===> How Quantifying Knowledge Wealth Reveals Knowledge Gap: A Case Study of Wikidata

Quantifying Knowledge Wealth in Knowledge Graphs: A Case Study of Wikidata

TBD[1], TBD[1], TBD[2] and TBD[3]

[1]TBD
[2]TBD
[3]TBD

Abstract

Along with the rapid development of data volumes, the need for machine-readable data is inevitable. As a result, the use of knowledge graph data structures becomes more popular. With its development, quality aspects of a knowledge graph need to be considered, one of which is knowledge wealth: the amount of information contained in a knowledge graph. A high level of knowledge wealth in a knowledge graph may indicate the high quality of a knowledge graph; conversely, a low level of knowledge wealth can be a sign of poor quality of a knowledge graph. However, there is no formal way to define knowledge wealth and how to measure and analyze it. This study proposes a framework to analyze knowledge wealth and the level of knowledge imbalance in the RDF knowledge graph by seeing how the knowledge wealth of an entity class is spread over the knowledge graph using statistical measures and visualization. To evaluate this framework, some use cases were conducted on several classes on Wikidata to detect bias as well as to explore how different definitions of type of wealth impact the magnitude of the Gini coefficient. It is hoped that the results of this study can assist in researching knowledge wealth in the knowledge graph and be used to optimize the efforts of editing and developing knowledge graph projects by the contributors.

Keywords

Knowledge graphs, knowledge wealth, knowledge imbalance analysis

1. Introduction

Wealth is the abundance of valuable possessions or money, or the state of having this abundance (Oxford English Dictionary). In economics, wealth is defined as the net value of all assets owned by individuals or households, calculated as total assets minus liabilities, and used to assess economic inequality (Saez & Zucman, 2016). When considering individual (human) wealth, the most common measure is net worth. On the other hand, in knowledge graphs (KGs), wealth can be defined as the amount of information (in terms of properties or links) an entity possesses.

Figure 1 shows three Wikidata entities of the type ethnic group: the Romani people, the Minangkabau, and the Ambonese, along with some information about them. For example, the entity Romani people includes information such as its class, flag, native languages, population, and associated Freebase ID. We can observe that a type of information can have a single value,

To be decided

tbd@tbd.com (TBD); tbd@tbd.com (TBD); tbd@tbd.com (TBD); tbd@tbd.com (TBD)

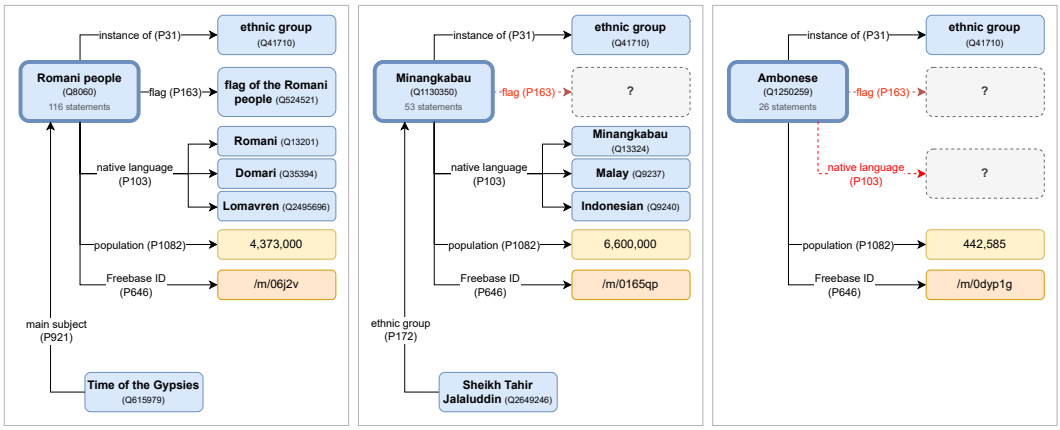


Figure 1: Data on Wikidata about the Romani people, the Minangkabau, and the Ambonese

such as the property *flag* for each entity, which is exactly one (or zero, if the information does not exist), or multiple value, such as the property *native languange*. Both of these properties have values that are other Wikidata entities. Additionally, there are other types of properties, such as *population*, which has a static numerical value, and *Freebase ID*, which provides a string used to identify the entity in the Freebase database.

While the previous examples focus on information with outgoing links, we can also consider the opposite perspective by examining incoming links. This approach does not directly show the possessions of an entity but rather indicates its popularity by showing how often it is mentioned elsewhere. This is illustrated by the Romani people being mentioned in *Time of the Gypsies* and the Minangkabau being associated with Sheikh Tahir Jalaluddin.

The example in Figure 1 provides a clear picture of how entities in Wikidata can have different kinds and amounts of information, which may lead to knowledge imbalance. If this issue is left unaddressed, it can be problematic for anyone utilizing open KGs as a data source. Data users may draw invalid inferences and conclusions based on incomplete or imbalanced data, such as the Minangkabau and Ambonese are less important than the Romani. Moreover, if contributors to open KGs cannot identify which entities or classes are lacking information, efforts such as editathons may not be effective, potentially widening the gap between information-rich and information-poor entities. For example, contributors might prioritize enriching the Romani people's data while overlooking gaps in the Minangkabau entry, leaving the *flag* property in the Minangkabau empty despite the well-documented existence of the Marawa flag of Minangkabau.

Existing approaches often focus on (...), lacking a way of quantifying the amount of information contained in KGs. Our study addresses this gap by proposing a formal model to define the knowledge wealth in the RDF knowledge graph. Specifically, we focus on three key contributions: (i) introducing three notions of quantifying knowledge wealth for knowledge graphs and demonstrating how they can be used to further characterize knowledge wealth; (ii) implementing the formal and insight model using Python and making it accessible for broader use; and (iii) conducting a case study on Wikidata classes, illustrating how biases can be identified, how different definitions of wealth impact the imbalance level of a class, and how

the composition of wealth varies between classes.

2. Related Work

(Pak FD)

Data Completeness Profiling Wisesa et al. (2019) presented ProWD, a framework and web application tool for profiling the completeness of Wikidata. It is used to provide insight on degree of attribute completeness of a class in Wikidata. The visualization provided in the we dashboard is equipped with single, compare, or multidimensional view to help in analyzing the facet at entity or class level.

Imbalance and Gap in Wikidata - Refo: Gini index

- Nio: gap property Ramadizsa1 et al. (2023) introduced the concept of gap properties that helps to characterize class-level knowledge gaps within knowledge graphs. The framework adapts association rule mining to determine ...

3. Knowledge Wealth Analytics Framework

3.1. Wealth Formal Model

Knowledge graphs follow Resource Description Framework (RDF) as a means of data organization. Without loss of generality of how the form of the URIs is, data is stored in the form of triple (s, p, o); a combination of a subject s, a predicate p, and an object o which can be visualized as nodes and directed-arc diagrams. For example, the statement "William Shakespeare's notable work is Romeo and Juliet" in human-readable URIs is mapped to the triple (*WilliamShakespeare*, *notableWork*, *RomeoAndJuliet*). Likewise, the statement in Wikidata, which uses ID-based URIs, is mapped to (Q692, P800, Q83186).

There are 3 kind of nodes: IRIs, literals, and blank nodes. A triple is in the form of $(s, p, o) \in G) \cap (I \cup B) \times I \times (I \cup B \cup L)$ where I is the node with type IRIs, B is the node with type of blank node, and L is the node with type of literals. In this study, we omit the usage of blank node, as this adds complexity to the analysis and may result in quality issue.

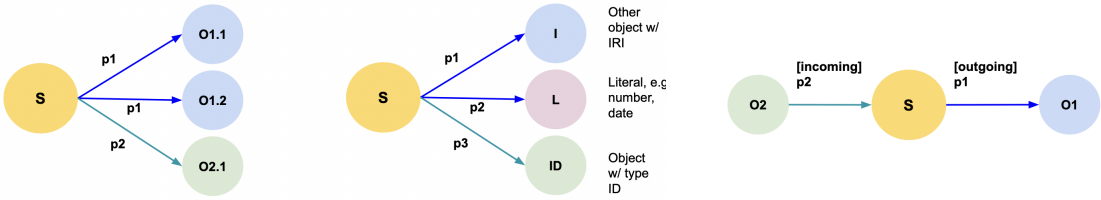
3.1.1. Class

In this study, we re-use the class model defined by Ramadizsa et al. (2023). A class is a group of entities that are the subject of the study. *Human*, *film*, and *taxon* are some examples of class. In general, entity s is an instance of class C is expressed by the triple (s, *instanceOf*, C). We can get a more narrow class inside the defined class by specifying additional conditions, each consisting of a particular property and value associated with it. Example of such conditions for human class is *gender* with associated value *male*, while example for a country would be *continent* with value *Asia*. For instance, the class of human with gender male that lived during English Renaissance is queried using ($?s$, {($?s$, *instanceOf*, *human*), ($?s$, *gender*, *male*), ($?s$, *timePeriod*, *EnglishRenaissance*)}).

3.1.2. Entity-Level Wealth: Knowledge Wealth Type and Definition

Let s be any entity in a knowledge graph G. We quantify the wealth of entity s in G as the amount of information about s available in G. Thus, the knowledge wealth of an entity is defined by the number of properties associated/linked to it. For example, the wealth of William Shakespeare (Q692) in Wikidata counts all triples describing Q692 in Wikidata, including those detailing his family, occupation, image, and so on.

There are several notion on how to calculate the knowledge wealth of an entity: (1) wealth based on the (non-)uniqueness of individual property; (2) wealth based on type of property; and (3) wealth based on the direction of the link. The wealth of s with regard to graph G for each wealth category is denoted by W, formalized and explained as follows.



(a) Illustration of bag of properties and set of properties

(b) Illustration of 3 types of property: object, literal, ID

(c) Illustration of incoming link vs. outgoing link

Figure 2: Three simple graphs

Wealth based on the (non-)uniqueness of individual properties The first measure of the knowledge wealth of s is bag of properties—the cardinality of set of all triples that has s in their subject position. In this definition, the triples (s, p_1, o_1) and (s, p_1, o_2) account for a wealth of 1 each, thus both have a total of 2. Let $N_{bag}(s, G)$ be a set that comprises all pair of predicate/property and object (p, o) that is connected to s. Then $W_{bag}(s, G)$ is the cardinality of $N_{bag}(s, G)$.

$$N_{bag}(s, G) = \{(p, o)|(s, p, o) \in G\}$$
$$W_{bag}(s, G) = |N_{bag}(s)|$$

Another way of measuring the wealth is by counting the number of distinct properties describing the entity, or set of properties. By this way, we are capturing the variety of information about an entity. In contrast to bag of properties, in set of properties (s, p_1, o_1) and (s, p_1, o_2) would be regarded as the "same" information because of the identical property p_1, thus they only account for a total wealth of 1. Let $N_{set}(s, G)$ be a set that comprises all predicate/property p that is connected to s. Then $W_{set}(s, G)$ is the cardinality of $N_{set}(s, G)$.

$$N_{set}(s, G) = \{p|\exists o, (s, p, o) \in G\}$$
$$W_{set}(s, G) = |N_{set}(s, G)|$$

By the above definition, the wealth of entity s in Figure 2a is 3 and 2, using bag of properties and set of properties respectively.

Wealth based on type of property Object properties are other entities besides s that is connected with s through a property p. Wealth of s using bag of properties with only considering the object properties is defined as:

$$W_{bag,object}(s,G) = |\{(p,o)|((s,p,o) \in G) \cap (o \in I)\}|$$

Literal properties are non-object properties that is connected with s through a property p. Wealth of s using bag of properties with only considering the literal properties is defined as:

$$W_{bag,literal}(s,G) = |\{(p,o)|((s,p,o) \in G) \cap (o \in L)\}|$$

An external ID is a special type of string that is used to represent an entity in an external source. In Wikidata, an ID is identifiable by property type *wikibase:ExternalId*. Just like any other property, an ID is connected with s through a property p. Let $C_{ID,G}$ be a set comprising ID property in graph G. Wealth of s using bag of properties with only considering the ID properties is defined as:

$$W_{bag,ID}(s,G) = |\{(p,o)|((s,p,o) \in G) \cap (o \in L) \cap (o \in C_{ID,G})\}|$$

Wealth based on the direction of the link In outgoing link type of wealth, the properties that are used in the wealth calculation of an entity s are those obtained from link with outwards direction from that particular entity s; that is where s appears to be the subject in the set of triples in graph G. All types of wealth defined before use the notion of outgoing link.

In incoming link type of wealth, the properties that are used in the wealth calculation of an entity s are those obtained from link with inwards direction to that particular entity s; that is where s appears to be the object in the set of triples in graph G. To illustrate, let $N_{bag}(s)$ be a set that comprises all pair of object and predicate/property (o,p) that is connected to s in incoming direction to s i.e., $N_{bag}(s) = \{(o,p)|(o,p,s) \in G\}$. Then the wealth of s using bag of properties and the view of incoming link is notated as $W_{bag,incoming}(s,G)$, and equal to the cardinality of $N_{bag}(s)$.

$$N_{bag,incoming}(s,G) = \{(o,p)|(o,p,s) \in G\}$$

$$W_{bag,incoming}(s,G) = |N_{bag,incoming}(s,G)|$$

Looking at in Figure 2c, the wealth of entity s is 1 using outgoing link, which is from the triple $(s,p_1,o1)$. Its wealth is also 1 and using incoming link, which comes from the triple $(o2,p_2,s)$.

Each definition above can be used simultaneously. For example, the wealth of entity s using set of properties, calculating object and data but not ID properties, and using the direction of outgoing link is denoted by $W_{set,outgoing,(object \cup data) \cap ID^{\complement}}(s,G) = |N_{set,outgoing,(object \cup data) \cap ID^{\complement}}(s,G)|$ with $N_{set,outgoing,(object \cup data) \cap ID^{\complement}}(s,G) = \{p|\exists o,(s,p,o) \in G, \cap (o \in ((I \cup L) \cap C^{\complement}_{ID,G})\}$

3.1.3. Class-Level Wealth

Let C be a class that consists of m distinct entities s_1, s_2, ... s_m in graph G. The wealth of C can be be quantified using its constituent entities. It can be described by several measures, such as entity count, mean and median of its entities' wealth, and percentile of its entities's wealth. To illustrate, let a class C consists of 4 entities s_1, s_2, s_3, and s_4 from Figure 3. We may say class C has a total wealth of 4 when we look from count of entities.

3.2. Insight Model

Exploratory Data Analysis (EDA): Descriptive Statistics Measures. Descriptive statistics is concerned with the description and summarization of data. It is a summary of a dataset that helps to describe features of data quantitatively (Ross, 2019). To have a general view of wealth distribution of a class, we use the following measures:

- measures of central tendency: mean, median, mode
- measures of frequency: count, cumulative frequency/percentage
- measures of position: quartile, percentile
- measures of dispersion: minimum, maximum, range, interquartile range, standard deviation, coefficient of variation, kurtosis
- measures of symmetry: skewness

Gini Coefficient Gini coefficient is a metric used to measure the economic wealth gaps between countries. A study by Akbar (2020) utilized the Gini coefficient to measure the level of knowledge imbalance in knowledge graphs, particularly Wikidata classes. To calculate the imbalance level of a Wikidata class using Gini coefficient, the researcher started by calculating the number of properties of each entity of that particular class and storing them in an array. The array will then be sorted in descending order, from the smallest to the largest i.e., $y_i \geq y_{i+1} \forall i \in \{1, 2, ..., n\}$. The Gini coefficient will be calculated from the sorted array using the Gini coefficient formula below.

$$G = 1 - \frac{1}{n^2 \mu} \sum_{i=1}^{n} \sum_{j=1}^{n} Min(y_i, y_j)$$

$$G = 1 + \frac{1}{n} - \frac{1}{n^2 \mu} - (y_1 + 2y_2 + ... + ny_n)$$

In economics context, n is the size of population of a country, μ is the average income, and y is an array containing data of each country's income. However, in the context of knowledge graph, n is the number of entities in the class, μ is the average knowledge wealth of the entities, and y is an array containing data of each entity's wealth, sorted in descending order.

For example, let's say we have a class that consists of 10 entities. After counting the number of properties of each entities (using the notion of bag of properties for wealth) and sorting them in descending order, we will have an array of $y = [10, 8, 8, 7, 4, 2, 2, 1, 1, 1]$. The length of the array is $n = 10$ and the average wealth is $\mu = 4.4$. Then, apply the Gini coefficient formula and we get $G = 1 + \frac{1}{10} - \frac{2}{10^2 \times 4.4}(1 \times 10 + 2 \times 8 + ... + 10 \times 1) = 0.414$

For another another example, let's say we have another array of 10 entities $z = [10, 9, 9, 9, 9, 9, 9, 9, 9, 5]$. By applying the same formula to z, we get a Gini coefficient value of 0.052.

The Gini coefficient has a value between 0 and 1. The higher the coefficient value, the greater the imbalance level. The value of 0 is achieved when all observed entities have the same amount of wealth. The value of 1 occurs when all income is owned solely by one entity and this phenomenon expresses full inequality.

Lorenz Curve Lorenz curve is a graphical representation of wealth inequality (The Lorenz Curve: What It Tells You About Economic Inequality, 2022). It shows how the wealth is cumulatively distributed, with data points sorted from the poorest to the richest. In Lorenz curve, the horizontal axis represents the fraction of the population, and the vertical axis represents the cumulative wealth. Therefore, if the point $(x, y) = (30, 15)$ lies on the curve, then we can interpret that the bottom 30% of the population account for 15% of the total wealth in that population. The Lorenz curve is usually drawn along with a straight diagonal line with a slope of 1. This straight line represents perfect equality in wealth distribution, i.e., each individual in the observed population has equal wealth. The Lorenz curve itself is drawn below the straight line. The ratio of the area between the Lorenz curve and the straight line of perfect equality to the triangular area below the straight line, is the Gini coefficient.

3.3. Sample Application of Formal and Insight Model

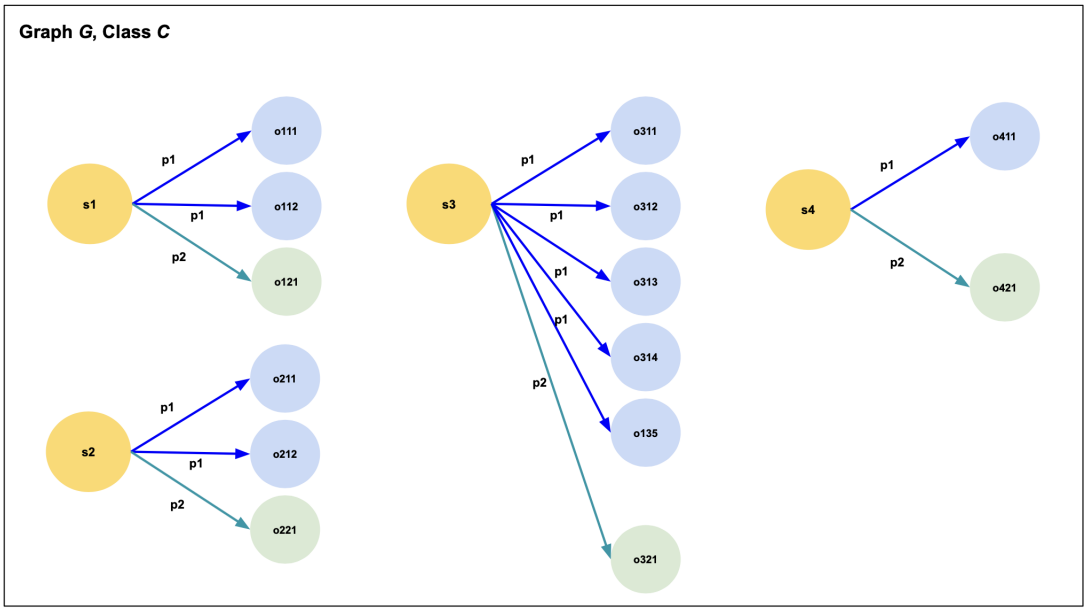


Figure 3: Sample knowledge graph G that contains class C with 4 entities

We provide small case in Figure 3 to illustrate how both models can be applied in quantifying

knowledge wealth. In this example, we will focus on the notion using bag of properties with outgoing direction of link to quantify the wealth.

A graph G has a class C, which consists of four entities s_1, s_2, s_3, and s_4. Each entity has two distinct properties p_1 and p_2. For example, entity s_1 is linked by property p_1 to objects o_{111} and o_{112}, and by property p_2 to object o_{121}. Using the bag of properties and outgoing link direction, the wealth of entity s_1 is 3 (2 accounted for by p_1 and 1 by p_2). Similarly, the wealth of entities s_2, s_3, and s_4 is 3, 6, and 2, respectively. Table 1 provides a statistical summary describing the wealth of class C. Entities in class C have a mean wealth of 3.5, a median wealth of 3, a mode wealth of 3, a minimum wealth of 2, and a maximum wealth of 6. Based on each individual entity's wealth, the imbalance measure of class C is quantified using the Gini coefficient, which has a value of 0.21.

Table 1: Statistical Summary of Wealth of Class C

Measure	Entity Count	Mean	Median	Mode	Minimum	Maximum	Gini
Value	4	3.5	3	3	2	6	0.21

This table shows some statistical measures to quantify the wealth of class C.

In addition to the Gini coefficient, the Lorenz curve for the wealth of entities in class C is shown in Figure 4. This figure illustrates that the wealth distribution within class C is very close to the diagonal line of perfect equality, which aligns with the small Gini coefficient value of 0.21.

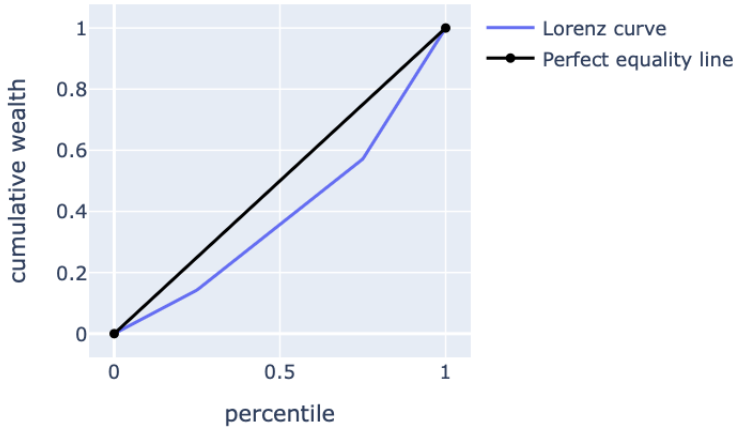


Figure 4: Lorenz curve of class C

4. Use Cases and Evaluation

This section will discuss use cases and evaluation of the research. The data used were taken from Wikidata and retrieved using Wikidata Query Service on <*input date here*>. We only

consider we omit the usage of blank node, as this adds complexity to the analysis and may result in quality issue. The data is then processed and analyzed using a Python library that we created.

4.1. Bias in Wikidata

In this subchapter, analysis is conducted to see whether any particular entity group in Wikidata is underrepresented compared to others. There are 2 analysis done: gender bias and western bias.

Gender Bias in Wikidata Gender bias analysis in Wikidata will be performed on 10 Wikidata classes: computer scientist, American singer, American actress/actor, badminton player, businessperson, lawyer, American politician, American writer, American researcher, and American journalist.

To analyze the bias, the first aspect that will be considered is the proportion of each gender in every class. We assume that there are equal numbers of males and females in real-world and this will be the basis to determine if there is any bias in the data. Pearson's chi-square test (goodness-of-fit) is then performed to test the null and alternative hypotheses with significance level of $\alpha = 5\%$ as follows:

H_0: The proportions of males and females in a particular class are equal to the real-world proportion

H_1: The proportions of males and females in a particular class are not equal to the real-world proportion

From Table 2, we can see that there are more male entities than female entities in all of the classes. In terms of entity count, the gender gaps in some classes such as American singer, American actress/actor, badminton player, and American writer, are slim. The gender gaps in some other classes are huge, and it can be observed in the classes of computer scientist, businessperson, lawyer, American politician, journalist, and researcher. This phenomenon can also be easily identified through visualization, as exhibited in Figure 5a, where the histogram of the female subclass is much smaller compared to the male. Looking at the chi-square test result, as p-value is well below the chosen significance level, the null hypothesis is rejected in all classes. Hence, we consider the difference of entity count to be significant and conclude that the proportions of males and females in each Wikidata class are not the same as the assumed real-world proportion of 50%-50%.

However, it is arguable that, for some classes, the gap in entity count between both genders is expected because, in reality, there are more men than women in the workforce, especially in particular fields such as engineering. As a consequence, it is not reasonable if we expect to have an equal number of males and females entities in Wikidata. Therefore, entity count may not be a good measure of bias because of the nature of the data itself. To address this, we need to evaluate other metrics which can quantify the bias on entity-level.

Table 2: Entity Count of 10 Wikidata Classes per Gender Category

Class Name	Entity	Male	Female	%Male	%Female	χ^2	p-value
American actress/actor	38087	21451	16636	0.56	0.44	608.72	2.13e-134
American journalist	17740	12223	5517	0.69	0.31	2534.97	0.0
American politician	92901	83007	9894	0.89	0.11	57539.86	0.0
American researcher	4867	3387	1480	0.70	0.30	747.21	1.63e-164
American singer	15712	9027	6685	0.57	0.43	349.09	6.67e-78
American writer	32573	19113	13460	0.59	0.41	981.07	2.34e-215
Computer scientist	17914	15229	2685	0.85	0.15	8783.74	0.0
Badminton player	25283	13377	11906	0.53	0.47	85.58	2.22e-20
Businessperson	74538	66706	7832	0.89	0.11	46501.76	0.0
Lawyer	91348	80639	10709	0.88	0.12	53533.79	0.0

This table shows the entity count of 10 Wikidata classes per Gender Category. Chi-square test result shows the significance of difference between the entity count of the two genders male and female.

The next metrics to be considered are the measures of central tendency and dispersion to see where the wealth distribution is concentrated and how the data spread.

From Table 3, female entities generally have lower values of measure of central tendency (mean, median, mode). These characteristics can also be observed from the histogram in Figure 5a: female histograms' peak and dense area are located on the left of the male's. The range of property count of females is generally also lower than males. However, there are some classes in which the richest entity is a female. An example for this is the class of American Singer, which is shown by Figure 5b. Though the value of mean, median, and mode of count of properties are lower for female compared to male, the richest entity on that class is a female entity Madonna (Q1744) with bag of property count of 687, with a significant difference with Michael Jackson (Q2831) with bag of property count of 574. We also observed positive values of skewness (skewness > 0) and high kurtosis values (kurtosis > 3) in all classes, denoting the wealth distribution is right skewed and leptokurtic.

Table 3: Measures of Central Tendency of 10 Wikidata Classes per Gender Category

Class Name	Mean (o/m/f)	Median (o/m/f)	Mode (o/m/f)
American actress/actor	38.96/39.85/37.80	29.00/30.00/28.00	19/19/22
American journalist	30.71/32.44/26.89	23.00/25.00/20.00	14/14/14
American politician	19.22/19.33/18.25	15.00/15.00/15.00	13/9/13
American researcher	23.97/25.00/21.63	20.00/21.00/18.00	12/12/12
American singer	42.78/42.99/42.51	31.00/33.00/30.00	18/24/15
American writer	38.86/42.76/33.33	30.00/33.00/26.00	19/22/19
Computer scientist	24.16/24.50/22.28	19.00/19.00/18.00	8/8/8
Badminton player	21.50/21.25/21.78	16.00/16.00/16.00	13/13/13
Businessperson	16.91/16.83/17.61	13.00/13.00/13.00	10/10/9
Lawyer	22.44/22.98/18.37	19.00/19.00/15.00	16/16/12

This table shows the measures of central tendency of 10 Wikidata classes per gender category. Each measure will have 3 values: o (overall), m (male), and f (female).

Table 4: Measures of Dispersion and Symmetry of 10 Wikidata Classes per Gender Category

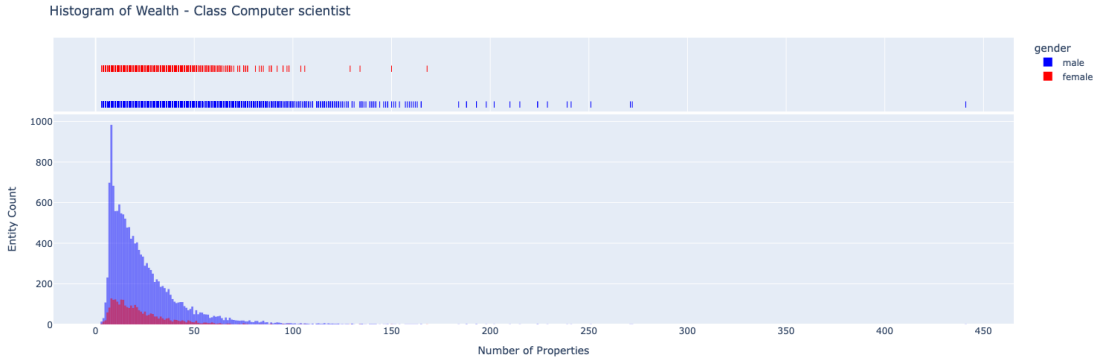
Class Name	Min (o/m/f)	Max (o/m/f)	Std. Deviation (o/m/f)	Skewness (o/m/f)	Kurtosis (o/m/f)
American actress/actor	4/4/4	687/574/687	33.89/32.75/35.28	4.23/3.54/4.94	30.74/21.59/39.53
American journalist	4/5/4	402/402/353	26.84/28.14/23.23	4.21/4.18/4.15	30.73/30.18/29.38
American politician	4/4/4	476/476/328	14.77/14.77/14.75	6.53/6.55/6.40	97.36/100.39/72.68
American researcher	4/4/4	222/222/171	16.89/18.19/13.17	3.86/3.79/3.45	25.67/23.50/26.78
American singer	4/4/4	687/574/687	39.35/35.33/44.21	4.09/3.16/4.61	29.38/18.24/33.19
American writer	4/4/4	476/476/425	34.12/37.19/28.30	3.56/3.29/4.05	20.51/17.35/27.82
Computer scientist	3/3/3	441/441/168	19.53/20.18/15.24	3.41/3.45/2.20	27.66/27.80/9.33
Badminton player	4/9/4	360/238/360	16.09/15.28/16.96	4.23/3.89/4.48	30.85/22.98/35.97
Businessperson	3/3/3	574/574/429	14.65/14.01/19.27	7.73/7.37/8.15	130.49/128.49/108.21
Lawyer	3/3/3	550/550/328	16.61/16.91/13.47	5.56/5.59/5.23	76.02/76.31/64.39

This table shows the measures of dispersion and symmetry of 10 Wikidata classes per gender category. Each measure will have 3 values: o (overall), m (male), and f (female).

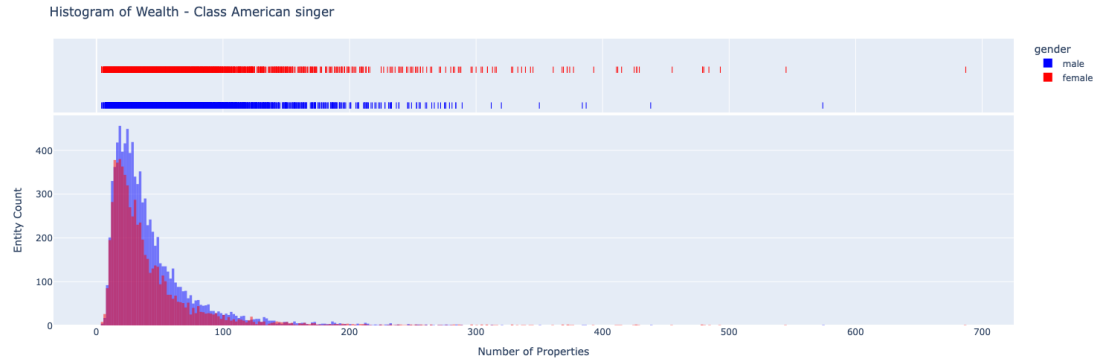
At a glance we saw female classes are poorer compared to the male classes. To test this, we will use t-test and Welch's test. First, we performed F-test to check if the male and female classes have equal variance. The result of F-test is then used to determine the appropiate test to be used in each class. Those with equal variance will use t-test; otherwise Welch's test is used. Then, we performed the tests to verify the null and alternative hypotheses with significance level of $\alpha = 5\%$ as follows:

H_0: The means of wealth of males and females in a particular class are equal

H_1: The means of wealth of males and females in a particular class are not equal



(a) Histogram and Marginal Distribution Plot of Wealth for Class Computer Scientist



(b) Histogram and Marginal Distribution Plot of Wealth for Class American Singer

Figure 5: Histogram of wealth

Table 5: F-Test, T-Test, and Welch's Test Result of 10 Wikidata Classes

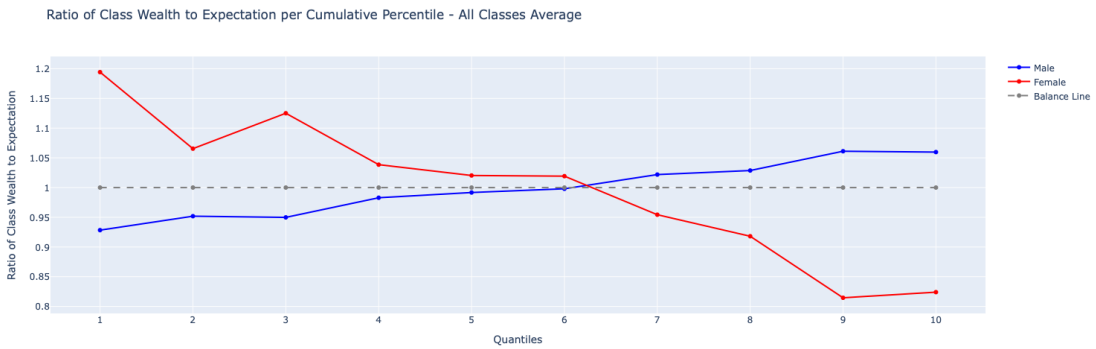
Class Name	F-Test statistic	F-Test p-value	T-Test statistic	T-Test p-value	Welch's Test statistic	Welch's p-value
American actress/actor	0.86	0.00	5.85	4.97e-09	5.80	6.89e-09
American journalist	1.47	1.00	12.80	2.45e-37	13.75	1.01e-42
American politician	1.00	0.55	6.86	6.90e-12	6.87	6.94e-12
American researcher	1.91	1.00	6.43	1.35e-10	7.28	4.15e-13
American singer	0.64	0.00	0.75	0.45	0.73	0.47
American writer	1.73	1.00	24.80	1.56e-134	25.98	2.96e-147
Computer scientist	1.75	1.00	5.43	5.57e-08	6.60	4.61e-11
Badminton player	0.81	0.00	-2.63	8.46e-03	-2.62	8.86e-03
Businessperson	0.53	0.00	-4.48	7.56e-06	-3.49	4.81e-04
Lawyer	1.58	1.00	27.06	1.16e-160	32.16	8.04e-220

From the test results in Table 5, we rejected the null hypothesis in 9 out of 10 class–American singer being the only exception. 7 out of 9 classes are in favor of male. The other 2 classes, American singer and . We concluded that female classes are more likely to have smaller means than male classes.

Here, a new measure is defined: top $x\%$ male/female relative to the expectation. The value of expectation of a gender in a class is equal to the percentage of that particular in the class. Top $x\%$ male relative to expectation is the ratio of percentage of male entities in the top $x\%$ to the expectation. Similarly, top $x\%$ female relative to expectation is the ratio of percentage of female entities in the top $x\%$ to the expectation.

When the shape of distribution of male and female of a class is the same (in other word, the wealth is distributed equivalently to male and female entities), then the value of top $x\%$ relative to expectation should be 1 for both male and female subclasses. A value higher than 1 indicates domination by that particular gender.

(a) Ratio of Class Wealth to Expectation per Cumulative Top Percentage - All Classes Average



(b) Ratio of Class Wealth to Expectation per Quantile - All Classes Average

Figure 6: Ratio of each gender wealth to expectaion

From Figure 6 the value of ratio between top $x\%$ potion to the expectation in the above tables, we can see that on average, the rich entities are dominated by male. Exceptions are held for 3 classes, that is classes American singer, businessperson, and badminton player. Moreover, as we

set bigger portions (higher percentage), the gap of ratio between the two ender in each class decreases i.e. the value of top $x\%$ relative to expectation of both genders converge to 1.

Western Bias in Wikidata Western bias analysis in Wikidata will be performed on 5 Wikidata classes: computer scientist, singer, memorial, university, and river. For each class, we collect the data for the western portion from 8 countries: Canada, France, Germany, Ireland, Poland, Switzerland, the United Kingdom (UK), and the United State of America (USA). For the non-western portion, we also chose 8 countries: China, Egypt, India, Indonesia, Japan, Morocco, Nigeria, and South Africa.

To analyze the bias, the first aspect that will be considered is the proportion of each regional category in every class. We assume that there are equal numbers of western and non-western and this will be the basis to determine if there is any bias in the data. Pearson's chi-square test (goodness-of-fit) is then performed to test the null and alternative hypotheses with significance level of $\alpha = 5\%$ as follows:

H_0: The proportions of western and non-western entities in a particular class are equal

H_1: The proportions of western and non-western entities in a particular class are not equal

Table 6: Entity Count of 5 Wikidata Classes per Regional Category

Class Name	Entity	Western	Non-western	%Western	%Non-western	χ^2	p-value
Computer scientist	6063	5446	617	0.90	0.10	3846.15	0.0
Singer	43240	31039	12201	0.72	0.28	8206.99	0.0
Memorial	4011	3836	175	0.96	0.04	3341.54	0.0
University	6124	2398	3726	0.39	0.61	287.98	1.37e-64
River	125567	70059	55508	0.56	0.44	1686.20	0.0

This table shows the entity count of 5 Wikidata classes per regional category. Chi-square test result shows the significance of difference between the entity count of the two genders male and female.

In terms of entity count, Table 6 shows that there are big gaps between the westerners and non-westerners in all of the classes. From Table 7, non-western entities generally have lower values of measure of central tendency (mean, median, mode). The range of property count of non-westerns is generally also lower than the westerns. Positive values of skewness (skewness > 0) and high kurtosis values (kurtosis > 3) in all classes denote the wealth distribution is right skewed and leptokurtic.

Table 7: Measures of Central Tendency of 5 Wikidata Classes per Regional Category

Class Name	Mean (o/w/n)	Median (o/w/n)	Mode (o/w/n)
Computer scientist	35.00/35.87/27.24	29.00/29.00/22.00	21/15/16
Singer	34.99/39.14/24.43	25.00/29.00/18.00	15/18/14
Memorial	11.04/11.04/11.13	9.00/9.00/9.00	9/9/9
University	23.11/31.61/17.63	17.50/24.00/16.00	6/6/6
River	7.69/8.55/6.60	7.00/7.00/6.00	7/7/7

This table shows the measures of central tendency of 5 Wikidata classes per regional category. Each measure will have 3 values: o (overall), w (western), and f (non-western).

Table 8: Measures of Dispersion and Symmetry of 5 Wikidata Classes per Gender Category

Class Name	Min (o/w/n)	Max (o/w/n)	Std. Deviation (o/w/n)	Skewness (o/w/n)	Kurtosis (o/w/n)
Computer scientist	4/4/5	441/441/145	25.15/25.67/18.15	3.02/3.02/2.37	20.90/20.83/8.49
Singer	3/4/3	687/687/379	33.27/36.60/18.94	4.49/4.28/3.15	35.56/31.09/21.59
Memorial	2/3/2	142/142/52	5.89/5.82/7.28	8.50/8.95/2.98	143.57/156.15/12.79
University	2/2/2	234/234/166	20.00/25.88/12.25	2.37/1.65/2.22	10.34/5.33/12.70
River	2/2/2	452/452/148	5.24/6.46/2.68	21.71/19.54/14.67	1152.84/868.56/465.42

This table shows the measures of dispersion and symmetry of 5 Wikidata classes per gender category. Each measure will have 3 values: o (overall), w (western), and f (non-western).

Out of 5 classes, class of memorial is the only class where the null hypothesis is not rejected. In the other 4 classes, we can see a significant difference between the mean of the two reginal categories, which all are in favor of the western.

Table 9: F-Test, T-Test, and Welch's Test Result of 5 Wikidata Classes

Class Name	F-Test statistic	F-Test p-value	T-Test statistic	T-Test p-value	Welch's Test statistic	Welch's p-value
Computer scientist	2.00	1.00	8.13	5.10e-16	10.66	3.64e-25
Singer	3.73	1.00	42.22	0.0	54.59	0.0
Memorial	0.64	0.00	-0.19	0.85	-0.16	0.88
University	4.46	1.00	28.40	8.23e-167	24.73	2.12e-123
River	5.83	1.00	66.91	0.0	72.64	0.0

(a) Ratio of Class Wealth to Expectation per Cumulative Top Percentage - All Classes Average

(b) Ratio of Class Wealth to Expectation per Quantile - All Classes Average

Figure 7: Ratio of each regional wealth to expectaion

4.2. Effect of Type of Wealth on Inequality Measure

In this subchapter, analysis is done to see how each wealth type affects the level of inequality of Wikidata classes. There are 2 ways this is done–quantitatively using Gini coefficient and qualitatively using Lorenz curve. The analysis is performed on 8 Wikidata classes, in which 4 of them are human-related class while the other 4 are not.

Table 10: Knowledge Wealth Type on Gini Coefficient

Class Name	Gini Bag	Gini Set	Gini Object	Gini Pure Literal	Gini ID	Gini Outgoing	Gini Incoming
American researcher	0.33	0.31	0.26	*0.65	0.50	0.33	0.77
American singer	0.40	0.39	0.29	0.36	0.53	0.40	0.82
Badminton player	0.29	0.14	0.30	*0.14	0.60	0.29	0.68
Computer scientist	0.41	0.37	0.36	*0.64	0.56	0.41	0.81
Historical painting	0.23	0.15	0.25	0.26	0.45	0.23	0.87
Memorial	0.22	0.19	0.20	0.31	0.41	0.22	0.99
Sci-fi book	0.30	0.21	0.31	0.33	0.44	0.30	0.82
University	0.44	0.41	0.39	0.49	0.53	0.44	0.91

This table shows the comparison of Gini coefficient of 8 Wikidata classes

When looking at the notion of wealth using the characteristics of (non-)uniqueness of individual properties, it is intuitive that the measure of the bag of properties will always give higher (or at least, equal) amount of wealth compared to the measure of set. Set of property will have an upper bound of number of unique property, while the bag of properties does not have any upper bound. Moreover, using the bag of properties, a large number of triples having the same property may inflate the wealth substantially–though this is not necessarily a problem nor an advantage. This characteristics has a direct impact on inequality measure and it is well depicted on the value of Gini coefficient. From Table 10, in all clasess, the Gini coefficient using the bag of properties is always higher than of set of properties.

Using the notion of wealth by type of property, in general the smallest Gini coefficient value comes from wealth using object properties, followed by literal properties, with ID properties having the highest value. This condition holds for the four non-human classes that are being observed. However, there are anomalies in the human classes, as illustrated in Figure 8.

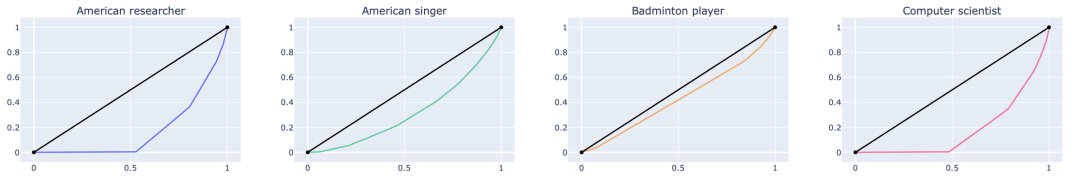


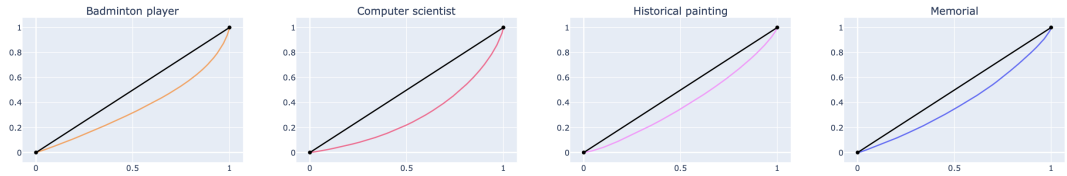
Figure 8: Lorenz curve of wealth using literal properties

In the classes of American researchers and computer scientists, the Gini coefficients for wealth using pure literal properties are significantly higher than those of other classes. This is due to approximately 50% of the entities in these classes having 0 pure literal properties–not even basic human-related literal properties such as *date of birth* (P569). For example, an American researcher *John Heidemann* (Q29354581) has a total wealth of 61 (using the bag of properties) and is among the top 4% entities in his class, yet this entity has 0 literal properties. In contrast, only 818 out of 16150 (just about 5%) entities of American singer class have 0 literal properties.

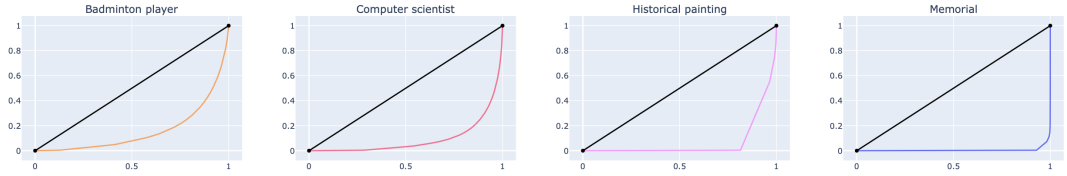
Moreover, an instance of American singer, Kris Allen (Q216927), who also has a total wealth of 61 (using the bag of properties), has 6 literal properties.

Another anomaly is observed in the class of badminton player where the Gini coefficient for wealth using pure literal is significantly lower than that of other classes. Upon inspection, three main reasons for this anomaly were identified. First, there are no entities in this class with 0 literal properties; all entites have at least 1 literal. Second, the range of wealth in this class is smaller, with a minimum value of 1 and maximum value of 10, in comparison to the classes of American researcher, American singer, and computer scientist which all have a minimum value of 0 and maximum values of 16, 26, and 54, respectively. This smaller range means the difference between the poorest and richest entities in the class of badminton player is minimal, and a small wealth range generally leads to a low Gini coefficient because inequality is limited by the narrow spread of wealth. The third reason is the large number of entities in this class, which totals 25,402. With such a high population size, the wealth of each individual—whether the poorest, middle, or richest entity—contributes only a tiny fraction to the total wealth of the class. Combined with the small wealth range, this further reduces the inequality measure, resulting in a low Gini coefficient.

Using the notion of wealth by the direction of the link, the Gini coefficient when using incoming link is always higher than using outgoing link. By inspecting the Lorenz curve, we can see that most entities do not have any incoming link, and only the small percentage of entites has some incoming link. Figure 9 shows the comparison of Lorenz curve of knowledge wealth based on the direction of the link from 3 Wikidata classes. The difference between the two is very significant, because in Figure 9a the Lorenz curves are closer to the perfect equality line, meanwhile in Figure 9b the diagonal and the Lorenz curve almost form a right triangle which is very close to maximum inequality.



(a) Lorenz curve of wealth using outgoing link



(b) Lorenz curve of wealth using incoming link

Figure 9: Comparison of Lorenz curve of wealth based on the direction of link

4.3. Contribution of Property Type to Knowledge Wealth

In this subchapter, analysis is done to see how much contribution of each property types to the knowledge wealth of Wikidata classes. The analysis is performed on 8 Wikidata classes, in which 4 of them are human-related class while the other 4 are not. The wealth of each classes is calculated using the concept of bag of properties and set of properties, for each property types (object, literal, ID, and their combinations).

Two methods of averaging are used in this analysis—global percentage and average contribution per individual entity. The first method is done by calculating the total sum of each property across all entities and then dividing it by the grand total of all properties combined, providing a holistic view of each property's overall contribution. The second method is done by first determining the percentage contribution of each property within each individual entity and then averaging these percentages across all entities, ensuring that every entity is equally represented regardless of its scale. While the global percentage method highlights absolute contributions, the average per-entity method captures relative contributions within each entity, making it more robust to variations in scale and less sensitive to outliers since every data point contributes equally to the final result.

Table 11: Contribution of Property Type to Knowledge Wealth Using Bag of Properties

Class Name	% Object Bag	% Literal Bag	% ID Bag	% Object + Literal Bag
American researcher	59.55 / 65.65	3.32 / 2.8	37.12 / 31.55	62.88 / 68.45
American singer	40.44 / 48.42	7.47 / 8.48	52.09 / 43.1	47.91 / 56.9
Badminton player	79.34 / 78.96	10.14 / 11.68	10.52 / 9.36	89.48 / 90.64
Computer scientist	59.36 / 65.93	4.19 / 3.71	36.44 / 30.36	63.56 / 69.64
Historical painting	66.35 / 66.46	25.48 / 25.85	8.17 / 7.69	91.83 / 92.31
Memorial	62.82 / 64.16	21.0 / 20.4	16.18 / 15.44	83.82 / 84.56
Sci-fi book	62.96 / 62.62	14.44 / 15.78	22.59 / 21.6	77.41 / 78.4
University	37.59 / 44.79	18.11 / 18.31	44.3 / 36.9	55.7 / 63.1

This table shows the contribution of each property type of 8 Wikidata classes based on the notion of bag of properties. Each record has 2 values separated by "/". The first value is the contribution rate when calculated with global percentage method, and the second one is when using the average contribution per individual entity

Table 12: Contribution of Property Type to Knowledge Wealth Using Set of Properties

Class Name	% Object Set	% Literal Set	% ID Set	% Object + Literal Set
American researcher	51.25 / 59.69	3.97 / 3.29	44.78 / 37.02	55.22 / 62.98
American singer	33.91 / 43.58	8.1 / 9.18	57.99 / 47.24	42.01 / 52.76
Badminton player	71.75 / 74.13	13.79 / 13.9	14.46 / 11.97	85.54 / 88.03
Computer scientist	50.53 / 60.54	4.98 / 4.27	44.49 / 35.19	55.51 / 64.81
Historical painting	60.57 / 61.91	28.88 / 28.61	10.55 / 9.48	89.45 / 90.52
Memorial	60.06 / 62.01	22.41 / 21.59	17.54 / 16.4	82.46 / 83.6
Sci-fi book	59.36 / 61.91	13.86 / 14.59	26.78 / 23.5	73.22 / 76.5
University	33.93 / 42.74	17.67 / 18.32	48.4 / 38.95	51.6 / 61.05

This table shows the contribution of each property type of 8 Wikidata classes based on the notion of set of properties. Each record has 2 values separated by "/". The first value is the contribution rate when calculated with global percentage method, and the second one is when using the average contribution per individual entity

From Table 11 and Table 12, generally, the usage of pure literal properties is lower than of object and ID. This aligns with the principle of knowledge graph to use URI for internal connections and ID for external resources.

We can say that object properties and literal properties combined represent wealth sourced internally from Wikidata, whereas ID properties originate from external sources. In general, more than half of the wealth is attributed to internal resources.

Typically, an ID property has exactly one associated value because it serves as a unique identifier in external resources, preventing ambiguity. In contrast, object properties can have multiple values. For example, the computer scientist *Noam Chomsky* (Q9049) has an ID property, *VIAF cluster ID* (P214), with exactly one value: *89803084*. At the same time, he has an object property, *work location* (P937), with four values: *Pennsylvania* (Q1400), *Cambridge* (Q350), *Tucson* (Q18575), and *Massachusetts* (Q771). When using the bag of properties, *VIAF cluster ID* (P214) contributes 1 to the total wealth, while *work location* (P937) contributes 4. Under the set of properties, both properties contribute equally, each accounting for 1. This explains why the percentage contribution of object properties to wealth is smaller when using the set of properties compared to the bag of properties. Since the overall property count is lower in the set of properties, the percentage contribution of ID properties to wealth is correspondingly higher than in the bag of properties.

5. Discussion

Generality of Framework The framework that is proposed in this study is applicable to any kind of knowledge graph. However, the library that we built for experimentation is specifically for Wikidata, because the query service, structure, and response is very unique for each knowledge graph. If further experimentation is to be conducted for other knowledge graph, then the library needs to be modified.

Issue of Incoming Link Wikidata is an entity-centric knowledge graph which means in the editathon efforts, the subject s is always be the subject of interest and the starting point to be edited by contributors. It is very unlikely that the opposite approach is done, that is, an object (o) is given and a pair subject and property (s, p) is to be searched. Not only from the contributors, the platform itself does not support the later view. This explains the phenomenon that we observed in subsection 4.2 regarding outgoing and incoming link. With existing subject-centric point of view, the number of new outgoing link introduced to the knowledge graph will grow in a much higher rate as opposed to incoming link. As a result, incoming link will be very rare, or even only present in certain entites.

Weight of Property Set of properties might be more suitable if the main concern is the presence of properties, instead of the abundance of information it contains. We will take *William Shakespeare* (Q692) in Wikidata as an example. It is reasonable if property like *date of birth* (P569) to be treated using set of properties, but for a well-known playwright and poet, we shall expect the property *notable work* (P800) to incorporate all or most of his well-known works. If only few works registered in G despite he has dozens of works, then we might conclude the entity is poor. For this case, treating the property using the notion of set is not preferable because it will fail to capture the aforementioned poor condition, while blatantly using bag of properties might skew the wealth amount.

Due to this, the notion of (non-)uniqueness can be extended to a weighted form. The weight is given independently for each property and can be defined in such a way that is most appropriate for the nature of the property. Example of weight is inverse of median of property count of each entity.

Let h_{p_i} be the weight of property p_i in graph G. Let $N_{bag}(s, G, p_i)$ be a set that comprises all pair (p_i, o), that is, property p_i and an object o that is connected to s. Then $W_{weighted}(s, G)$ is the sum of $N_{bag}(s, G, p_i)$ multiplied by the associated weight h_{p_i}.

$$N_{bag}(s, G, p_i) = \{(p_i, o) | (s, p_i, o) \in G\}$$

$$W_{weighted}(s, G) = \sum_i |N_{bag}(s, G, p_i)| * h_{p_i}$$

Using Figure 3, we can show how the notion of weighted wealth can be calculated. Let's define $h_{p_1} = 1/median$ and $h_{p_2} = 1$. For property p_1, $\{1, 2, 2, 4\}$ is the sorted amount of information contributed from p_1 in each individual entity from s_1 to s_4, from which we get $h_{p_1} = 1/median = 1/2$. With the above definition, $W_{weighted}(s_1, G) = 2$, $W_{weighted}(s_2, G) = 2$, $W_{weighted}(s_3, G) = 3.5$, $W_{weighted}(s_4, G) = 1.5$.

Another phenomena that should also be considered is that an entity might have several occurrences of the same property, but this property might just be 'trivial'. This is similar to a document having a lot of 'the' or 'a' (stopwords). Conversely, an entity might just have a single occurrence of a property, but it is a non-trivial one. Perhaps, the property is a highly relevant one for the entity's class. For example, *time period* (P2348) for *William Shakespeare* (Q692) will be a highly relevant one for prominent poets. However, the property *sibling* (P3373) might not be too relevant for Shakespeare's career. One idea to handle such trivial is to regard those porperties using the notion of set of properties–thus we care only on its existence rather

than its actual count–or, we can use threshold function to set a tolerance for how much of such trivial we allow.

Thus, the notion of (non-)uniqueness can be generalized to accommodate the bag, set, and weighted concept. Let s_1, s_2, ... s_m be m distinct entities in graph G, which all of them collectively create a class C. Let $N_{bag}(s, G, p_i)$ be a set that comprises all pair of a particular property p_i and object, (p_i, o), that is connected to s. We define T_{C,p_i} a multiset consisted of the number of non-unique properties of each entity of C contributed from property p_i, i.e., multiset of cardinality of $N_{bag}(s, G, p_i)$ for all s in C. For each p_i, let f_{p_i} be a multivariate function with 2 arguments: T_{C,p_i} and $N_{bag}(s, G, p_i)$. w_{p_i} is the amount of wealth of s attributed from property p_i, which is calculated by function f_{p_i}. Then $W(s, G)$ is the sum of w_{p_i}.

$$
\begin{aligned}
N_{bag}(s, G, p_i) &= \{(p_i, o) | (s_j, p_i, o) \in G\} \\
T_{C,p_i} &= \{|N_{bag}(s, G, p_i)| \ | s \in C\} \\
w_{p_i} &= f_{p_i}(T_{C,p_i}, |N_{bag}(s, G, p_i)|) \\
W(s, G) &= \sum_i w_{p_i}
\end{aligned}
$$

Using Figure 3, we can show how the generalized form can be utilized. From the definition, we get $T_{C,p_1} = \{2, 2, 5, 1\}$ and $T_{C,p_2} = \{1, 1, 1, 1\}$. Property p_1 illustrates the former issue of trivial. Here, we define $f_{p_1}(x, y) = \begin{cases} x & \text{if } x \leq 1 \\ \frac{x}{2} + 1 & \text{if } 1 < x \leq 4 \\ 3 & \text{if } x > 4 \end{cases}$. For p_2, we want to handle it using the notion of set of properties, thus we define it using a constant function $f_{p_2}(x, y) = 1$. With the above definition, $W(s_1, G) = 3$, $W(s_2, G) = 3$, $W(s_3, G) = 4$, $W(s_4, G) = 2$.

Multiclass Entity and Misclassification In data classification, entities may belong to multiple classes or be misclassified due to inconsistencies in data sources. For instance, the richest entity in the computer scientist class based on pure literal properties count, *Bruno Darcet* (Q71055086), has a total wealth of 54, which accounts from 1 value in *date of birth* (P569) and 53 values in *Elo rating* (P1087). The same case is also observed in the second richest entity, *Daniel José Queraltó* (Q23906907). From their profiles, we may say these 2 entities are not actually computer scientists, highlighting an occupational misclassification issue. A similar problem occurs with the classification of American researcher, where the wealthiest by pure literal count, *Alexander Julien* (Q61249179), is not a university researcher. In the case of badminton players, *Dirk Stikker* (Q194654), the richest individual in the class based on ID properties count, is more of a politician with badminton affiliations rather than a prominent player. These examples illustrate that an entity can belong to multiple classes, or be entirely misclassified. These classification inconsistencies reveal data quality issues, necessitating careful data querying, cleaning, and validation to ensure accuracy prior to analysis.

6. Conclusions

TBD

Acknowledgments

TBD

References