Class Cardinality Comparison as a Fermi Problem

Shrestha Ghosh
Max Planck Institute for Informatics
Saarland University
Saarbruecken, Germany
ghoshs@mpi-inf.mpg.de

Simon Razniewski
Max Planck Institute for Informatics
Saarbruecken, Germany
srazniew@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbruecken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Questions on class cardinality comparisons are quite tricky to answer and come with its own challenges. They require some kind of reasoning since web documents and knowledge bases, indispensable sources of information, rarely store direct answers to questions, such as, "Are there more astronauts or Physics Nobel Laureates?" We tackle questions on class cardinality comparison by tapping into three sources for absolute cardinalities as well as the cardinalities of orthogonal subgroups of the classes. We propose novel techniques for aggregating signals with partial coverage for more reliable estimates and evaluate them on a dataset of 4005 class pairs, achieving an accuracy of 83.7%.

KEYWORDS

Entity classes, Comparative questions, Class cardinality estimation

1 INTRODUCTION

Motivation and Problem. Are there more astronauts or Physics Nobel laureates? More nuclear power plants or catholic cathedrals? More lakes or rivers? More airports on this planet than satellites in orbit, or vice versa? Comparative questions of this kind tickle our curiosity, yet are often surprisingly hard to answer. For some comparisons, there are authoritative official sources that provide reliable (albeit not necessarily up-to-date) numbers. In most cases, though, the natural resort is to tap into online sources like knowledge bases (KBs, e.g., Wikidata), search engine (SE, e.g., Bing) results, or large language models (LMs, e.g., GPT). However, all of these come with biases in what they cover and what not, and often give treacherous signals that lead humans to wrong conclusions.

For example, while we found the number of rivers to be close to 300,000 [9], the number of castles is quite inaccessible and lies for Europe alone between 400,000 to 1.3 million[1].

---

[1] https://quora.com/How-many-castles-are-there-in-Europe/answer/Michael-Burke-339; https://dw.com/does-germany-really-have-25000-castles/a-42350502

Wikidata suggests that there are more than 400,000 rivers and roughly 23,000 castles. Major search engines, when probed with different query formulations, pick up on the number of rivers as 250,000 (in the United States) and 10,000 (medieval European) castles, suggesting that there are far more rivers than castles, contradicting reality. Even the GPT-3 language model, which performs well on question answering tasks, is completely off. When prompted with three example questions and their answers, followed by the question asking for the number of rivers, it returns *"an estimated 1.3 million rivers on Earth"*, and for castles it returns, *"There are around 900 castles in the world."* Obviously, neither of these major sources is a true mirror of reality. The online world is inherently hampered by incompleteness and bias.

Smart humans, on the other hand, are sometimes able to judiciously select online sources as cues, then combine multiple cues in a clever way, and eventually arrive at reasonable estimates at class cardinalities (at least the right order of magnitude) and relative comparisons between classes (just asserting which is bigger). Enrico Fermi, a Physics Nobel Laureate from the early 20th century, was known to be a master of such estimates; hence this kind of problem is also known as Fermi Problem [26].

In this paper, we aim to emulate a smart human's approach. We introduce and study a variety of online signals that could be applied, gaining insights on strengths and weaknesses for different domains of entity classes (e.g., occupations vs. creative works vs. organizations). Moreover, we propose novel techniques for aggregating signals with partial coverage into more reliable estimates on which of the two given classes has more real-world instances.

Approach and Contribution. This paper focuses on *dominance estimation*: which of two classes has the higher cardinality. We obtain cues for the numeric cardinalities from three sources: the Wikidata KB via SPARQL queries [24], the Bing search engine with judicious queries using the CoQEx method [7], and the GPT-3 language model [2] with various prompts. Absolute cardinalities from these sources are often completely wrong; so we interpret them merely as signals to be used for further inference.

The key idea of mitigating these bias effects is to additionally inspect *subgroups* of classes, such as actors or airports by country or geo-region (e.g., North America, East Asia etc.). Such subgroups are orthogonal to the classes under comparison. The hypothesis that we study is that the estimates for subgroups of classes can give more reliable cardinality estimates, at least for some subgroups and for relative comparison. This larger set of finer-grained signals are then aggregated using different techniques proposed in this paper.

The evaluation dataset consists of 4005 class pairs from 6 diverse domains. Our major finding is that the novel technique of aggregating signals substantially improves the dominance estimation, achieving over 80% accuracy compared to direct source signals.

2 RELATED WORK

We focus on bias in the digital world and completeness in information sources, which would most affect dominance estimation.

Bias in the Digital World. Wikipedia, a major source of general knowledge, used in automatic construction of KBs, suffers from implicit and explicit bias [10] and specific demographic biases such as gender bias [22]. The common crawl [3] is another massive dataset of textual information, almost 50 times greater than Wikipedia (around 5.6 TB) and the BookCorpus (6 GB). It requires significant processing before it can be used for pre-training large LMs [18]. A case study shows that the filtered C4 dataset disproportionately affects minorities [4]. Another work addresses the different biases on the web itself and its potential effects [1]. Representation in KBs is unbalanced due to multiple reasons, including but not limited to reporting bias [8], data, schema and inferential bias [11, 21].

Sources of Information. It is well known that general-knowledge KBs are incomplete [20, 25] even with their increased coverage over time [19]. Species estimation techniques from biology have been used to estimate cardinalities [15, 23]. Techniques like mark and recapture assume sampling from the real-world and, applying them to KB edit history makes severe underestimations, for instance predicting that there are roughly 4M humans [15][2]. Current powerful SEs are now capable of providing structured answers from their internal KBs, and highlighting the most probable answer in the top snippet. Even so, structured answers are not the norm for more complex or less popular questions. This low recall has prompted research on answering entity counts from multiple SE snippets [6, 7] and large-scale mining of quantities from the web [5]. LMs have been shown to be effective in recalling factual information [13, 17]. Nevertheless, they lack scrutability, rely on high quality prompts, [12], and are known to struggle with numeric/count information [14].

3 CLASS CARDINALITY COMPARISON

Let us continue with the question: *Are there more rivers than castles?* In principle, such questions can be decomposed into cardinality questions: *how many rivers are there?* and *how many castles are there?*, the answers of which are then compared. We therefore identify three problem statements related to comparison questions, of inversely related informativeness and difficulty.

1. *Cardinality Estimation*: What are the absolute cardinalities of the classes *rivers* and *castles*?
2. *Proportionality Estimation*: What is their ratio?
3. *Dominance Estimation*: Are there more *rivers* or *castles*?

While cardinality estimation provides the most information, this is also the most challenging task. Estimation methods may give impractical results that are orders of magnitude away from the ground-truth, and for some classes, there might not even exist a widely agreed ground-truth.

Looking at the relation of pairs of classes reduces the impact of uncertainty: Even if there is no widely agreed count for the number of castles, most estimates might agree that there are more castles than rivers. Determining the actual proportion would be desirable, though it is also subject to uncertainty.

In this paper, we focus on the most approachable task of *dominance estimation*: to determine whether one class is bigger than the other. Intuitively, as humans, we can deduce that there are more castles than rivers from our observation that along a river, there are typically several castles. However, *machines are incapable of such reasoning*. Nevertheless, there exists evidence in the form of KB entities and actual counts in web documents, which can be leveraged to predict the bigger of the two classes. We identify cardinality signals to predict whether a class A is greater than B through the output variable $O_{|A|>|B|}$, defined as:

$$O_{|A|>|B|} = \begin{cases} 1 & \text{if } |A| \text{ is predicted to be bigger than } |B|, \\ -1 & \text{if } |A| \text{ is predicted to be smaller than } |B|, \\ 0 & \text{if the predictor abstains.} \end{cases} \quad (1)$$

3.1 Basic Cardinality Signals

We obtain cardinality signals from three sources, and of two types.

Signal sources. We use three different signal sources: KBs, SEs, and LMs. Each provides a different angle: of entities covered in a KB, of what is popular on the web, and of what information has been distilled by LMs. Specifically we look into Wikidata [24], the top-50 search results by Bing [16] and GPT-3 [2].

1. **Knowledge base (KB).** Formulate SPARQL queries to retrieve the count of entities per class from Wikidata, preferring hand-annotated queries over KB-QA systems. KB-QA systems perform worse than hand-annotated queries due to the available system's inability to construct the most accurate SPARQL queries.
2. **Search engine (SE).** Return the most confident cardinality of a class using the CoQEx system for inferring counts from top-50 SE result snippets [6].
3. **Language model (LM).** In a few-shot setting, given n cardinality questions and their answers as a prompt, retrieve the cardinality of the class in the $(n + 1)^{th}$ question. Specifically, we provide three examples along with the intended question to the GPT-3 model. The examples and the model parameters remain constant for all queries. We process the text output to extract the class cardinality using Python Quantulum3 library[3].

Signal types. We consider two types of signals.

1. **Root signals.** Signals for the count of the class of interest itself, e.g., *rivers* (worldwide).
2. **Subgroup signals.** Signals for counts of subgroups orthogonal to the classes being compared, e.g., *castles in Germany*.

Unlike subclasses, which divide a class based on a type hierarchy and are by design class-specific, we use subgroups which are applicable across classes. In principle, subgroups signals can be computed for a range of subgroups, e.g., by 195 countries, by year/decade, by status, etc. Thus, subgroups provide a dimension for comparing two classes from the same or different domains. If we restrict the domains, we could also focus on specific subgroups, such as subgroups by gender for occupations. In the following we focus on the G20-group of countries [27], as for these, sources tend to have more reliable information.

Note also that subgroup signals normally *do not sum up* to root signals, both for pragmatic reasons (data for certain subgroups is

---

[2]https://cardinal.exascale.info/

[3]https://pypi.org/project/quantulum3/

unavailable/incomplete), as well as principled reasons (an entity belonging to several subgroups, or to none).

3.2 Signal Aggregation

Besides using the root signals directly to predict class comparison, we aggregate basic signals in three levels: First we aggregate subgroup signals, after which we include root signals, and then aggregate the resulting signals by sources.

I. Subgroup aggregation.

1. By majority: $O^M_{|A|>|B|}$ is 1, if at least θ_M percent of the subgroups of A have more entities than the those of B, -1 if at least θ_M have less, else 0.

2. By significance: We perform a one-sided t-test, testing for A bigger than B when the mean of the subgroup cardinalities of A is greater than that of B. We test for A less than B when the reverse is true. In the former case, if the p-value $\leq \alpha$, the subgroup distribution of A is significantly greater than B and $O^S_{|A|>|B|}$ is 1. If the p-value $\leq \alpha$ in the latter case, then the reverse is true and $O^S_{|A|>|B|}$ is -1. If the p-value $> \alpha$ in either of the cases, the prediction is 0.

II. Root and subgroup aggregation. Here, the final prediction is the average over the predictions obtained from i) comparing root signals, and ii) majority vote over subgroup signals, and iii) significance test over subgroup signals. We use weights, $W \in [0, 1]$ for the majority and significance predictions, such that,

$$Ensemble^{source}_{|A|>|B|} = \frac{1}{3}(O^{source}_{|A|>|B|} + W_M \cdot O^{source,M}_{|A|>|B|} + W_S \cdot O^{source,S}_{|A|>|B|}) \quad (2)$$

Hereby, the weight W_M is the majority ratio, when majority is $> \theta_1$, i.e. A bigger than B, and, $W_M = (1 - majority\ ratio)$ when B is greater. The significance prediction is weighted by $(1 - p\text{-value})$, such that the lower the p-value, higher the significance.

III. Source aggregation. In the last level of aggregation, we combine the predictions from all sources by:
Majority vote: if any two sources agree, that label is selected, else the predictor abstains.
Weighted vote: we train a Logistic Regression classifier for each of the signal aggregation cases, to learn the weights of each source.

$$ln(p(|A| > |B|)) = -ln(1 + e^{-\sum_{source} W_{source} \cdot Ensemble^{source}_{|A|>|B|}}) \quad (3)$$

$$Ensemble^{weighted}_{|A|>|B|} = \begin{cases} 1 & \text{if } p(|A|>|B|) > p(|A|<|B|), \\ -1 & \text{if } p(|A|>|B|) < p(|A|<|B|), \\ 0 & \text{if } p(|A|>|B|) = p(|A|<|B|) = 0.5. \end{cases} \quad (4)$$

4 EXPERIMENT

Dataset. We create a ground-truth (GT) dataset of 90 classes spanning 6 domains (Table 1)[4]. Our evaluation set comprises $\binom{90}{2} = 4005$ combinations of class pairs, of which $6 \times \binom{15}{2} = 630$ are in-domain pairs, i.e., both classes belong to the same domain and the remaining 3375 class pairs are inter-domain, i.e., both classes belong to different domains. In order to assess the difficulty of the task, we compute the order of magnitude of the ratio of the ground-truth cardinalities of all class pairs. We argue that class

[4]Dataset link: https://github.com/ghoshs/class_cardinality_comparison.

Table 1: Domains and example classes.

Domain	#Classes	Examples
creative work	15	film, board game, book
geographical entities	15	lake, castle, dam
man-made object	15	satellite, submarine
occupation	15	politician, actor, physicist
organization	15	university, football club
species	15	snake, insect, fish

pairs with close cardinalities would be more difficult to predict than class pairs whose cardinalities differ by several orders of magnitude. The dataset has more pairs with close cardinalities, and less than 8% of the class pairs have cardinality ratios more than 10^4 orders of magnitude. For instance, airlines (5,000) and national parks (3,369) has a cardinality ratio of 1.4 : 1 (higher:lower) while airlines to politicians (6,500,000) has a cardinality ratio of 1.3×10^3 : 1.

Metrics. We primarily measure the performance of signals by *accuracy*, i.e., the percentage of correct predictions relative to all samples. To analyse if low accuracies stem from abstentions or wrong predictions, we also report the *rate of abstention*, and, *precision*, as the percentage of correct predictions relative to non-abstentions.

Parameters. We use the *text-curie-001* model of GPT-3 with temperature set to 0 and maximum tokens to generate set to 15. The subgroup aggregation parameters are θ_M, which we set to 0.5 (more than 50% majority) and α, which we set to 0.05. We train the Logistic Regression classifiers, by dividing the dataset into train and test splits (80:20). We perform a 5-fold cross-validation on the training data to determine the regularization hyperparameter. We evaluate on the whole dataset of 4005 class pairs, except for the weighted vote ensemble, which is evaluated on the test set of 800 pairs.

State-of-the-art Baselines. They are the root signals obtained from the three sources: KB, SE, LM.

Human Baseline. We sampled 100 class pairs, 50 in-domain and 50 inter-domain, to evaluate human performance, i.e., general non-expert users. Each class pair was annotated by three MTurkers, who were given brief descriptions of the classes, and were asked two multiple-choice questions. i) Which class has more entities in real life? Options: *Class 1, Class 2, Equal, Cannot determine*, and ii) How certain are you? Options: *Very sure, Estimation, Guess, No idea*.

In addition, annotators could justify their responses in a free text field. The task had two settings (i) the closed-book setting, where annotators should answer without consulting external sources, and (ii) the open-book setting, where the annotators were encouraged to perform web research. For the closed-book settings, annotators had 3 minutes per question, for the open-book setting, 8 minutes.

The accuracy in both settings was comparable (75% vs. 76%), although the precision increased substantially more (from 79% to 85%), i.e., in the open-book setting, the additional evidence made annotators more often abstain, instead of guessing.

Result 1: Performance by Source. Table 2 shows the accuracy of the sources by their signals. Of the three sources, only LM surpasses the human baseline. We find that an ensemble over the root and the aggregated subgroup signals performs well for all three sources. Aggregating over subgroup and root signals reduces the rate of abstention from 0.5% to 5%, down to less than 1% consistently across sources. Upon further inspection we find that the t-test subgroup

Table 2: Accuracy (in %) of cardinality signals. Baselines are `highlighted`.

Source	Root (1)	Subgroup Aggregations		Root and Subgroup Signal Ensembles		
		Majority (2)	T-test (3)	(1)+(2)	(1)+(3)	(1)+(2)+(3)
KB	64.7	57.0	36.0	61.5	**65.9**	61.8
SE	65.4	67.2	39.6	65.7	65.4	**68.4**
LM	74.4	75.8	57.3	77.1	75.8	**79.4**
Ensemble over Sources (KB, SE, LM)						
Majority Vote	77.8	76.3	42.9	76.8	78.7	**78.9**
Weighted Vote	78.2	76.2	81.2	79.3	**83.7**	81.3
Non-expert human baseline						
Closed-book	75.0					
Open-book	76.0					

Table 3: Accuracy (in %) of aggregation over root and subgroups by domain.

Domain	KB	SE	LM	Best
Creative work	62.8	54.2	**83.8**	LM
Geographical entity	**77.1**	60.9	70.4	KB
Man-made object	26.6	77.1	**96.1**	LM
Occupation	57.1	**80.0**	74.2	SE
Organization	58.0	72.3	**88.5**	LM
Species	61.9	**78.0**	63.8	SE
Interdomain	62.6	68.0	**79.4**	LM
All	61.8	68.4	**79.4**	LM

Table 4: Examples from our dataset, ordered from easier to harder.

Class 1	Class 2	GT ratio	Root signals			Basic signal ensembles*			Source ensembles**		Comment
			KB	SE	LM	KB	SE	LM	Majority	Weighted	
Websites	Religious texts	$6 \times 10^7 : 1$	✓	✓	✓	✓	✓	✓	✓	✓	Strong signals from all sources.
Bacteria species	Bee species	$3.15 \times 10^5 : 1$	✓	x	x	✓	x	x	x	x	KB root signal ratio of 4.8:1 gives correct prediction.
Books	Paintings	$1.3 \times 10^3 : 1$	x	x	✓	x	x	✓	x	✓	Weighted vote picks up weak signal from LM.
School teachers	Hospitals	$6.07 \times 10^2 : 1$	x	x	x	✓	✓	✓	✓	✓	Correct signals from subgroup agg. in SE and LM.
Cities	Islands	9.7 : 1	x	✓	x	x	✓	✓	✓	x	Weighted vote fails due to strong incorrect KB + weak correct LM signals.
Actors	Architects	1.48 : 1	✓	x	x	✓	x	x	x	✓	Weighted vote leverages strong KB signal.

* *Best ensemble for KB: (1)+(3); SE and LM: (1)+(2)+(3) from basic signals: Root (1), Majority agg. (2) and T-test agg. (3).*

** *Best source ensemble for majority vote: (1)+(2)+(3); weighted vote: (1)+(3).*

aggregations have the lowest accuracy, despite good precision (LM: 87%, SE: 83%), due to very high abstention rates (LM:34%, SE: 52%).

Result 2: Aggregation over Sources. From Table 2, we see that learning weights using supervised learning increases robustness across aggregation strategies, as the accuracy does not drop below 76%, hence performing better than any individual source. On the contrary, majority vote over sources is most effective only on root signals. *Notably*, in the weighted vote strategy, subgroup aggregations by t-test does remarkably well achieving 81% when aggregating just the t-test signals and 83.7% when aggregating the root and t-test signals. In both cases, we noticed high coefficient for LM, followed by lower coefficients for SE and then KB.

Result 3: Performance by Domain. We analyse the class pairs by each of the 6 domains on the ensemble over root and subgroup signals in Table 3. Notice that there is no one winning source and relying on the best source per domain gives an accuracy above 77%. Estimates of everyday objects such as bicycles or smartphones, are easily available on the web, as reflected in SE (77%) and LM (96%) accuracies. KB is most accurate on geographical entities (77.14%), outperforming the other two sources by a large margin.

5 DISCUSSION AND CONCLUSION

Direct comparison questions are more challenging for current SEs and LMs. For instance, Bing returns irrelevant answers to the question *are there more rivers or lakes?*, while GPT-3 returns a definitive answer that there are more lakes. When prompted for an explanation, the answer is reversed. In Table 4 we highlight a few examples from our dataset and the predictions of root signals and the best performing ensembles from Table 2.

In this work we tackle dominance estimation of two classes in the real-world by using signals from varied information sources. We propose techniques to combine cardinality signals by aggregating over orthogonal subgroups and over multiple sources. From experiments, we observe ensembles over sources using subgroup aggregations to perform very well in dominance estimations.

REFERENCES

[1] Ricardo Baeza-Yates. 2018. Bias on the web. In *Communications of the ACM*.
[2] Tom Brown et al. 2020. Language models are few-shot learners. In *NeurIPS*.
[3] Common Crawl. 2008. Common crawl dataset. https://commoncrawl.org/.
[4] Jesse Dodge et al. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *EMNLP*.
[5] Yanai Elazar et al. 2019. How Large Are Lions? Inducing Distributions over Quantitative Attributes. In *ACL*.
[6] Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. 2022. Answering Count Queries with Explanatory Evidence. In *SIGIR*.
[7] Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. 2022. Answering Count Questions with Structured Answers from Text. In *JoWS*.
[8] Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *AKBC*.
[9] Günther Grill et al. 2019. Mapping the world's free-flowing rivers. In *Nature*.
[10] Christoph Hube. 2017. Bias in Wikipedia. In *WWW Companion*.
[11] Krzysztof Janowicz et al. 2018. Debiasing Knowledge Graphs: Why Female Presidents are not like Female Popes. In *ISWC Workshops*.
[12] Zhengbao Jiang, Frank F. Xu, J. Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know?. In *Transactions of the ACL*.
[13] Vladimir Karpukhin et al. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP*.
[14] Bill Yuchen Lin et al. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *EMNLP*.
[15] Michael Luggen et al. 2019. Non-parametric class completeness estimators for collaborative knowledge graphs—the case of wikidata. In *ISWC*.
[16] Microsoft. 2022. Bing search API. https://www.microsoft.com/en-us/bing/apis/bing-custom-search-api.
[17] Fabio Petroni et al. 2019. Language Models as Knowledge Bases?. In *EMNLP-IJCNLP*.
[18] Colin Raffel et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.. In *JMLR*.
[19] Simon Razniewski and Priyanka Das. 2020. Structured Knowledge: Have we made progress? An extrinsic study of KB coverage over 19 years. In *CIKM*.
[20] Simon Razniewski, Fabian M. Suchanek, and Werner Nutt. 2016. But What Do We Actually Know?. In *AKBC*.
[21] Tara Safavi et al. 2021. Report on the first workshop on bias in automatic knowledge graph construction at AKBC 2020. In *SIGIR Forum*.
[22] Jiao Sun and Nanyun Peng. 2021. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *ACL-IJCNLP*.
[23] Beth Trushkowsky, Tim Kraska, Michael J. Franklin, and Purnamrita Sarkar. 2013. Crowdsourced enumeration queries. In *ICDE*.
[24] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. In *Communications of the ACM*.
[25] Gerhard Weikum et al. 2021. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. In *Foundations and Trends in Databases*.
[26] Wikipedia. 2022. Fermi Problem. https://en.wikipedia.org/wiki/Fermi_problem.
[27] Wikipedia. 2023. G20 countries. https://en.wikipedia.org/wiki/G20.