Representativeness of Knowledge Bases with the Generalized Benford's Law

Arnaud Soulet, Arnaud Giacometti, Béatrice Bouchou Markhoff, Fabian M. Suchanek

Representativeness of Knowledge Bases with the Generalized Benford's Law

Arnaud Soulet[1], Arnaud Giacometti[1], Béatrice Markhoff[1], and Fabian M. Suchanek[2]

[1] Université de Tours, LIFAT
firstname.lastname@univ-tours.fr
[2] Telecom ParisTech, LTCI
suchanek@telecom-paristech.fr

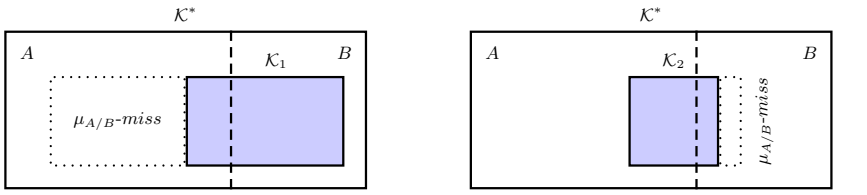
Abstract. Knowledge bases (KBs) such as DBpedia, Wikidata, and YAGO contain a huge number of entities and facts. Several recent works induce rules or calculate statistics on these KBs. Most of these methods are based on the assumption that the data is a representative sample of the studied universe. Unfortunately, KBs are biased because they are built from crowdsourcing and opportunistic agglomeration of available databases. This paper aims at approximating the representativeness of a relation within a knowledge base. For this, we use the generalized Benford's law, which indicates the distribution expected by the facts of a relation. We then compute the minimum number of facts that have to be added in order to make the KB representative of the real world. Experiments show that our unsupervised method applies to a large number of relations. For numerical relations where ground truths exist, the estimated representativeness proves to be a reliable indicator.

1 Introduction

One of the undisputed successes of the Semantic Web is the construction of huge knowledge bases (KBs). Several recent works use these KBs to derive new knowledge by calculating statistics or deducing rules from the data [7,26,27,29]. For instance, according to DBpedia, 99% of the places in Yemen have a population of more than 1,000 inhabitants. Thus, we could conclude that Yemeni cities usually have more than 1,000 inhabitants. But is that true in the real world?

Naturally, the reliability of such conclusions depends on the quality of the knowledge base [34] namely its correctness (accuracy of the facts) and its completeness. It is well known that KBs are highly incomplete. This is usually not a problem in statistics and in machine learning, where it is rare to have a complete description of the universe under study. Most approaches work on a sample of the data. In such cases, it is crucial that this sample is representative of the entire universe (or at least, that the bias of this sample is known). For example, it is not a problem if the KB contains only half of the cities of Yemen, if their distribution across different sizes corresponds roughly to the distribution in the real world. Figure 1 illustrates this: there is an ideal knowledge base $\mathcal{K}^*$ divided

into two classes A and B that correspond respectively to the places with less than 1,000 inhabitants and other places. The KB $\mathcal{K}_1$ is more complete than the KB $\mathcal{K}_2$. However, $\mathcal{K}_2$ better reflects the distribution between the two classes.



(a) More complete, less representative (b) Less complete, more representative

Fig. 1. Completeness vs Representativeness

Unfortunately, it is not clear whether the data in KBs is representative of the real world. For example, several large KBs, such as DBpedia [2] or YAGO [28], extract their data from Wikipedia. Wikipedia, in turn, is a crowdsourced dataset. In crowdsourcing, contributers tend to state the information that interests them most. As a result, Wikipedia exhibits some cultural biases [6,33]. Inevitably, these biases are reflected in the KBs. For instance, 3,922 entities in DBpedia concern the American company "Disney", which is almost as much as the 4,493 entities concerning Yemen (a country with more than 26 million inhabitants). Wikidata [32], likewise, is the result of crowdsourcing, and may exhibit similar biases. In particular, it is likely that countries such as Yemen are less evenly covered than places such as France – due to the population of contributors. Even if the information in these KBs is correct [13], it is not necessarily representative. If we knew how representative a certain KB is, then we could know whether it is reasonable or not to exploit it for deriving statistics. Such an indication should, for example, prevent us from drawing hasty conclusions about the distribution of the population in the cities of Yemen. But, how to estimate whether a knowledge base is representative or not?

This paper proposes to study the representativeness of knowledge bases by help of the generalized Benford's law. This parameterized law indicates the frequency distribution expected by the first significant digit in many real-world numerical datasets. We use this law as a gold standard to estimate how much data is missing in the KB. More specifically, our contributions are as follows:

- We present a method to calculate a lower bound for the number of missing facts for a relation to be representative. This method works in a supervised context (where the relation is known to satisfy the generalized Benford's law), and in an unsupervised context (where the parameter of the law has to be deduced from the data).
- We prove that, under certain assumptions, the calculated lower bounds are correct both in the supervised and the unsupervised context.

– We show with experiments on real KBs that our method is effective for supervised contexts as well as for unsupervised contexts. The unsupervised method, in particular, can audit 63% of DBpedia's facts.

This paper is structured as follows. Section 2 reviews some related work. Section 3 introduces the basic notions of representativeness. In Section 4, we propose our method for approximating representativeness based on the generalized Benford's law. Section 5 provides experimental results. We conclude in Section 6.

2 Related Work

To the best of our knowledge, the representativeness of knowledge bases with respect to the real world has not yet been studied. Nevertheless, as mentioned in the introduction, this problem is related to the completeness of KBs.

Completeness. Several recent works have studied the completeness of KBs [25,34]. Some works propose to manually add information about the completeness relations [8]. Other approaches mine rules on the data [12] (e.g., people usually live in the city where they work) and propose to add this information where it is missing. For this purpose, the work of [12] makes the Partial Completeness Assumption (PCA): It assumes that, if the KB contains at least one object for a given relation and a given subject, then it contains all of the objects for this context. The PCA has been shown to be reasonably accurate in practice [12]. Newer approaches for rule mining take into account the cardinality of the relations, if it is known [30]. Other work aims to determine more generally whether all objects of a certain relation for a certain subject are present in the KB [11]. For this, the approach uses oracles, such as the PCA and the popularity of the subject in Wikipedia. Again other work [1,14,17,31] mines class descriptions. Such approaches are able to determine that a certain attribute is obligatory for a class – and then allow estimating the number of missing facts per class.

All of these approaches are concerned with completeness in terms of facts with respect to the present entities. Our approach, in contrast, also considers the facts of entities that are missing. Furthermore, none of the above works studies the representativeness of the KB, i.e., whether or not the distribution of entities in the KB corresponds to the distribution in the real world.

Representative sample. Completeness is an important notion for estimating the quality of a knowledge base, but it is not necessarily the best indicator when one wants to measure the quality of a distribution. In statistics, several resampling techniques [9] exist to estimate the quality of a sample (median, variance, quantile), in particular by analyzing the evolution of a measure on a subsample or by permuting labels. None of these techniques can be used to check whether a single sample is representative, if the ground truth is unknown – as it is the case in our scenario.

Benford's law. When the data is complete, Benford's law [4] is regularly used to detect inconsistencies within the data [22]. If the distribution of the first significant digit of some numerical dataset does not satisfy Benford's law, then

the data is assumed to be faulty. For this reason, Benford's law is regularly used to detect frauds in various kind of data: in accounts [23], in elections [19], or in wastewater treatment plant discharge data [3]. However, in all of these cases, Benford's law is used only to estimate the correctness of the data – not its completeness. The work cannot be used, e.g., to decide how many facts are missing in a KB, or whether a KB is representative of the real world.

3 Preliminaries

3.1 Representativeness of knowledge bases

For our purposes, a knowledge base (KB) over a set of relations $\mathcal{R}$ and a set of constants $\mathcal{C}$ (representing entities and literals) is a set of *facts* $\mathcal{K} \subseteq \mathcal{R} \times \mathcal{C} \times \mathcal{C}$. We write facts as $r(s, o) \in \mathcal{K}$, where r is the relation, s is the subject, and o is the object. The set of facts for the relation r in $\mathcal{K}$ is denoted by $\mathcal{K}_{|r} = \{r(s, o) \in \mathcal{K}\}$. Given a relation r, $r^{-1}(o, s) \in \mathcal{K}$ means that $r(s, o) \in \mathcal{K}$ where r^{-1} is the inverse relation of r.

In line with the other work in the area [11,17,18,21,24], we denote with $\mathcal{K}^*$ a hypothetical ideal KB, which contains all facts of the real world. Then, the completeness (also called recall) of $\mathcal{K}$, denoted $comp(\mathcal{K})$, is the proportion of facts of $\mathcal{K}^*$ present in $\mathcal{K}$: $comp(\mathcal{K}) = |\mathcal{K} \cap \mathcal{K}^*|/|\mathcal{K}^*|$. For our work, we will make the following assumption:

Assumption 1 (Correctness) *Given a knowledge base $\mathcal{K}$, we assume that all facts of $\mathcal{K}$ are correct i.e., $\mathcal{K} \subseteq \mathcal{K}^*$.*

The correctness assumption is a strong assumption. It has been investigated in [28,34]. In our work, we use it mainly for our theoretical model. Our experiments will show that our method delivers good results even with some amount of noise in the data. Let us now introduce the notion of a *uniform-sampling invariant measure*. A measure μ maps a knowledge base $\mathcal{K}$ to a frequency vector $(f_1, \ldots, f_n) \in \mathbb{R}^n_{\geq 0}$ where each component f_i is the number of observations of the ith characteristic in $\mathcal{K}$. Given a non-zero frequency vector $F = (f_1, \ldots, f_n)$, $\overline{f_i}$ denotes the normalized ith component of F where $\overline{f_i} = f_i / \sum_{i=1}^{n} f_i$. We use the mean absolute deviation (MAD) for comparing two non-zero frequency vectors $F = (f_1, \ldots, f_n)$ and $F' = (f'_1, \ldots, f'_n)$:

$$MAD(F, F') = \frac{1}{n} \sum_{i=1}^{n} \left| \overline{f_i} - \overline{f'_i} \right|$$

F and F' are similar for $\epsilon \ll 1$ iff $MAD(F, F') \leq \epsilon$. In such case, we write $F \sim_\epsilon F'$, or simply $F \sim F'$. A measure μ is uniform-sampling invariant iff for any uniform sample $\mathcal{K}'$ from $\mathcal{K}$ such that $|\mathcal{K}'| \gg 1$, we have $\mu(\mathcal{K}') \sim \mu(\mathcal{K})$. For instance, in Figure 1, counting the number of places with less than 1,000 inhabitants (in part A) and more than 1,000 inhabitants (in part B) is a measure with two characteristics (denoted by $\mu_{A/B}$). The measure $\mu_{A/B}$ is uniform-sampling

invariant because whatever the uniform sample of a knowledge base $\mathcal{K}$, the proportion of cities with more (or less) than 1,000 inhabitants remains the same. In the following, we consider only uniform-sampling invariant measures.

A knowledge base is *representative* if each measure returns a frequency vector that is proportional to the frequency vector on $\mathcal{K}^*$:

Definition 1 (Representative KB). *A knowledge base $\mathcal{K}$ is representative of $\mathcal{K}^*$ iff $\mu(\mathcal{K}) \sim \mu(\mathcal{K}^*)$ for any uniform-sampling invariant measure μ.*

If a knowledge base $\mathcal{K}$ is unrepresentative, there is at least one measure μ such that $\mu(\mathcal{K}) \nsim \mu(\mathcal{K}^*)$. In this case, since all the facts of $\mathcal{K}$ are correct (Assumption 1), it would be necessary to add new facts to the knowledge base to make it representative for μ. Formally, this number of missing facts of $\mathcal{K}$ for the measure μ, denoted by $\mu\text{-}miss(\mathcal{K})$, is defined as:

$$\mu\text{-}miss(\mathcal{K}) = \min\{|F| : F \subseteq \mathcal{K}^* \wedge \mu(\mathcal{K} \cup F) \sim \mu(\mathcal{K}^*)\}$$

The number of missing facts in $\mathcal{K}$, denoted by $miss(\mathcal{K})$, is the minimum number of facts that have to be added to make the KB representative (whatever the considered measure μ): $miss(\mathcal{K}) = \max_\mu \mu\text{-}miss(\mathcal{K})$. The representativeness of $\mathcal{K}$ estimates whether $\mathcal{K}$ is a representative sample of $\mathcal{K}^*$:

Definition 2 (Representativeness). *The representativeness of $\mathcal{K}$, denoted $rep(\mathcal{K})$, is defined as:*

$$rep(\mathcal{K}) = \frac{|\mathcal{K}|}{|\mathcal{K}| + miss(\mathcal{K})}$$

Interestingly, a KB can be representative without being complete. The representativeness of $\mathcal{K}$ is an upper bound of the completeness: $rep(\mathcal{K}) \geq comp(\mathcal{K})$.

3.2 Problem statement

The goal of this paper is to approximate the representativeness of a relation r in $\mathcal{K}$ (i.e., the representativeness of $\mathcal{K}_{|r}$) without having a reference knowledge base $\mathcal{K}^*_{|r}$ (which is the most common case in a real-world scenario). This task is ambitious because the calculation of the representativeness of a knowledge base requires to know the distribution of any measure μ on an unknown knowledge base $\mathcal{K}^*_{|r}$. It is obviously not possible to know the distribution $\mu(\mathcal{K}^*_{|r})$ for any measure. In order to calculate an approximation, we propose to use the following observation, which holds for all measures μ:

$$\mu\text{-}miss(\mathcal{K}_{|r}) \leq miss(\mathcal{K}_{|r})$$

This result (which follows from the definition of $miss(\mathcal{K}_{|r})$) means that it is possible to get a lower bound l of the number of missing facts $miss(\mathcal{K}_{|r})$, if some distributions $\mu_i(\mathcal{K}^*_{|r})$ are known. Such a lower bound is useful for calculating an upper bound of the representativeness and the completeness of the knowledge base: $|\mathcal{K}_{|r}|/(|\mathcal{K}_{|r}| + l)$.

Given a knowledge base $\mathcal{K}$ and a relation r, we aim at estimating the representativeness of the relation r in the knowledge base $\mathcal{K}$ by finding a lower bound l such that $l \leq miss(\mathcal{K}_{|r})$.

4 Our Approach

4.1 The generalized Benford's law for KBs

The challenge is to find a set of measures whose distribution is known on the ideal knowledge base $\mathcal{K}^*$. To this end, we propose to rely on Benford's law [4]. This law says that, in many natural datasets, the first significant digit of the numbers is unevenly distributed: Around 30% of numbers will start with a "1", whereas only 5% of numbers will start with a "9". This somehow surprising result follows from the fact that many natural numbers follow a multiplicative growth pattern. For example, a city of 1000 inhabitants may grow by 30% each year, thus passing by the values of 1300, 1690, 2197, 2856, 3712, 4826, 6274, 8157, 10604. These values already show a skewed distribution of the first digit, which will repeat itself in the coming years. There are other reasons for such patterns, and Benford's law has since been observed not just for population sizes, but also for prices, stock markets, death rates, lengths of rivers, and many other real-world phenomena [4] – although not all [20]. Technically, Benford's law is a statistical frequency distribution on the first significant digit of a set of numbers, which may or may not apply to a given dataset. In this paper, we use the generalized Benford's law [16], which is parametrized and can thus apply to more datasets.

Definition 3 (Generalized Benford's Law [15]). *A set of numbers is said to satisfy a generalized Benford's law (GBL) with exponent $\alpha \neq 0$ if the first digit $d \in [1..9]$ occurs with probability:*

$$B_d^\alpha = \frac{(1+d)^\alpha - d^\alpha}{10^\alpha - 1}$$

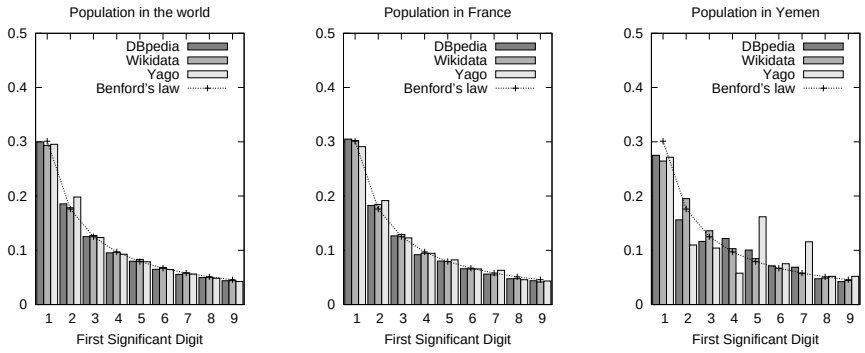


Fig. 2. First significant digit distribution for population

The parameter α adds a great flexibility since the choice of this value makes it possible to find Benford's law ($\alpha \to 0$) and the uniform law ($\alpha = 1$). Data

that follows a power law ax^{-k} also follows the GBL approximately with $\alpha = -1/k$ [15]. This is, e.g., the case for the out-degree of Web pages [5], with $k = 2.6$.

The GBL can be applied to KBs. Let us look at the relation `pop`, which links a geographical place to its number of inhabitants (`populationTotal` in DBpedia, `P1082` in Wikidata, and `hasNumberOfPeople` in YAGO). Figure 2 shows the distribution of first digits of this relation, drilled down to places in the world, in France, and in Yemen. We see that the distribution in the KB roughly follows the GBL. Interestingly, the GBL applies better to the French population than to the Yemeni population. We will now take advantage of this information to measure representativeness.

Technically, Figure 2 presents the frequency vector $(f_1, \ldots, f_9)$ of the first digits of the relation `pop`. Of course, it is not possible to directly calculate the ideal frequency vector $(f_1^*, \ldots, f_9^*)$ of $\mathcal{K}^*$. However, in many cases, we know at least the distribution of the ideal frequency vector (thanks to the GBL). If we do not know the distribution, then our idea is to *learn* the exponent α of the GBL from the observed vector. Once the ideal distribution has been determined, we can use the difference between the observed distribution and the estimated distribution to bound the number of missing facts (Figure 3).

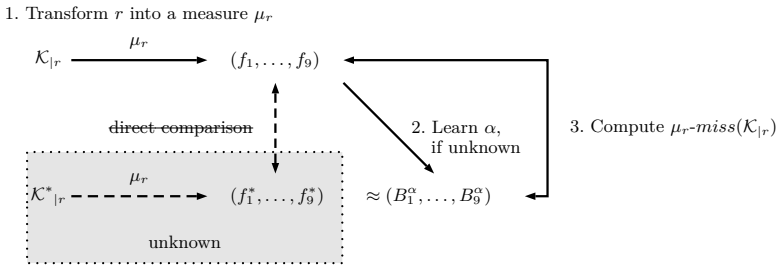


Fig. 3. Overview of the method

More precisely, we propose to proceed as follows:

1. **Transforming a relation into a measure:** Benford's law can only work on numerical datasets. Some relations (such as `pop`) are already numerical. Other relations will have to be transformed into numerical datasets (Section 4.2).
2. **Parameterizing the GBL:** To use the GBL, we have to know the parameter α. We distinguish two contexts. In a *supervised* context, the parameter α is known upfront in the real world (as it is the case for the population). Otherwise, in an *unsupervised* context, we learn the parameter α that best fits the facts in $\mathcal{K}_{|r}$ assuming it is close to the ideal parameter α^* on $\mathcal{K}^*{}_{|r}$ (Section 4.3).
3. **Estimating the number of missing facts:** As the knowledge base is correct, only the addition of new facts would make the frequency vector $(f_1, \ldots, f_9)$ coincide with the distribution of $(B_1^\alpha, \ldots, B_9^\alpha)$ which is (approxi-

mately) proportional to $(f_1^*, \ldots, f_9^*)$. The objective of this last step is to calculate the minimum number of facts to add so that $(f_1, \ldots, f_9) \sim (B_1^\alpha, \ldots, B_9^\alpha)$ (Section 4.4).

In the following, when we consider a relation r, $\mathcal{K}$ implicitly refers to $\mathcal{K}_{|r}$.

4.2 Transforming relations into measures

We show in this section how to transform a relation r into a measure μ_r. The key idea is to transform each relation r into a set of numbers N_r that is a kind of digital signature. Then, we derive a measure μ_r that counts the frequency of each number in N_r having d as first significant digit:

$$\mu_r(\mathcal{K}) = (\#n : \text{the first significant digit of } n \in N_r(\mathcal{K}) \text{ is equal to } d)_{d \in [1..9]}$$

In our example with the relation `pop`, the measure $\mu_{\texttt{pop}}$ counts the number of places that have a population with d as first significant digit. Let us now generalize this principle to two common types of relations:

- **Numerical transformation:** Given a numerical relation r, the numerical transformation keeps all the numbers different from 0:

  $$N_r^{num}(\mathcal{K}) = \{number : r(s, number) \in \mathcal{K} \wedge number \neq 0\}$$

  Figure 2 illustrates this transformation for relation `pop` by showing the frequency vector resulting from $\mu_{\texttt{pop}}$.
- **Counting transformation:** Given a relation r, the counting transformation returns for each object o how many facts it has:

  $$N_r^{count}(\mathcal{K}) = \{\#s : r(s, o) \in \mathcal{K} \text{ such that } o \text{ is an object of a fact in } \mathcal{K}_{|r}\}$$

  For example, for the relation `starring`, we can count the number of movies for each actor. The left hand-side of Figure 4 illustrates the resulting frequency vector. We choose to count the number of subjects rather than the number of objects, because relations tend to have more subjects per object than vice versa [12]. However, we can also count the number of objects per subject by applying the above method to r^{-1}. Figure 4 shows two other histograms, one for the relation `team` (number of players per team) and for `birthPlace` (number of births per place).

This list of transformations is not exhaustive. For instance, it would be possible to count the number of days since today for a date (e.g. for the birth date relation) or to consider the length of strings. Besides, it is possible to transform the same relation in several ways. In this way, it is possible to obtain more frequency vectors.

4.3 Parameterizing the generalized Benford's law

The previous section has given us a measure μ_r that we can apply on the knowledge base $\mathcal{K}$ to calculate a distribution. Now, we want to compare this distribution with the distribution on the ideal KB $\mathcal{K}^*$. This requires knowledge of the

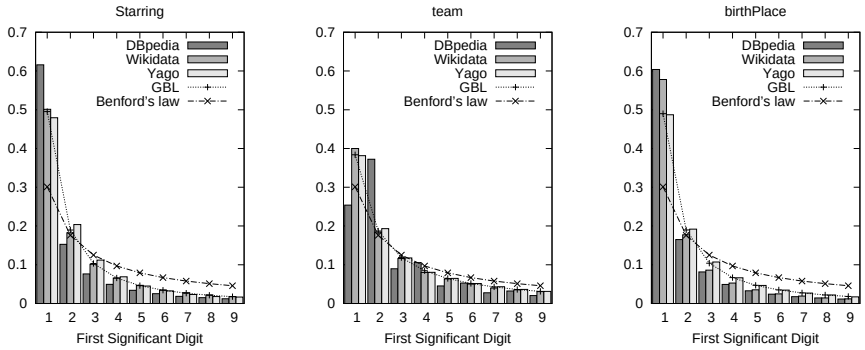


Fig. 4. Examples of measures resulting from counting transformation

parameter α, which depends on the unknown distribution $\mu_r(\mathcal{K}^*)$. We distinguish two settings.

Supervised setting. In some cases, it is known that $\mu_r(\mathcal{K}^*)$ follows the GBL in the real world with a certain parameter α. For instance, the population of places, the length of rivers, etc. conform to the GBL in the real world with an exponent tending to 0 (see Table 2 below). In that case, the GBL is already parametrized.

Unsupervised setting. If it is not known whether $\mu_r(\mathcal{K}^*)$ follows the GBL, or if its parameter α is not known, we propose to estimate it from the KB. For this purpose, we make the following assumption:

Assumption 2 (Transferability) *Given a knowledge base $\mathcal{K}$, we assume that if $\mathcal{K}$ conforms to the GBL with exponent α, then the ideal knowledge base $\mathcal{K}^*$ also conforms to the GBL with exponent α.*

This assumption may seem strong. However, it is verified in several cases where we have a ground truth available (see experiments in Section 5). The assumption allows us to learn the parameter α that best fits the facts in $\mathcal{K}$. Let us denote by $(f_1, \ldots, f_9)$ the characteristic vector resulting from $\mu_r(\mathcal{K})$ i.e., f_d is exactly the number of occurrences in $N_r(\mathcal{K})$ with d as first significant digit. Let us denote $N = \sum_{d=1}^{9} f_d$. To choose the right parameter α, we use the WLS measure (probability weighted least square or Chi square statistics) as goodness-of-fit measure [15]:

$$WLS_{(f_1,\ldots,f_9)}(\alpha) = \sum_{d=1}^{9} \frac{\left(B_d^\alpha - \frac{f_d}{N}\right)^2}{B_d^\alpha}$$

Now, choosing the right parameter α means minimizing the WLS measure for the frequency vector $(f_1, \ldots, f_9)$. For this, we use the gradient descent algorithm. For instance, Figure 4 shows the gap between the GBL and Benford's law for the three relations. For `starring`, α is -1.156 (in DBpedia), -0.759 (in Wikidata)

and -0.750 (in YAGO). Once the parameter α has been obtained, we have to assess whether the frequency vector $\mu_r(\mathcal{K})$ conforms to the generalized Benford's law. For this, we use the mean absolute deviation (MAD) defined in Section 3.1. To know whether the GBL can be used according to the MAD estimator, we distinguish four cases [16,22]: close conformity (C) when $MAD \leq 0.006$, acceptable conformity (AC) when $0.006 < MAD \leq 0.012$, marginal conformity (MC) when $0.012 < MAD \leq 0.015$, and nonconformity (NC) otherwise. In our running examples, the measure $\mu_{\texttt{pop}}$ gives rise to a nonconformity only for Yemeni places in YAGO, because $\alpha = 0.351$ and $MAD(\mu_{\texttt{pop}}(\mathcal{K}), B^{0.351})$ equals 0.035 (> 0.015). If a measure μ_r leads to a nonconformity, then it is not possible to apply the GBL at all. In all other cases, we can estimate the number of missing facts for the relation r as explained in the next section.

4.4 Estimating the number of missing facts

The purpose of this section is to estimate the number of missing facts for a relation r, knowing that we have an approximation of the expected distribution $(B_1^\alpha, \ldots, B_9^\alpha)$ that is proportional to $(f_1^*, \ldots, f_9^*)$. We assume that all the facts of the knowledge base $\mathcal{K}$ are correct (Assumption 1). Therefore, only the addition of facts can bring the observed distribution of facts $(f_1, \ldots, f_9)$ closer to the expected distribution $(B_1^\alpha, \ldots, B_9^\alpha)$.

Numerical transformation. When a relation is numerical, the only way to have a number with a given first significant digit is to add a new fact. Intuitively, it is then enough to add facts for each of the digits where the measured frequency is lower than the expected frequency. The following theorem formalizes this idea:

Theorem 1. *Given a knowledge base $\mathcal{K}$ and a measure μ_r^{num} such that $\mu_r^{num}(\mathcal{K}^*)$ satisfies a generalized Bendford's law with exponent α, the number of missing facts for the relation r is:*

$$\mu_r^{num}\text{-}miss(\mathcal{K}) = \max_{d \in [1..9]} \frac{f_d}{B_d^\alpha} - N$$

where $(f_1, \ldots, f_9) = \mu_r(\mathcal{K})$ and $N = \sum_{d=1}^{9} f_d$.

This follows from the fact that the expected distribution $f_d/(N+\mu_r^{num}\text{-}miss(\mathcal{K}))$ must be less than B_d^α for each digit d. Table 1 indicates the number of missing facts estimated for the relation `pop` with the unsupervised method, and deduces an approximation of the representativeness. Interestingly, the approximation $\mu_r^{num}\text{-}miss$ for Yemeni places of YAGO is very close to what we obtain in a supervised context (where we know that $\alpha \to 0$) – even though the measure is non-conform for that case. In the supervised context, we calculate that 181 facts are missing, while our estimation tells us that 127 facts are missing. Whatever the KB, our estimation of representativeness confirms our intuition mentioned in the introduction: the population of Yemeni places is less well informed than that of French ones.

	Missing facts			Representativeness		
Measure	DBpedia	Wikidata	YAGO	DBpedia	Wikidata	YAGO
$\mu^{num}_{\texttt{pop}}$ in World	15,789	13,720	44,223	0.954	0.961	0.895
$\mu^{num}_{\texttt{pop}}$ in France	1,153	1,546	18,829	0.970	0.963	0.918
$\mu^{num}_{\texttt{pop}}$ in Yemen	78	4,281	127 (NC)	0.829	0.888	0.577 (NC)
$\mu^{count}_{\texttt{starring}}$	51,179	10,370	2,703	0.892	0.989	0.979
$\mu^{count}_{\texttt{team}}$	41,484	3,373	463	0.980	0.997	0.999
$\mu^{count}_{\texttt{birthPlace}}$	38,664	25,691	470	0.971	0.986	0.998

Table 1. Representativeness of relations in three KBs (unsupervised context)

Counting transformation. For this transformation, the estimation of the number of missing facts is more complicated, because the addition of a fact for an object can change its first significant digit. For instance, if a number starting with 5 is missing, an object with 5 facts has to be added. One can imagine to add 5 new facts for a new object, to add four new facts for an object that has already 1 fact, to add 3 facts for an object that has already 2 facts, etc. We choose the solution that minimizes the total number of added facts:

Theorem 2. *Given a knowledge base $\mathcal{K}$ and a measure μ_r^{count} such that $\mu_r^{count}(\mathcal{K}^*)$ satisfies a generalized Bendford's law with exponent α, the number of missing facts for the relation r is:*

$$\mu_r^{count}\text{-}miss(\mathcal{K}) = \sum_{d=1}^{9} ((B_d^\alpha \times m) - f_d) \times d$$

where $m = \max_{d\in[1..9]} \frac{\sum_{i\geq d} f_i}{\sum_{i\geq d} B_i^\alpha}$ *and* $(f_1, \ldots, f_9) = \mu_r(\mathcal{K})$.

This follows from the fact that $\sum_{i\geq d} f_i/m \leq \sum_{i\geq d} B_i^\alpha$ for each digit d. For the unsupervised context, Table 1 indicates the number of missing facts estimated for the relations `starring`/ `team`/ `birthPlace` with our method and deduces an approximation of the representativeness.

Note that for the same relation r, under the two transformations leading to μ_r^{num} and μ_r^{count}, the number of missing facts is bounded by the maximum result: $\max\{\mu_r^{num}\text{-}miss(\mathcal{K}); \mu_r^{count}\text{-}miss(\mathcal{K})\} \leq miss(\mathcal{K})$. Under the same transformation, the missing facts for two distinct relations r_1 and r_2 can be added together: $(\mu_{r_1}\text{-}miss(\mathcal{K}) + \mu_{r_2}\text{-}miss(\mathcal{K})) \leq miss(\mathcal{K})$. We will use these properties in Section 5.3 for DBpedia analysis.

4.5 Limitations of our approach

Using Theorems 1 and 2, our approach approximates the representativeness of some relation r in the knowledge base $\mathcal{K}$ by finding a lower bound $\mu_r\text{-}miss(\mathcal{K})$ such that $\mu_r\text{-}miss(\mathcal{K}) \leq miss(\mathcal{K}_{|r})$ as requested in Section 3.2. This approach

works only if Assumption 1 (Correctness) holds. For the unsupervised setting, we also need Assumption 2 (Transferability).

Furthermore, for the GBL to be applicable, the set of numbers N_r has to meet the following two conditions. First, the numbers of N_r have to be distributed across several orders of magnitude: $\log_{10} \max(N_r) - \log_{10} \min(N_r) \geq 1$. For instance, the height of people does not meet this criterion because it is between 100 and 199 centimeters for most people. In that case, a numerical transformation would lead to a lot of "1" and "2" as first significant digits. For the same reason, it is also not possible to apply the counting transformation to an inverse functional relation r because in that case, each object has only one subject (i.e., $N_r^{count} = \{1, 1, 1, \dots\}$) and then, its prevalence is 0. Second, the cardinality of N_r has to be sufficiently high: $|N_r| \gg 1$. If we do not have enough numbers in N_r, the derived distributions $\mu_r(\mathcal{K})$ will not be reliable enough to learn the parameter α. The next section will show where our method can be applied.

5 Experiments

These experiments answer the following three questions: Is the unsupervised method reliable? Is the representativeness estimated by our method correct? Is the GBL sufficiently effective to be useful for auditing a knowledge base?

All experimental data (the queries, the distributions, the experimental results, and details of the learning method), as well as the source code, are available here: `http://www.info.univ-tours.fr/~soulet/prototype/iswc18`.

5.1 Verification of the transferability assumption

Assumption 2 (Transferability) is a central assumption in the unsupervised approach for learning the GBL parameter. Our first experiment aims to verify if this assumption is true. For this, we compare the parameter α that we obtained by the unsupervised approach to the parameter α of the real world. We found seven relations under the numerical transformation that are known to verify Benford's law in the real world, and that exist in DBpedia and Wikidata. We also found one relation under the counting transformation that exists in our KBs and that is known to follow the GBL in the real world: the out-degree of Wikipedia pages, where $\alpha = -1/2.6 = -0.385$ [5].
Table 2 shows the results obtained for representativeness by Theorem 1 in both supervised and unsupervised contexts. The last column indicates the GBL compliance between the supervised and unsupervised case according to the MAD test (Section 4.3). We see that the learned parameter conforms to the ground truth in all cases: it is very close to zero and does not deviate to values that would have a distorting impact (e.g., $\alpha > 2$, or $\alpha > 5$). For the out-degree of Wikipedia pages, the learned parameter also corresponds well to the real parameter. In addition, the estimator of MAD always indicates a very good conformity (≤ 0.012). This entails that the representativeness that we compute in the unsupervised approach is very similar to the supervised value. In all cases except one,

		Sup.		Unsup.		
Relation	KB	α^*	Rep.	α	Rep.	$MAD(B^{\alpha}, B^{\alpha^*})$
Population of places	DBpedia	0.001	0.949	-0.020	0.954	C
Elevation of places	DBpedia	0.001	0.750	-0.083	0.765	C
Area of places	DBpedia	0.001	0.535	0.143	0.624	AC
Length of water streams	DBpedia	0.001	0.887	0.001	0.887	C
Discharge of water streams	DBpedia	0.001	0.938	-0.105	0.930	AC
Number of deaths	Wikidata	0.001	0.909	-0.106	0.908	AC
Number of injured	Wikidata	0.001	0.883	-0.119	0.875	AC
Out-degree of Wikipedia page	DBpedia	-0.385	0.999	-0.486	0.999	AC

Table 2. Conformity of the unsupervised method with the supervised one

there is less than 1% difference. Even for the least correct prediction (`areaTotal`) the difference is at most 10%[3].

Finally, we also applied the unsupervised method to numerical relations whose numbers should *not* verify the GBL. In such a situation, the method should have a MAD test that indicates a nonconformity (i.e. > 0.015). This is indeed the case for the following relations: Wikipedia page ID (with MAD 0.029), runtime of films (0.077) or albums (0.090), and weight of persons (0.070).

5.2 Validity of representativeness

In Section 3, we postulated that representativeness is an upper bound for completeness. To test this postulation, we simulate an unrepresentative KB as a sample of a known KB. For this purpose, we use the number of inhabitants of French cities from DBpedia as gold standard, because we know that these numbers verify the GBL. We then apply three approaches to degrade this KB:

- **Most-populated:** We removes cities, starting from the least populated to the most populated. This biased sample simulates a KB of Yemeni cities, where only the most populated cities are present.
- **Least-populated:** We remove the most populated cities first. This approach is the opposite of the previous one.
- **Random:** We randomly removes cities. The retained sample of facts is therefore uniformly drawn and it is representative of the original KB.

Our first step is to verify whether our samples conform to Benford's law (Section 4.3). This is indeed the case for 100% of samples for the most-populated approach and the random approach, and for 99% of the samples for the least-populated approach. This validates Assumption 2, and makes our approach applicable. Figure 5 plots the representativeness for the three approaches according to the number of preserved cities in a supervised and unsupervised context. We also plot the real completeness of the sample (w.r.t the original KB).

[3] Different from α, the representativeness varies only between 0 and 1.

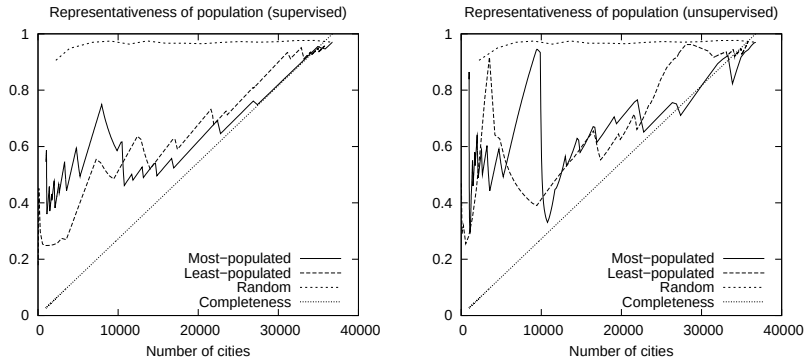


Fig. 5. Impact of incompleteness on French cities using `dbo:populationTotal`

We observe that whatever the approach and the context, representativeness is indeed an upper bound for completeness, as postulated. There is only a single major violation at the point of around 34,000 cities for the most-populated approach, which is due to a wrong approximation of the parameter α in that particular sample. Surprisingly, the representativeness is a very good approximation of completeness for the most-populated and the least-populated approaches. In the case of the supervised context, considering a sample $\mathcal{C} = \mathcal{K}_{|\texttt{pop}}$ with more than 22,000 cities, the estimated number of cities (i.e., $P = |\mathcal{C} + \mu^{num}_{\texttt{pop}}\text{-}miss(\mathcal{C})|$) approximates the true number of cities in $\mathcal{K}^*$ (i.e., $T = |\mathcal{K}^*_{|\texttt{pop}}|$) with less than 5% error: $|P - T|/P \leq 0.05$.

Finally, we observe that as long as the number of cities remains large enough (i.e., greater than 2,500), the representativeness of the random approach is high (around 0.95). This is expected for any large random sample from a complete relation, because a random sample has to be representative in our sense.

5.3 Effectiveness of the GBL for a KB

We considered in DBpedia (France) all the relations with at least 100 facts. We applied the numerical transformation and the counting transformation. We removed all relations whose numbers are not distributed across several orders of magnitude i.e., $\log_{10} \max(N_r) - \log_{10} \min(N_r) < 1$. Table 3 gives a general overview of the resulting 2,920 relations: the number of considered relations, the number of compliant relations (i.e., with $MAD \leq 0.015$), the number of facts, the proportion of facts in DBpedia, the estimated number of missing facts and finally, the estimated representativeness. Clearly, the counting transformation concerns more relations and facts than the numerical transformation. All in all, our analysis covers about 63% of the facts in DBpedia and we estimate its representativeness at 0.719. To make DBpedia's current relations representative, at least 46 million facts would have to be added.

Trans.	# of rel.	# of comp. rel.	# of facts	% of DBpedia	Missing facts	Rep.
Counting	2,920	1,461	117,349,802	0.633	45,869,202	0.719
Numerical	108	43	329,853	0.002	109,603	0.751
Total	2,920	1,487	117,461,855	0.634	45,972,923	0.719

Table 3. Overview of the representativeness of DBpedia (France)

6 Conclusion

In this paper, we have introduced the first method to analyze how representative a knowledge base is for the real world. We believe that representativeness is a dimension of data quality in its own right (along with correctness and completeness), because it is essential for applying statistical or machine learning methods. Our approach quantifies a minimum number of facts that must complement the knowledge base in order to make it representative. Experiments on DBpedia validate our proposal in a supervised and unsupervised context on several relations. Using our method, we estimate that at least 46 million facts are missing for DBpedia to be a representative knowledge base. In future work, we would like to take into account representativeness to correct the result of queries on knowledge bases much like this has been done recently for completeness [10].

References

1. Alam, M., Buzmakov, A., Codocedo, V., Napoli, A.: Mining Definitions from RDF Annotations Using Formal Concept Analysis. In: IJCAI (2015)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)
3. Beiglou, P.H.B., Gibbs, C., Rivers, L., Adhikari, U., Mitchell, J.: Applicability of benfords law to compliance assessment of self-reported wastewater treatment plant discharge data. Journal of Environmental Assessment Policy and Management p. 1750017 (2017)
4. Benford, F.: The law of anomalous numbers. Proceedings of the American philosophical society pp. 551–572 (1938)
5. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer networks 33(1-6), 309–320 (2000)
6. Callahan, E.S., Herring, S.C.: Cultural bias in wikipedia content on famous persons. Journal of the Association for Information Science and Technology 62(10), 1899–1915 (2011)
7. de la Croix, D., Licandro, O.: The longevity of famous people from hammurabi to einstein. Journal of Economic Growth 20(3) (Sep 2015)
8. Darari, F., Razniewski, S., Prasojo, R.E., Nutt, W.: Enabling fine-grained RDF data completeness assessment. In: ICWE. pp. 170–187. Springer (2016)
9. Efron, B.: The jackknife, the bootstrap, and other resampling plans, vol. 38. Siam (1982)

10. Galárraga, L., Hose, K., Razniewski, S.: Enabling completeness-aware querying in SPARQL. In: Proceedings of the 20th International Workshop on the Web and Databases. pp. 19–22. ACM (2017)
11. Galárraga, L., Razniewski, S., Amarilli, A., Suchanek, F.M.: Predicting completeness in knowledge bases. In: WSDM. pp. 375–383. ACM (2017)
12. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.M.: Fast rule mining in ontological knowledge bases with AMIE++. The VLDB Journal 24(6), 707–730 (2015)
13. Giles, J.: Internet encyclopaedias go head to head (2005)
14. Hellmann, S., Lehmann, J., Auer, S.: Learning of OWL class descriptions on very large knowledge bases. Int. J. Semantic Web Inf. Syst. 5 (04 2009)
15. Hürlimann, W.: A first digit theorem for powers of perfect powers. Communications in Mathematics and Applications 5(3), 91–99 (2014)
16. Hürlimann, W.: Benfords law in scientific research. Int J Sci Eng Res 6(7), 143–148 (2015)
17. Lajus, J., Suchanek, F.M.: Are All People Married? Determining Obligatory Attributes in Knowledge Bases . In: WWW (2018)
18. Levy, A.Y.: Obtaining complete answers from incomplete databases. In: VLDB (1996)
19. Mebane Jr, W.R.: Election forensics: Vote counts and benfords law. In: Summer Meeting of the Political Methodology Society, UC-Davis, July. pp. 20–22 (2006)
20. Morzy, M., Kajdanowicz, T., Szymański, B.K.: Benfords distribution in complex networks. Scientific reports 6, 34917 (2016)
21. Motro, A.: Integrity = Validity + Completeness. TODS (1989)
22. Nigrini, M.: Benford's Law: Applications for forensic accounting, auditing, and fraud detection, vol. 586. John Wiley & Sons (2012)
23. Nigrini, M.J.: A taxpayer compliance application of benford's law. The Journal of the American Taxation Association 18(1), 72 (1996)
24. Razniewski, S., Korn, F., Nutt, W., Srivastava, D.: Identifying the extent of completeness of query answers over partially complete databases. In: SIGMOD (2015)
25. Razniewski, S., Suchanek, F., Nutt, W.: But what do we actually know? In: Proceedings of the 5th Workshop on Automated Knowledge Base Construction. pp. 40–44 (2016)
26. Rebele, T., Nekoei, A., Suchanek, F.M.: Using YAGO for the Humanities . In: WHISE workshop (2017)
27. Schich, M., Song, C., Ahn, Y.Y., Mirsky, A., Martino, M., Barabási, A.L., Helbing, D.: A network framework of cultural history. Science 345(6196) (2014)
28. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: WWW. pp. 697–706. ACM (2007)
29. Suchanek, F.M., Preda, N.: Semantic culturomics. Proceedings of the VLDB Endowment 7(12), 1215–1218 (2014)
30. Tanon, T.P., Stepanova, D., Razniewski, S., Mirza, P., Weikum, G.: Completeness-aware rule learning from knowledge graphs. In: ISWC. pp. 507–525. Springer (2017)
31. Völker, J., Niepert, M.: Statistical schema induction. In: ESWC (2011)
32. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
33. Wagner, C., Garcia, D., Jadidi, M., Strohmaier, M.: It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In: ICWSM. pp. 454–463 (2015)
34. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web 7(1), 63–93 (2016)