The rest of this paper is structured as follows: In Section 2, we analyze work related to graph models. Section 3 introduces definitions used later. In Section 4, we present the asymmetric bipartite graph model for KGs (algorithm and theoretical analysis). Section 5 presents a synthesis of experimental results obtained with the model. Section 6 discusses several practical applications of our model, and Section 7 concludes this work with perspectives.

## 2. Related Work

This section reviews the main categories of existing approaches for generating graphs, and more specifically, knowledge graphs, highlighting both the similarities and differences with our approach, and clarifying what distinguishes our research from existing studies. Given the rise of deep learning in recent years, and the fact that the model we propose re-generates an artificial structure for knowledge graph (KG) relationships that mirrors real-world patterns, we reference in Section 2.1 various studies that use deep learning to model graph structures. In doing so, we emphasize the differences in objectives between those works and ours, addressing their limitations relative to our goals. Additionally, we examine in Section 2.2 heuristic methods for regenerating synthetic KGs, discussing their general limitations, particularly in the context of our specific objectives. More importantly, as the use of graph models for understanding knowledge-related data has already been widely studied (as shown for instance in survey [24]), but not for KGs, we introduce those modeling techniques. For example, in the seminal paper [8], Barabási and Albert devised a fundamental model (named Barabasi-Albert model) to better understand networks such as the Web, or citation networks, by proving that the distribution of degrees follows a power law. Similarly, Ferrer i Cancho and al. [26] model the organization of syntax as a scale-free network. This model explains the structural similarities in syntax between different languages. Also, Halpin and al. [31] studied collaborative tagging using a model with a tripartite graph (users/tags/documents). Although the final distribution was not formally defined, their experiments also concluded that a stable structure emerges among tags, despite the diversity of contributors. All these studies clearly show the importance of graph models in understanding the structure of crowdsourced knowledge. They are useful for predicting the evolution of knowledge structure and evaluating its stability. Thus, Section 2.3 provides a detailed examination of this type of models. Unfortunately, as we will see in the following subsections, none of these existing approaches are suitable for our objectives or adequately address the research questions we have posed. We provide in Table 1 a synthesis that highlights this fact.

| Methods | Scope | Parameters | Growth | Attachment |
|---|---|---|---|---|
| Deep Learning methods | | | | |
| global [15, 22, 52] | simplex | training set | no | learned |
| sequential [29, 35, 36, 59] | simplex | training set | **yes** | learned |
| Heuristic methods | | | | |
| schema [54] | **multiplex** | schema+stat.+dist. | no | constrained |
| schema + data [30] | **multiplex** | schema+stat. | no | constrained |
| schema + data [6, 25, 38] | **multiplex** | schema+stat.+dist. | no | constrained |
| Stochastic models for complex networks | | | | |
| [8] | simplex | **stat.** | **yes** | **pref.** |
| [16, 21, 27] | simplex | **stat.** | **yes** | **pref.** |
| [14] | duplex | **stat.** | **yes** | **pref.** |
| [20, 43] | simplex | **stat.** | no | **rec.** |
| **Our model** | **bipartite** | **stat.** | **yes** | **unif./pref.** |

Table 1
Comparison of our model with related methods.

### 2.1. Deep Learning methods

With the growing interest in deep learning, numerous approaches have been proposed for generating graphs [58] either by generating the whole graph at once (global approach) or by generating the links one by one (sequential approach). At first sight, it might seem interesting to apply such approaches, which have worked successfully in several fields (e.g., molecular graphs [22]), to knowledge graphs. First, *global approaches* aim at reproducing graphs in one shot by benefiting from various architectures like autoencoder [52] or generative adversarial networks [15, 22]. But, these methods do not reproduce the successive addition of facts that is the main mechanism of crowdsourced knowledge graphs. More interestingly, *sequential approaches* [29, 35, 36, 59] generate a graph by adding at each iteration nodes and links. Unfortunately, deep learning methods are well suited only to small simplex networks (where all links are of the same type) such as molecule networks [22, 29, 52]. In contrast, KGs are multiplex networks with millions of nodes and links of various types. Furthermore, all these approaches require a set of graphs as a training set, which is very restrictive for reproducing a given KG (necessarily unique). Finally, the learned model is a black box that does not enable us to understand and analyze the evolution of a graph. Conversely, using simple statistics, we will see that KRELM accurately explains the distribution of entities within a relationship, and its evolution.

### 2.2. Heuristic methods

Several heuristic methods have been proposed for generating synthetic KGs that can be used as benchmarks for evaluating existing query engines. Their strength lies in their ability to take into account the specific features of KGs, by considering several types of relationships (in the manner of multiplex networks) and integrating a schema part (or T-box). Thereby, schema parameterization has become increasingly complex, from simple statistics [30] to rules extracted from the KG to be reproduced [38]. However, the generation of the assertion part (or A-Box) has received much less attention. For instance, Theoharis and al. [54] even ignore the generation of these assertions, because their aim was to reproduce the schema only. Other heuristic methods [6, 25, 30, 38] focus on ensuring that the assertions generated are consistent with the schema. However, one of the earliest methods only uses unrealistic uniform draws [30]. More recent approaches allow to control the degree distributions to be reproduced by adding constraints as input parameters (e.g., uniform or truncated exponential distributions [38] or joint-distributions [25]). These heuristic methods focusing on benchmark building do not aim to generate a crowdsourcing-like process. In particular, they have the disadvantage of being static and they do not allow a graph to be gradually completed to simulate the continuous arrival of new knowledge. This explains why these methods take distributions as input parameters, instead of reproducing these features by construction. As a result, they do not really model the work of editors for understanding the structuring of KG relationships. For this reason, our work is complementary to these

approaches. By tackling only the reproduction of one relationship at a time, our model could be injected into these methods so that we no longer need to specify the expected distributions.

### 2.3. Stochastic models for complex networks

Introduced in [8], preferential attachment is one of the main mechanisms explaining the emergence of networks. It consists in associating each new link with an existing node, favoring those with larger connectivity. More precisely, the probability of adding a new link connecting an existing node $n$ will be proportional to its degree $k_n$ (its connectivity). This mechanism has been extended to directed graphs [16] and non-linear preferential attachments [21, 27]. Note that [20, 43] also devised a method for generating graphs with complex distributions by performing recursive operations on the adjacency matrix, but they do not consider a continuous growth. In all these existing approaches (dedicated to the Web, citation network, and so on), all nodes have the same nature, and all links also have the same nature. In contrast, a KG groups a wide variety of entities linked by different types of relationships, distinguished by their name, each with its own semantics. There is a model for multiplex networks [14] (also called multi-layer or multi-dimensional) where each layer could represent a relationship, but it is limited to only two layers (duplex), with no distinction between nodes. Rather than having a coarse-grain model to represent the whole graph at once, our proposal aims at a fine-grain model to represent one relationship at a time. We will see in the next section that this means generating bipartite graphs, for which there is no existing method.

To conclude, Table 1 summarizes the characteristics of each method, highlighting their strengths (in bold). Importantly, none of the methods allows us to understand the underlying mechanisms of the knowledge graphs we are targeting with research questions RQ1 and RQ2. Deep Learning and heuristic methods do not seek to model but to reproduce data, and stochastic models are not suited to the representation of a knowledge graph.

## 3. Preliminaries

### 3.1. Basic definitions and notations

*Knowledge graph [32]* A knowledge graph of a set of edges $\mathcal{R}$ (representing relationships) and a set of vertices $\mathcal{E}$ (representing entities) is a set of labeled edges $\mathcal{K} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ (representing facts). It is important to note that this work focuses on entity modeling while ignoring literals. We write the facts in the form $\langle s, r, o \rangle \in \mathcal{K}$, where $r$ is the relationship, $s$ is the subject and $o$ is the object. For instance, $\langle \texttt{Fann Wong}, \texttt{place of birth}, \texttt{Singapore} \rangle$ and $\langle \texttt{Shanghai Knights}, \texttt{cast member}, \texttt{Fann Wong} \rangle$ state respectively that Fann Wong is born in Singapore and she starred in the movie "Shanghai Knights". Given a relationship $r$, $\langle s, r^{-1}, o \rangle \in \mathcal{K}$ means that $\langle o, r, s \rangle \in \mathcal{K}$ where $r^{-1}$ is the inverse relationship of $r$. For instance, we have $\langle \texttt{Singapore}, \texttt{place of birth}^{-1}, \texttt{Fann Wong} \rangle$.

Given a relationship $r$, $\mathcal{K}_r$ is the set of facts in $\mathcal{K}$ having $r$ as relationship: $\mathcal{K}_r = \{\langle o, r', s \rangle \in \mathcal{K} \ : \ r' = r\}$. Thereafter, we worked mostly on a single relationship $r$ at a time (e.g., $\mathcal{K}_{\texttt{place of birth}}$ selects all the facts about birthplaces). Given this implicit relationship $r$, its number of facts in $\mathcal{K}_r$ is denoted by $n$, and its number of subjects (resp. objects) in $\mathcal{K}_r$ 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.

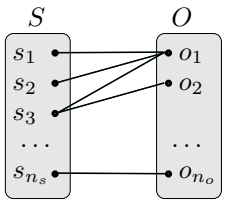


Fig. 2. Representing a relationship by a bipartite graph

*Bipartite graph [17]* It is possible to represent a relationship $r$ by a bipartite graph $(S, O, F)$ by considering two sets of vertices: one for subject entities $S \subseteq \mathcal{E}$ and one for object entities $O \subseteq \mathcal{E}$. Typically for the birthplace relationship, subjects and objects are people and cities respectively. Note that we do not require $S$ and $O$ to be disjoint, which is useful for relationships such as `part of` where a same place can be both subject or object of this relationship.

As shown in Figure 2, each edge $(s_i, o_j) \in F$ represents a fact $\langle s_i, r, o_j \rangle$. We denote the probability that an entity $e$ has $k$ facts at a time $t$ by $t$: $P(k_e(t) = k)$ (i.e., $k_s$ for outgoing degree and $k_o$ for incoming degree), and the probability mass function by $P(k)$. For knowledge graphs, the degree $k$ of an entity $e$ is a key indicator, as it provides information about the amount of knowledge about the entity, i.e. the number of facts involving it with other entities in the graph. In this way, the distribution of degrees $P(k)$ indicates the distribution of facts within the graph, highlighting entities that are poorly represented and those that concentrate the majority of knowledge. In the case of the birthplace relationship, the outgoing degree will be 1 for most subjects (i.e., people), while the incoming degree for objects (i.e., cities) will be very unbalanced as shown in Figure 1.

### *3.2. Methodology*

In the rest of the paper, we will follow the standard *Modeling-Analysis-Verification* methodology [40] to answer the two research questions identified in the introduction. In each step, original contributions are provided.

*Modeling* For research question RQ2, we propose a stochastic model, named KRELM, to mimic the behavior of editors so as to reproduce the construction of a knowledge graph relationship. Drawing inspiration from literature, we isolate two key mechanisms: the continuous addition of new entities (continuous growth) and the choice of different fact attachments for subjects and objects (asymmetric attachment) (see Section 4.1). From this model we derive an algorithm that generates data to reproduce real-world knowledge graph relationships (see Section 4.2).

*Analysis* We analyze in Section 4.3 the stochastic model KRELM to determine the probability mass function of both subject entities (i.e., outgoing degree $P(k_s)$) and object entities (i.e., incoming degree $P(k_o)$). These analytical results aim to demonstrate a stable distribution of facts within the entities involved in a relationship. This analysis therefore answers research question RQ1.

*Verification* We evaluate in Section 5 our stochastic model KRELM by comparing the distributions it generates with real-world distributions from several large knowledge graphs. For this purpose, the usefulness of the model's two ingredients is assessed by comparing the results with an equivalent model where they have been removed in turn (ablation study). Verifying that the generated distributions match the real-world distributions is essential for validating our theoretical model. In this way, the results of the theoretical analysis can be deemed realistic for application on top of knowledge graphs.

## 4. KRELM: Knowledge Relationship Model

### *4.1. Key ingredients of the model*

Our goal is to propose a model to generate a random bipartite graph simulating the crowdsourcing of a relationship by its editors. One might think that subjects and objects would behave identically because of their definition within a relationship which plays a dual role. As some ontologies (e.g. [12]) systematically define an inverse relationship for any relationship, subjects and objects should exhibit similar behaviors. However, this does not seem to be the case in practice, particularly in crowdsourced knowledge graphs, as measured by experiments in Section 5. Instead, relationships are oriented from a more recent subject entity to an older object entity; from the more specific to the more general, and from the more concrete to the more abstract. Naturally, when a person is added to a knowledge graph, he or she is linked to existing entities such as place of birth or gender. Similarly, all the actors in a film already exist when we add it. We will see that this intuition about the importance of adding new entities and their order of arrival leads to an asymmetrical model. More precisely, the addition of a new fact by an editor can either lead to the

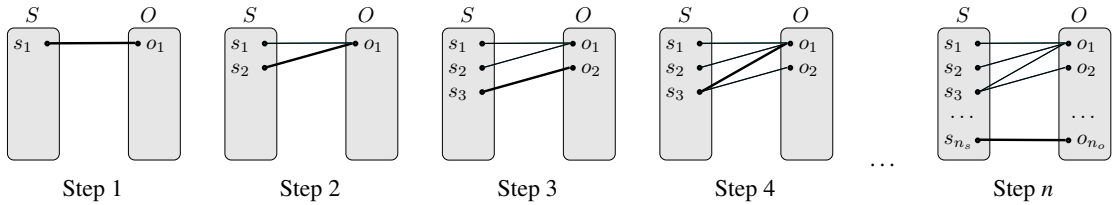


Fig. 3. Dynamic process of the bipartite graph representing the editing of the relationship

addition of new entities or just complete the information of existing ones. For example, Figure 3 shows the editing process that led to the complete bipartite graph of Figure 2 by adding each fact one by one. It is easy to see that Steps 1 and 3 add both new subjects and new objects, but Step 4 with the addition of $(s_3, o_1)$ does not add a new entity. To simulate editors, we introduce two key ingredients: Growth probability and Asymmetric attachment.

*Growth probability* With the addition of a new fact, editors regularly add new entities for both subjects and objects. We propose to model this proportion of new entities by one probability for subjects $p_s$ and one probability for objects $p_o$. Indeed, most relationships do not have the same behavior for subjects and objects. The choice of the probability $p_e$ (i.e. $p_s$ or $p_o$) is obviously crucial to reproduce a relationship by controlling the proportion of entities that will be created according to the number of facts. For instance, a probability of 1 associates each new fact with a new entity while a probability close to 0 concentrates all facts on the same few existing entities. It is therefore possible to exploit the number of expected facts and the number of expected entities to parameterize the model. If we wish to create $n$ facts with $n_e$ entities, we must choose the probability $p_e = n_e/n$. For instance, the application of this model to reproduce the birthplace relationship requires[1] $p_s \approx 1$ (whatever the KG) and $p_o = 0.081$ for Wikidata, $p_o = 0.101$ for DBpedia and $p_o = 0.085$ for YAGO4.

*Asymmetric attachment* When an editor does not create a new entity (i.e., with a probability $1 - p_e$), he has to choose one among those that already exist. We propose to use a different type of attachment for subjects and objects.

- *Uniform attachment for subjects*: in practice, attachment is often weak for subjects as growth probability is often close to one. Nevertheless, when there is an attachment, we choose one subject at random with a uniform probability. The intuition is that all subjects have the same chance of being described by a new fact. For the cast member relationship, for instance, movies have a similar number of actors, in other words, none concentrates all the actors. In Figure 3, the subjects' attachment is exploited only in Step 4, where the fact $(s_3, o_1)$ links back to the subject $s_3$ that had already been added in Step 3. Since the subjects' attachment is uniform, the probability of drawing $s_3$ was the same as drawing $s_1$ or $s_2$.
- *Linear preferential attachment for objects*: we choose one object at random with a probability proportional to the number of facts already describing the objects. The key idea is that an object that has already received a lot of facts will be more likely to receive others. For the birthplace relationship, it is clear that cities with many births are more likely to have a new birth. In the same way, popular actors who have already played in many films are more likely to be approached for a new film. In Figure 3, this means that in Step 3, the object $o_1$ is twice as likely to be selected than $o_2$ because its in-degree is twice as large. After this step, the preferential attachment of object $o_1$ will be further reinforced with a probability 3 times greater than that of object $o_2$. Unlike above, where the uniform attachment distributes facts evenly between subjects, the linear preferential attachment tends to concentrate them on the objects created first.

As mentioned in the state of the art, our model KRELM differs strongly from the heuristic methods [6, 25, 38, 54] since we do not directly inject the expected shape of the final degree distribution. Furthermore, KRELM simulates the dynamic evolution of the relationship by reproducing the successive additions of the different entities and facts instead of directly producing the final structure.

[1] As observed in the KGs we considered: we calculate $p_o$ by dividing the number of distinct objects (in this case, the cities) by the number of facts (in this case, the stated births). For example, for Wikidata, $p_o$ = 244,455 / 3,017,563 = 0.081.

**Algorithm 1** KRELM: Knowledge Relationship Model

**Require:** A number of subjects $n_s$, a number of objects $n_o$, a number of facts $n$
**Ensure:** A bipartite graph $(S, O, F)$
1: $S := \emptyset$; $O := \emptyset$; $F := \emptyset$
2: **while** $|F| < n$ **do**
3: $\quad F := F \cup (\textsc{Draw}(S, n_s/n, \mathit{unif}), \textsc{Draw}(O, n_o/n, \deg))$
4: **return** the bipartite graph $(S, O, F)$

5: **function** $\textsc{Draw}(E, p_e, a_e)$
6: $\quad$ **if** $\mathit{unif}(0,1) \leqslant p_e \vee E = \emptyset$ **then**
$\quad\quad$ // Insert a new entity in $E$
7: $\quad\quad e := e_{|E|+1}$
8: $\quad\quad E := E \cup \{e\}$
9: $\quad$ **else**
$\quad\quad$ // Draw an entity $e \in E$ proportionally to $a_e(e)$
10: $\quad\quad e \sim a_e(E)$
11: $\quad$ **return** $e$

### *4.2. Stochastic algorithm for asymmetric bipartite graph generation*

Knowledge RELationship Model (KRELM in short, see Algorithm 1) generates a bipartite random graph containing exactly $n$ facts, $n_s$ subjects on average, and $n_o$ objects on average. The main program (lines 1-4) generates $n$ facts separately by inserting them into the set $F$. To do this, it relies on the function DRAW to independently select one subject entity from $S$ and one object entity from $O$ (line 3). This function also regularly adds new entities to these two sets initialized with the empty set (line 1). More precisely, the function DRAW returns either a new entity $e$ with probability $p_e$ (or if $E$ is empty) inserted in $E$ (line 7-8), or an existing entity $e$ in $E$ drawn randomly with a probability proportional to $a_e(e)$ (line 10)[2]. As explained previously, the draw is uniform for subjects and proportional to a degree for objects, with respectively *unif* and $\deg$ as attachment function $a_e$ (line 3). The random draw proportionally to the degree is implemented by uniformly drawing a fact $(s, o)$ from $F$ and returning $o$ (see [11] for more details).

Figure 4 shows the out-degree distribution of subjects and the in-degree distribution of objects resulting from our asymmetric model KRELM applied on the birthplace and cast member relationships of Wikidata. For instance, we use Algorithm 1 for reproducing the birthplace relationship with $n = 3,017,563$ facts, $n_s = 3,001,229$ subjects (i.e., number of persons) and $n_o = 244,455$ objects (i.e., number of places) as parameters. Interestingly, this distribution of our model KRELM (in magenta) is close to the real-world distribution (in green) with a low Jensen–Shannon Divergence (see Section 5.1 for a definition) between 0.006 and 0.143, showing the relevance of the bipartite graph model.

### *4.3. Theoretical analysis of KRELM*

Algorithm 1 reproduces the main characteristics of the targeted relationship specified with the input parameters. Obviously, the number of edges $|F|$ corresponds exactly to $n$. The next property underlines that the number of subjects and objects is also comparable to those provided as input parameters:

**Property 1.** *Under the bipartite graph model KRELM, the expected number of entities is $n_e$.*

Property 1 follows from the fact that after $n$ calls of the function DRAW, the set $E$ contains $n \times p_e$ entities on average (because $p_e = n_e/n$). We therefore obtain $n_s$ subjects in $S$ and $n_o$ objects in $O$ on average.

[2]Let $\Omega$ be a population and $f : \Omega \to [0; 1]$ be a measure, the notation $x \sim f(\Omega)$ means that the element $x$ is drawn randomly from $\Omega$ with a probability distribution $\pi(x) = f(x)/Z$ where $Z$ is a normalizing constant.

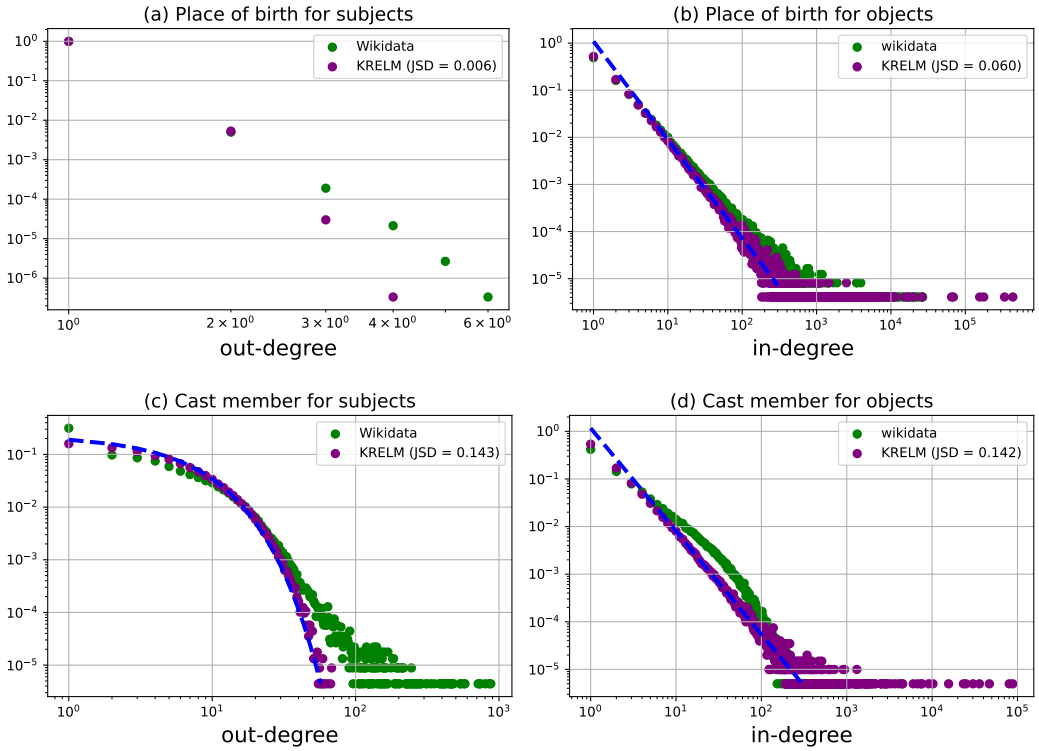


Fig. 4. Place of birth and Cast member relationships.

Beyond finding the main characteristics, the theoretical study of KRELM enables us to better characterize the original relationship, especially its degree distribution. As knowing the distribution of degrees is crucial because it means knowing the distribution of knowledge on entities.

Intuitively, the attachment mechanism adds new facts to entities already described by many facts. Formally, the degree $k_e(t)$ of the entity $e$ increases over the time $t$ with the following acquisition rate:

$$\frac{\partial k_e(t)}{\partial t} = (1 - p_e) \times \frac{a_e(t)}{t}$$

where $a_s(t) = 1/p_s$ (uniform attachment) and $a_o(t) = k_e(t)$ (linear preferential attachment). From this key equation on an entity, the following theorems infer the asymmetric distribution of facts for subjects and objects:

**Theorem 1** (Exponential law for subjects)**.** *Under the asymmetric bipartite graph model KRELM, the out-degree distribution of subjects follows an exponential law with the rate parameter $\beta$ such that:*

$$\beta = \frac{1}{1 - n_s/n} - 1$$

*Proof.* To prove this result, we benefit from the mean-field theory already used in network analysis [9]. The out-degree $k$ of a subject increases over the time $t$ with the following differential equation (when $p_s$ is not too close to 1):

$$\frac{\partial k(t)}{\partial t} = (1 - p_s) \times \frac{1}{p_s t} \tag{1}$$

where $p_s = n_s/n$ and $(1-p_s)$ is the attachment probability for subjects and $1/(p_s t)$ reflects the uniform attachment over the $p_s t$ existing subjects. We rewrite the differential equation and we calculate the corresponding integrals:

$$\int_1^{k_s} \partial k = \frac{1-p_s}{p_s} \int_{t_s}^{t} \frac{1}{t} \partial t \tag{2}$$

where $t_s$ is the time at which the subject $s$ is added and $k_s$ is its out-degree. We integrate both sides leading to the following equation: $k_s = \frac{1-p_s}{p_s} \ln \frac{t}{t_s} + 1$. It is possible to rewrite this equation for isolating $t_s(t)$:

$$t_s(t) = t \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right) \tag{3}$$

The probability that a subject $s$ has an out-degree smaller than $k$ (i.e., its cumulative degree distribution) $P[k_s(t) < k]$ can be rewritten as $P[t_s > t \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right)]$. Assuming that we add the facts at equal time intervals, leading to $P[t_s > t \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right)] = 1 - P[t_s \leqslant t \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right)] = 1 - \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right)$. The probability mass function $P(k)$ can then be obtained by derivation:

$$P(k_s) = \frac{\partial P[k_s(t) < k]}{\partial k} \tag{4}$$

$$= \frac{\partial \left[1 - \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right)\right]}{\partial k} \tag{5}$$

$$= \frac{p_s}{1-p_s} \times \exp\left(\frac{p_s}{1-p_s}(1-k_s)\right) \tag{6}$$

This gives an exponential law with the rate parameter $\beta = \frac{n_s/n}{1-n_s/n}$ proving Theorem 1.

□

**Theorem 2** (Power law for objects)**.** *Under the asymmetric bipartite graph model KRELM, the in-degree distribution of objects follows a power law with the exponent $\gamma$ such that:*

$$\gamma = 1 + \frac{1}{1-n_o/n}$$

*Proof.* We can apply the same proof scheme. The in-degree $k$ of an object increases over the time $t$ with the following differential equation (when $p_o$ is not too close to 1):

$$\frac{\partial k_o(t)}{\partial t} = (1-p_o) \times \frac{k_o(t)}{t} \tag{7}$$

where $p_o = n_o/n$ and $(1-p_o)$ is the attachment probability for objects and $k_o(t)/t$ reflects the preferential attachment over the objects. It leads to the following integrals (as proposed by Equation 2):

$$\int_1^{k_o} \partial k = (1-p_o) \int_{t_o}^{t} \frac{k_o(t)}{t} \partial t \tag{8}$$

We integrate both sides and rewrite this equation:

$$t_e(t) = t/k_e^{\frac{1}{1-p_o}} \tag{9}$$

The probability mass function $P(k)$ can be obtained with the same reasoning:

$$P(k) = \frac{1}{(1-p_o)} \times k^{-\left(1+\frac{1}{1-p_o}\right)} \tag{10}$$

This gives a power law with the exponent parameter $\gamma = 1 + \frac{1}{1-n_o/n}$ proving Theorem 2. □

These theorems are an important result since they qualify the structure of the relationships according to the bipartite graph model KRELM. Figures 4-b, 4-c, and 4-d show the blue dashed lines generated through the formulas of the above two theorems. These lines match very well with the distributions generated using KRELM, which illustrates the validity of our theoretical proof. The line of Figure 4-a is not drawn because these theorems are based on the hypothesis that facts come regularly and continuously to an entity over time. This hypothesis holds when $p_e \leqslant 0.7$. Otherwise, when $p_e$ is close to 1 (which is the case of the Place of birth relationship for subjects), in most steps of the algorithm, new entities are added with only one associated fact per entity, which makes $P(1) \approx 1$. This brings us to a case where it is not really important to apply the model since the distribution is straightforward.

The asymmetry between subject and object is a truly surprising result. Yet experimental results show that our model faithfully reproduces most relationships for both subjects and objects. This means that for a relationship, the distribution of facts for subjects follows an exponential law, while the distribution of facts for objects follows a power law (with an exponent $\gamma > 2$). This is a new result, as no previous work has studied the distribution of facts with the fine granularity of relationships. In addition to injecting our stochastic algorithm to generate more realistic knowledge graphs, these distributions are of interest for data analysis [41] and query optimization (instead of resorting to heuristics [4]). In Section 6, we discuss such concrete applications of this new theoretical result with several immediate use cases.

Regarding the example of the birthplace relationship, applying the exponent formula of Theorem 2 on the objects in relationship `birthPlace`, we obtain $\gamma = 2.088$ for Wikidata, $\gamma = 2.113$ for DBpedia and $\gamma = 2.092$ for YAGO4. These three exponents are really very close showing that the underlying structure is similar between different KGs. Similarly, the structure of subjects in the cast member relationship in Wikidata remains approximately the same over the time with $\beta_{2015} = 0.197$, $\beta_{2017} = 0.236$ and $\beta_{2022} = 0.190$. For this reason, we will see in the next section that our model predicts Wikidata's evolution relatively accurately.

## 5. Experiments

The aim of this experimental study is to evaluate our model's performance and compare it with two baselines. Firstly, after presenting the followed protocol in Section 5.1 we examine whether our model KRELM accurately reproduces the underlying structure of relationships present in the data, by comparing the generated synthetic bipartite graph's degree distribution with the real relationship's degree distribution (see Section 5.2). Secondly, we investigate the model's ability to predict the evolution of a KG (see Section 5.3). This entails analyzing how well our generated graphs align with the observed changes over time in the real graph.

As it generates the relations containing the most facts in just a few seconds on a personal computer, we do not present runtime results of KRELM. It is implemented in Java. Its source code, the description of relationships for each dataset, and results are publicly available: https://scm.univ-tours.fr/habdallah/KRELM/

### 5.1. Protocol

*Processing of KGs* We rely on four crowdsourced KGs: DBnary [51], DBpedia [5], Wikidata [56] and YAGO4 [44], that are available on the Web. We especially focus on Wikidata for the longitudinal study for which we have a snapshot for each year. We filtered each dump to remove literals and external entities because our model aims at understanding the internal topology of the entities belonging to a given KG. Literal values such as dates, strings or images have therefore been removed. To focus on the graph, we only consider the entities whose Uniform Resource Identifier (URI) is prefixed by http://dbpedia.org/, http://www.wikidata.org/ or http://yago-knowledge.org for DBpedia, Wikidata and YAGO4 respectively. Table 2 indicates the main statistics of these four crowdsourced KGs after

| KG | DBnary | DBpedia | Wikidata | YAGO4 |
|---|---|---|---|---|
| #rel. $\|\mathcal{R}\|$ | 52 | 15,100 | 1,484 | 85 |
| #facts | 198,355,239 | 1,082,635,010 | 2,404,397,928 | 282,056,110 |
| #subjects | 54,540,338 | 245,113,095 | 431,875,708 | 48,957,049 |
| #objects | 38,069,118 | 93,130,240 | 136,870,611 | 19,214,906 |
| $\%r : p_s \geqslant 0.7$ | 78.84% | 71.00% | 75.94% | 72.94% |
| $\%r : p_o \geqslant 0.7$ | 21.15% | 67.25% | 36.52% | 17.64% |

Table 2
Main statistics of the four KGs

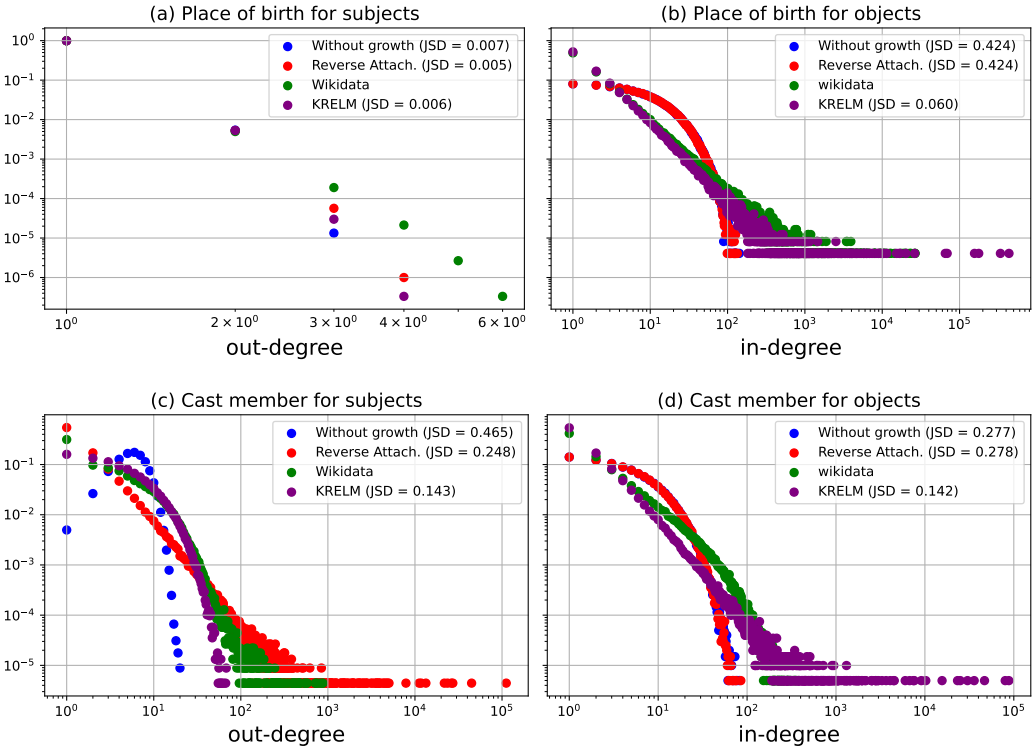


Fig. 5. Place of birth and Cast member relationships.

the preprocessing: number of relationships $|\mathcal{R}|$, number of facts, number of distinct subjects and number of distinct objects. The last two lines indicate the proportion of relationships having a growth probability $p_e = n_e/n$ greater than 0.7 for subjects and objects respectively. All statistics were computed by making ten passes on the dumps (five passes for in-degrees and five passes for out-degrees) in order not to overload the memory. The first pass consists in computing the number of entities and the number of facts reported in Table 2. The other four passes are used to compute the in/out-degree ground truth distributions per relationship. More precisely, the relationships are then divided into four groups $(R_i)_{i\in\{1,\dots,4\}}$. For each group $R_i$, we repeated a pass on the data that computes the number of entities $m_r$ and the number of facts $l_r$ for each relationship $r \in R_i$.

*Baselines* As mentioned in related work, there is no method in the literature specifically designed for generating bipartite graphs from basic statistics (see Table 1). In particular, it is not possible to directly compare our method with existing stochastic models dedicated to traditional simplex graphs. Other methods require different parameters (more specific than simple statistics), making comparison impossible. On the one hand, deep learning methods cannot be applied because they require a training set and are limited to small graphs. For instance, to reproduce

the birthplace relationship, we would need to retrieve equivalent relationships from many knowledge graphs. On the other hand, heuristic methods [6, 25, 38, 54] require the desired final distribution as an input parameter of their algorithms. For instance, for the birthplace relationship, this would be equivalent to estimating the exponent of the power law from real data, denoted as $\alpha_o$. Subsequently, entities and facts would be generated in a manner to follow the exponent $\alpha_o$. Therefore, to evaluate our method, we have conducted an ablation study, which is an experimental approach used to assess the contribution of each component of a model by removing or altering them. The main is to highlight the importance of the main ingredients of the model that we propose. This was done by introducing two baselines, one by eliminating continuous growth and the other by reversing the attachment functions. Here is how the baselines are created:

- Reverse attachment: This baseline takes our model and reverses the method of attachment between subjects and objects: linear preferential attachment for subjects and uniform attachment for objects. This baseline will enable us to assess the need for asymmetrical attachment between subjects and objects. It will also assess the gap between uniform attachment and linear preferential attachment. Growth in this baseline is applied as in our model.
- Without growth: This baseline takes our model by removing the notion of growth. It means all entities are initially generated with 1 fact (for instance, 281,253 towns for birthplace), the remaining facts (3,394,984 - 281,253 for birthplace) are generated one by one using attachment. Like the heuristic methods of Table 1, it generates all entities and then directly uses the final set to associate facts with them. In the model that we propose, continuous arrival implies that preferential attachment does not depend on the final distribution. Attachment in this baseline is applied as in our model (uniform for subjects and preferential for objects).

Note that these two baselines are illustrated for Place of birth and Cast member relationships in Figure 5 (in red for Reverse attach. and in blue for Without growth). Clearly, the two baselines do not reproduce real-world data as well, except for the reverse attachment considering the subjects of the birthplace relationship. In this case, the growth probability $n_s/n$ is close to 1 meaning that a subject is almost always added with a fact (i.e., the attachment has no impact thus the baseline and KRELM perform almost the same). We will see in the rest of this section that the performance of KRELM found for these two relationships is generalized to all relationships in different knowledge graphs. More generally, the comparison with these two baselines aims to evaluate whether the two main components of our model are crucial for achieving excellent results. This approach is common for validating a stochastic model in network science [8].

*Evaluation measures* To assess the effectiveness of the different approaches, we employ the Jensen-Shannon Divergence ($D_{JS}$) measure [39] to compare the degree distributions generated by a model with those obtained from real knowledge graphs. $D_{JS}$ is a symmetrized and smoothed version of the Kullback-Leibler Divergence, which quantifies the difference between two probability distributions. The Jensen-Shannon Divergence between two probability distributions, $P$ and $Q$, is calculated as follows:

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

where $D_{KL}(P||Q)$ represents the Kullback-Leibler Divergence between distributions $P$ and $Q$: $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$, and the distribution $M$ is the average of $P$ and $Q$: $M = (P + Q)/2$. The JSD measure robustly assesses the difference between degree distributions, providing a value between 0 and 1. This measure allows us to quantitatively evaluate how accurately our model captures the distribution of degrees observed in real-world KGs. We set a threshold of 0.2 for the Jensen-Shannon Divergence, which serves as a reference value of successful performance. In other words, we consider that the degree distribution $P$ generated by our model is similar to that of real-world $Q$ if $D_{JS}(P\|Q) \leqslant 0.2$. The chosen threshold of 0.2 is deemed effective as it visually demonstrates a significant resemblance (see supplementary materials). Obviously, varying this threshold modifies the numerical results; nevertheless, the conclusions remain the same (in particular, the superiority of KRELM over baselines).

When the number of entities in a relationship is low, the distribution of facts $P$ is unstable. Comparing the distributions $P$ and $Q$ then becomes less meaningful as there are likely to be large discrepancies. To take this phenomenon into account, we distinguish relationships with at least 500 entities from other relationships: $\mathcal{R}^* =$

$\{r \in \mathcal{R} : n_e \geqslant 500\}$. Thus, the *precision* of our model is the proportion of relationships in $\mathcal{R}^*$ successfully regenerated:

$$Prec(\mathcal{R}) = \frac{|\{r \in \mathcal{R}^* : D_{JS}(P_{\tilde{r}} \| P_r) \leqslant 0.2\}|}{|\mathcal{R}^*|}$$

where $P_{\tilde{r}}$ and $P_r$ are the degree distributions of the generated relationship and the original relationship $r$ respectively. Nevertheless, it remains informative to know the *coverage* of our model that is the proportion of relationships successfully regenerated among all the KG's relationships:

$$Cov(\mathcal{R}) = \frac{|\{r \in \mathcal{R} : D_{JS}(P_{\tilde{r}} \| P_r) \leqslant 0.2\}|}{|\mathcal{R}|}$$

Obviously, the closer precision and coverage are to 1, the better the model performs.

### 5.2. KG regeneration

For each KG, we regenerate relationships in order to measure the precision and coverage of our model and compare it to the two baselines. Tables 3 (a) and 3 (b) show the results of our model, Reverse attach. and Without growth by distinguishing between the degree distributions of subjects and objects. For each row, the best approach for precision and coverage is marked in bold.

We first observe that our model KRELM is better than the two baselines for subjects in Table 3 (a) (except for the precision in Wikidata and the coverage in DBpedia), with excellent coverage ranging between 0.734 and 0.929 and precision varying from 0.682 to 0.956. As expected, the baseline Without growth is significantly worse. Nonetheless, the improvement of our model for subjects compared to the baseline Reverse attach. is less impressive. Indeed, for subjects, the probability $p_s = n_s/n$ is close to 1 for many relationships (see the proportion of relationships with a growth probability $p_s$ larger than 0.7 in Table 2). Consequently, in such cases, our model and Reverse attach. behave similarly by adding new subjects in most iterations, with the attachment functions playing a minimal role. Nevertheless, in the case of YAGO4, there is a significant difference in precision/coverage (more than 5%), showing the benefit of uniform attachment for the subjects. Additionally, Table 4 (a) divides the precision of the model we propose and the two baselines into two categories. The first category represents precision calculated only for relations with $p_e < 0.7$, while the second category calculates precision only for relations with $p_e \geqslant 0.7$. This approach allows us to highlight the importance of the asymmetrical attachment function of the model. Indeed, for relations belonging to the second category (where the attachment is ignored), our model is close to the two baselines, with an average precision of 0.993 for subjects.

In contrast, when the attachment is really used (i.e., when $p_e < 0.7$), our model KRELM performs significantly better than both the Reverse attach. and the Without growth baselines, with an average precision of 0.634 for subjects across the four knowledge graphs.

More impressively, our model significantly outperforms the baseline for objects (except for the coverage in DBpedia), with coverage varying between 0.365 and 0.708 depending on the KG, and a very good precision ranging from 0.642 to 0.806. Also, for objects, our model outperformed the two baselines for relations with $p_e < 0.7$, achieving an average precision of 0.617 across the four knowledge graphs, and 0.996 for relations with $p_e \geqslant 0.7$. The excellent precision of our model KRELM means that it works very well for a large proportion of relationships that contain a sufficient number of entities. This means that it aids in understanding the behavior of the crowd when adding elements to these relationships and helps to identify the resulting structure from this process.

### 5.3. Longitudinal study on Wikidata

*Evolution of precision and coverage* For each year, we reproduced the Wikidata KG to assess the quality of our asymmetric bipartite graph model and compare it to the two baselines. Figure 6 plots the coverage and precision of our model (solid line) and the baselines (dashed line for Reverse attach. and dotted line for Without growth) on

| | KRELM | | Reverse attach. | | Without growth | |
|---|---|---|---|---|---|---|
| KG | Prec. | Cov. | Prec. | Cov. | Prec. | Cov. |
| DBnary | **0.956** | **0.923** | **0.956** | 0.903 | 0.826 | 0.788 |
| DBpedia | **0.682** | 0.734 | 0.638 | 0.717 | 0.619 | **0.811** |
| Wikidata | 0.912 | **0.846** | **0.917** | **0.846** | 0.846 | 0.811 |
| YAGO4 | **0.948** | **0.929** | 0.870 | 0.858 | 0.753 | 0.729 |
| Average | **0.874** | **0.858** | 0.845 | 0.831 | 0.761 | 0.785 |

(a) subjects

| | KRELM | | Reverse attach. | | Without growth | |
|---|---|---|---|---|---|---|
| KG | Prec. | Cov. | Prec. | Cov. | Prec. | Cov. |
| DBnary | **0.772** | **0.365** | 0.545 | 0.250 | 0.545 | 0.326 |
| DBpedia | **0.642** | 0.708 | 0.423 | 0.688 | 0.417 | **0.739** |
| Wikidata | **0.806** | **0.572** | 0.503 | 0.444 | 0.501 | 0.456 |
| YAGO4 | **0.681** | **0.588** | 0.287 | 0.247 | 0.301 | 0.270 |
| Average | **0.725** | **0.558** | 0.439 | 0.407 | 0.441 | 0.447 |

(b) objects

Table 3

Precision and coverage for four crowdsourced KGs for KRELM and the two baselines, for subjects (a) and objects (b)

| | KRELM | | Reverse attach. | | Without growth | |
|---|---|---|---|---|---|---|
| KG | Prec: $\frac{n_s}{n} < 0.7$ | Prec: $\frac{n_s}{n} \geqslant 0.7$ | Prec: $\frac{n_s}{n} < 0.7$ | Prec: $\frac{n_s}{n} \geqslant 0.7$ | Prec: $\frac{n_s}{n} < 0.7$ | Prec: $\frac{n_s}{n} \geqslant 0.7$ |
| DBnary | **0.777** | **1.0** | **0.777** | **1.0** | 0.222 | 0.973 |
| DBpedia | **0.372** | 0.981 | 0.276 | **0.988** | 0.264 | 0.962 |
| Wikidata | 0.609 | 0.993 | **0.621** | **0.996** | 0.308 | 0.990 |
| YAGO4 | **0.777** | **1.0** | 0.444 | **1.0** | 0.055 | 0.966 |
| Average | **0.634** | 0.993 | 0.523 | **0.996** | 0.212 | 0.973 |

(a) subjects

| | KRELM | | Reverse attach. | | Without growth | |
|---|---|---|---|---|---|---|
| KG | Prec: $\frac{n_o}{n} < 0.7$ | Prec: $\frac{n_o}{n} \geqslant 0.7$ | Prec: $\frac{n_o}{n} < 0.7$ | Prec: $\frac{n_o}{n} \geqslant 0.7$ | Prec: $\frac{n_o}{n} < 0.7$ | Prec: $\frac{n_o}{n} \geqslant 0.7$ |
| DBnary | **0.615** | **1.0** | 0.230 | **1.0** | 0.230 | **1.0** |
| DBpedia | **0.538** | **0.991** | 0.254 | 0.987 | 0.253 | 0.967 |
| Wikidata | **0.707** | **0.994** | 0.250 | 0.984 | 0.248 | 0.984 |
| YAGO4 | **0.611** | **1.0** | 0.129 | **1.0** | 0.166 | **1.0** |
| Average | **0.617** | **0.996** | 0.216 | 0.993 | 0.224 | 0.988 |

(b) objects

Table 4

Precision considering properties that have $\frac{n_e}{n} < 0.7$ or $\geqslant 0.7$ for four crowdsourced KGs for KRELM and the two baselines, for subjects (a) and objects (b).

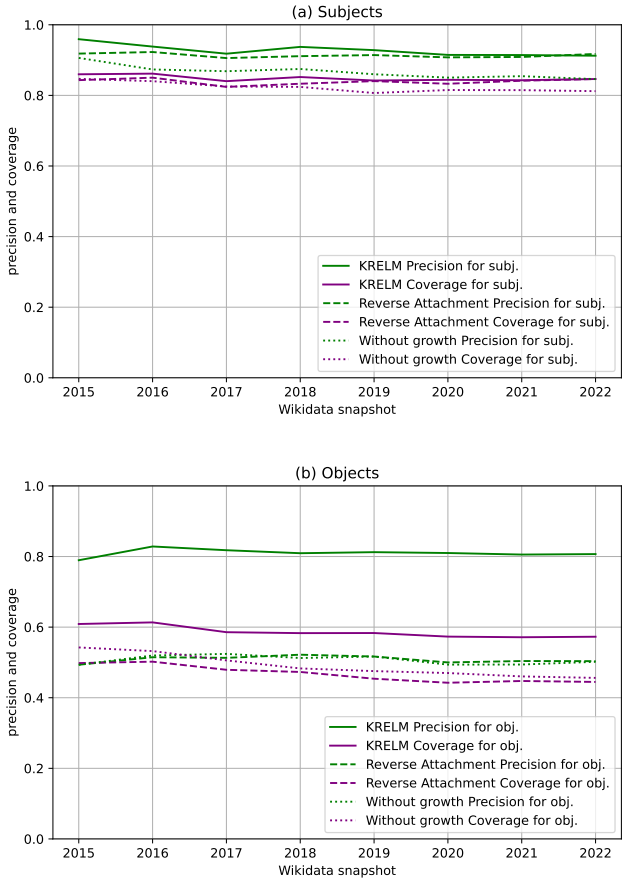


Fig. 6. The evolution of precision and coverage metrics of KRELM and the two baselines in Wikidata over time.

Wikidata over the years, applying it to both the degree distribution of subjects (a) and the degree distribution of objects (b). In Figure 6 (a), we can see that our model and the baseline Reverse attach. remain stable over time for subjects with high and similar values for both precision and coverage – our model is even slightly better. Again, this similarity can be explained by the fact that the probability of attachment is often close to 1, making the two approaches identical. In contrast, the baseline Without growth deteriorates significantly over time. It is known that this model is close to the model with growth at the beginning of the process but with time, the Without growth model tends towards a Gaussian distribution [9]. Considering the case of objects in Figure 6 (b), it is evident that our model outperforms the two baselines with better precision and coverage. Behind this stability, precision improves on relationships containing many facts, but it is offset as new relationships arrive. Conversely, we observe that Reverse attach. and Without growth deteriorate because they poorly reproduce relationships containing a lot of facts. It is essential to highlight that knowledge in crowdsourced knowledge graphs is incomplete. Nevertheless, our model significantly captures the structure of this knowledge, as evidenced by the high precision over the years. In general, our model is the most robust one over time.

*Predicting the evolution of Wikidata* Now, we want to check whether our model is capable of reproducing the dynamics of a KG by predicting the evolution of Wikidata. For each relationship $r \in \mathcal{R}^*$, we parameterized our model and the baseline Reverse attach. with probabilities $p_e^{2015} = n_e^{2015}/n^{2015}$, and then we generated bipartite graphs containing successively the number of facts for 2015 (i.e., $n^{2015}$), then the number of facts for 2016 (i.e., $n^{2016}$), and so on. For instance, for 2016, Algorithm 1 was applied with $n^{2016}$ facts and $p_e^{2015}$ for both subjects and objects. Note that the baseline Without growth has no year-related parameters like the growth probability $p_e^{2015}$.

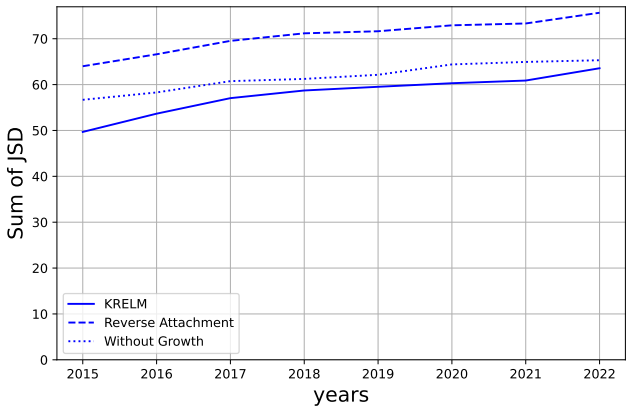


Fig. 7. Simulating Wikidata evolution over time.

Figure 7 reports the sum of the divergences $D_{JS}$ of the relationships for each annual version generated against the actual version for all entities (i.e., subjects and objects). Our model (solid line) has a significantly smaller sum of $D_{JS}$ across all years than the two baselines. Surprisingly, Without growth (dotted line) behaves better than Reverse attach. (dashed line) emphasizing the importance of uniform attachment for subjects and preferential attachment for objects. It is clear that the divergence of our model increases slightly, indicating that the generated facts slowly diverge from reality. Nevertheless, this sum, which significantly increases until 2018, then seems to stabilize despite Wikidata's strong evolution. This result is all the more impressive given that the number of subjects has risen from just 37M in 2015 to 432M in 2022 (similarly, there were 3M objects in 2015 versus 137M in 2022). This means that it is realistic to rely on our model and Wikidata's current parameters to predict its evolution once the volume of facts has increased in the future.

## 6. Discussion

The experimental study in the previous section has validated the ability of our model to reproduce real-world crowdsourced KGs at a given point in time. It also validates that the model reproduces the evolution in time, with the longitudinal study. We can therefore confidently answer Research Question 2: *The accumulation of facts in knowledge graphs regularly adds entities (growth), distributing facts evenly across subjects (uniform attachment) and, conversely, concentrating facts on certain objects (linear preferential attachment)*. Furthermore, the theoretical study of our model in response to RQ2 has enabled us to provide a precise answer to Research Question 1: *The structure of relations is globally stable for both subjects (see Theorem 1) and objects (see Theorem 2)*.

Now, we want to underline the importance of these fundamental results by discussing several practical use cases:

*Benchmark improvement* Our approach can improve methods aiming at generating synthetic graph data similar to real knowledge graphs, making them simpler and more realistic. Despite the strengths of the existing methods, they have limitations. Some methods [30] focus on creating benchmarks but overlook crucial aspects — the continuous growth in knowledge graphs, and the asymmetrical attachment so they use only one type of attachment for graph entities, whether they are subjects or objects. These aspects are important and removing them will alter the results as demonstrated in the experiment section. Some methods [6, 25, 38, 54] inject the desired distribution in the parameters of the algorithm (for example a power law with an exponent calculated from real data) and then aim to reproduce this distribution. Therefore, we believe that integrating our model into these existing methods will significantly enhance their results and optimize their algorithms, so they no longer need to specify the expected distributions and try to reproduce them.

*Model-based computation* At present, most approaches to knowledge graph analysis rely on data as an approximation to the actual distribution. This data is often costly to retrieve and manipulate. But above all the data itself remains a sample of an underlying distribution (i.e., the power or exponential laws we have identified) that it is often preferable to handle directly. For instance, the Gini coefficient [19] is a measure of statistical dispersion (useful in many applications including gap discovery [49]). When calculating the Gini coefficient for the distribution of facts related to a relationship like birthplace, determining the number of births for each city can be challenging due to current limitations in the Wikidata query service of time-out exceptions when a query needs high processing time. However, leveraging Theorem 2 and Newman's [41] formula, we can directly estimate the Gini coefficient for a relationship by executing only simple SPARQL queries, which request just the number of objects and the number of facts. This light approach simplifies the calculation process, especially for methods that need to calculate the Gini coefficient for multiple relations. We demonstrated the usefulness of this observation by building a method that automatically generates ranking indicators from Wikidata [1].

*Topological analysis* Moreover, the knowledge about how the bipartite graph of a relation is growing brings insights to many tasks performed on crowdsourced KGs. For instance, to the best of our knowledge, systems supporting Wikidata "patrollers" (who fight vandalism) are based on the topological characteristics of entities and do not consider the topological characteristics of relations. In network science, complex network models serve to evaluate the robustness of a network. In fields where they are applied, robustness is essentially about checking whether the removal of nodes or links breaks the network's connectedness [3, 33]. For example, removing central servers may prevent some remote machines from communicating with each other. In the case of a KG, we think that connectedness itself is less meaningful. Corrupting a KG is more about deleting information or inserting fake information. Intuitively, information is less vulnerable if it is based on a large number of facts. For example in Wikidata, by deleting the fact $\langle$`Fann Wong`, `place of birth`, `Singapore`$\rangle$, the information about Fann Wong' place of birth is permanently lost, but the information that Singapore is a place of birth persists with its numerous other births. The robustness of information about an entity is therefore proportional to its degree, and KRELM can be used to determine which parts of the KG are robust and which ones need to be monitored more closely. Another area where the focus is mostly on entities is data completion. For example, the Relative completeness indicator (RECOIN) tool [7] only considers the entity's description. With the model that we propose, it might be possible to suggest relations that are particularly lacking/abundant in facts or show an unexpected distribution.

## 7. Conclusion

This paper introduces the first complex network model for simulating relationships constructions in crowdsourced knowledge graphs. It completes the results already established by works focused on social editing, knowledge engineering, and KG profiling. Comparing to deep learning and heuristic methods for generating synthetic KGs, this new model includes continuous growth and asymmetrical attachment, which are more realistic for replicating KGs relationships. The model's theoretical study leads to two major new results: facts are distributed *on subjects* according to an exponential law, and they are distributed *on objects* according to a power law with an exponent greater than 2. The extensive experimental study on four KGs, with a longitudinal study concerning Wikidata, demonstrates the effectiveness of KRELM in reproducing current and predicting future relationships' structures. Experiments also demonstrate the model's generality in generating property distributions very similar to a large proportion of existing relationships, with excellent precision and coverage. They also show that it behaves better compared to two baselines, one without growth, and one with reverse attachments (evaluating the two main ingredients of our model). The source code of the model, description of relationships, and experimental results are provided: https://scm.univ-tours.fr/habdallah/KRELM/. This work provides a fundamental understanding of KGs that paves the way for numerous research directions. For instance, although KRELM ignores the interrelationships phenomena, it works very well for individual relationships. Thus, it could be used to generate benchmarks [6, 25, 30, 38, 54] with more realistic characteristics and less parameters by injecting our model. In data analysis [41] applied to KGs, having such a model makes it possible to analyze data evolution in greater detail, which is useful for anomaly detection and prediction. These are not the only directions that may be studied, another application field could also be

in query engine optimization: our model could be useful for refining cost models often relying on heuristics [2, 4]. A primary direct application of this model is presented in [1], where we use KRELM to optimize the complexity cost of computing the Gini coefficient.

## Acknowledgement

We thank Louise Parkin who recently joined our team for her rigorous proofreading and numerous suggestions for improvement.

## References

[1] H. Abdallah, B. Markhoff, and A. Soulet. Ranking Indicator Discovery from Very Large Knowledge Graphs. *Proceedings of the VLDB Endowment*, 18(4):1183–1195, 2024.

[2] I. Abdelaziz, R. Harbi, Z. Khayyat, and P. Kalnis. A survey and experimental comparison of distributed sparql engines for very large RDF data. *VLDB*, 10(13):2049–2060, 2017.

[3] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.

[4] W. Ali, M. Saleem, B. Yao, A. Hogan, and A.-C. N. Ngomo. A survey of rdf stores & sparql engines for querying knowledge graphs. *The VLDB Journal*, pages 1–26, 2022.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *ISWC*, pages 722–735, 2007.

[6] G. Bagan, A. Bonifati, R. Ciucanu, G. H. Fletcher, A. Lemay, and N. Advokaat. gmark: Schema-driven generation of graphs and queries. *IEEE Transactions on Knowledge and Data Engineering*, 29(4):856–869, 2016.

[7] V. Balaraman, S. Razniewski, and W. Nutt. Recoin: relative completeness in wikidata. In *Companion Proceedings of the The Web Conference 2018*, pages 1787–1792, 2018.

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[9] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999.

[10] A.-L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[11] V. Batagelj and U. Brandes. Efficient generation of large random networks. *Physical Review E*, 71(3):036113, 2005.

[12] C. Bekiari, G. Bruseker, M. Doerr, C.-E. Ore, S. Stead, A. Velios, M.-P. Blain, F. Bricaud, C. Crevier-Lalonde, S. Hart, et al. Cidoc crm. *International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Version*, 7(1), 2021.

[13] M. Ben Ellefi, Z. Bellahsene, J. G. Breslin, E. Demidova, S. Dietze, J. Szymański, and K. Todorov. Rdf dataset profiling–a survey of features, methods, vocabularies and applications. *Semantic Web*, 9(5):677–705, 2018.

[14] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics reports*, 544(1):1–122, 2014.

[15] A. Bojchevski, O. Shchur, D. Zügner, and S. Günnemann. Netgan: Generating graphs via random walks. In *International conference on machine learning*, pages 610–619. PMLR, 2018.

[16] B. Bollobás, C. Borgs, J. T. Chayes, and O. Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.

[17] J. A. Bondy and U. S. R. Murty. *Graph theory*. Springer Publishing Company, Incorporated, 2008.

[18] S. Cebiric, F. Goasdoue, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika. Summarizing Semantic Graphs: A survey. *The VLDB Journal*, 28:295–327, 2018.

[19] L. Ceriani and P. Verme. The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini. *The Journal of Economic Inequality*, 10:421–443, 2012.

[20] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-mat: A recursive model for graph mining. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 442–446. SIAM, 2004.

[21] F. Chung, F. R. Chung, F. C. Graham, L. Lu, et al. *Complex graphs and networks*. Number 107. American Mathematical Soc., 2006.

[22] N. De Cao and T. Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[23] G. Demartini. Implicit bias in crowdsourced knowledge graphs. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 624–630, 2019.

[24] M. Drobyshevskiy and D. Turdakov. Random graph modeling: A survey of the concepts. *ACM computing surveys (CSUR)*, 52(6):1–36, 2019.

[25] Z. Feng, W. Mayer, K. He, S. Kwashie, M. Stumptner, G. Grossmann, R. Peng, and W. Huang. A schema-driven synthetic knowledge graph generation approach with extended graph differential dependencies (gdd x s). *IEEE Access*, 9:5609–5639, 2020.

[26] R. Ferrer i Cancho, O. Riordan, and B. Bollobas. The consequences of zipf's law for syntax and symbolic reference. *Proc. R. Soc. B*, 272:561–565, 2005.

[27] F. Giroire, S. Pérennes, and T. Trolliet. A random growth model with any real or theoretical degree distribution. *Theoretical Computer Science*, 940:36–51, 2023.

[28] F. Goasdoue, P. Guzewicz, and I. Manolescu. RDF graph summarization for first-sight structure discovery. *The VLDB Journal*, 29(5):1191–1218, 2020.

[29] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[30] Y. Guo, Z. Pan, and J. Heflin. Lubm: A benchmark for owl knowledge base systems. *Journal of Web Semantics*, 3(2-3):158–182, 2005.

[31] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, pages 211–220, 2007.

[32] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37, 2021.

[33] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han. Attack vulnerability of complex networks. *Physical review E*, 65(5):056109, 2002.

[34] K. Kellou-Menouer, N. Kardoulakis, G. Troullinou, Z. Kedad, D. Plexousakis, and H. Kondylakis. A survey on semantic schema discovery. *The VLDB Journal*, pages 1–36, 2021.

[35] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pages 1945–1954. PMLR, 2017.

[36] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018.

[37] X. Liu, Z. Tu, Z. Wang, X. Xu, and Y. Chen. A crowdsourcing-based knowledge graph construction platform. In *International Conference on Service-Oriented Computing*, pages 63–66. Springer, 2020.

[38] A. Melo and H. Paulheim. Synthesizing knowledge graphs for link and type prediction benchmarking. In *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28–June 1, 2017, Proceedings, Part I 14*, pages 136–151. Springer, 2017.

[39] M. Menéndez, J. Pardo, L. Pardo, and M. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

[40] M. Newman. *Networks*. Oxford university press, 2018.

[41] M. E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[42] A. Oelen, M. Stocker, and S. Auer. Tinygenius: intertwining natural language processing with microtask crowdsourcing for scholarly knowledge graph creation. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5, 2022.

[43] H. Park and M.-S. Kim. Trilliong: A trillion-scale synthetic graph generator using a recursive vector model. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 913–928, 2017.

[44] T. Pellissier Tanon, G. Weikum, and F. Suchanek. Yago 4: A reason-able knowledge base. In *ESWC*, pages 583–596, 2020.

[45] G. Piao and W. Huang. Learning to predict the departure dynamics of Wikidata editors. In *International Semantic Web Conference*, pages 39–55. Springer, 2021.

[46] A. Piscopo and E. Simperl. Who models the world? collaborative ontology creation and user roles in Wikidata. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, 2018.

[47] A. Polleres, R. Pernisch, A. Bonifati, D. Dell'Aglio, D. Dobriy, S. Dumbrava, L. Etcheverry, N. Ferranti, K. Hose, E. Jiménez-Ruiz, et al. How does knowledge evolve in open knowledge graphs? *Transactions on Graph Data and Knowledge*, 1(1):11–1, 2023.

[48] K. Rabbani, M. Lissandrini, and K. Hose. Extraction of validating shapes from very large knowledge graphs. *PVLDB*, 16(5):1023–1032, 2023.

[49] M. Ramadizsa, F. Darari, W. Nutt, and S. Razniewski. Knowledge gap discovery: A case study of wikidata. 2023.

[50] C. Sarasua, E. Simperl, N. F. Noy, A. Bernstein, and J. M. Leimeister. Crowdsourcing and the semantic web: A research manifesto. *Human Computation*, 2(1), 2015.

[51] G. Sérasset. DBnary: Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web*, 6(4):355–361, 2015.

[52] M. Simonovsky and N. Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, pages 412–422. Springer, 2018.

[53] E. Simperl and M. Luczak-Rösch. Collaborative ontology engineering: a survey. *The Knowledge Engineering Review*, 29(1):101–131, 2014.

[54] Y. Theoharis, G. Georgakopoulos, and V. Christophides. Powergen: A power-law based generator of RDFS schemas. *Information Systems*, 37(4):306–319, 2012.

[55] F. Van Harmelen, V. Lifschitz, and B. Porter. *Handbook of knowledge representation*. Elsevier, 2008.

[56] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[57] D. Vrandečić, L. Pintscher, and M. Krötzsch. Wikidata: The making of. In *the ACM Web Conference 2023*, pages 615–624, 2023.

[58] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[59] J. You, R. Ying, X. Ren, W. Hamilton, and J. Leskovec. GraphRNN: Generating realistic graphs with deep auto-regressive models. In *International conference on machine learning*, pages 5708–5717. PMLR, 2018.

[60] J. Zhang. Knowledge learning with crowdsourcing: A brief review and systematic perspective. *IEEE/CAA Journal of Automatica Sinica*, 9(5):749–762, 2022.