645 Final Project: Implementation of *Extracting Top-K Insights from Multidimensional Data*

Josh Sennett
ID: 31751015
jsennett@umass.edu

1. Problem Statement

A common task in data analytics is to search for interesting points and trends in data by aggregating a measure across dimensions. Since data can be aggregated in many ways, data analysts face an enormous space of possible analyses. Since manually analyzing all possible aggregate measures is infeasible, analysts must specify which group-by attributes and aggregate measures to look for, causing valuable insights to go unnoticed.

The authors of *Extracting Top-K Insights from Multidimensional Data* [1] propose a system to automate extracting insightful observations. To do so, the system enumerates over possible *subspaces* (subsets of the data with certain dimension values), and calculates aggregate measures for each subspace of a *sibling group* (where siblings of a sibling group take on each value of a particular *dividing dimension*) to extract a *result set*. Each *result set* is analyzed using statistical tests to determine if a significant trend or point stands out. An *insight score* is calculated as the product of the *impact* (market share) of the subspace and the significance score of the point or trend.

Further, result sets are recursively aggregated; a *composite extractor* specifies how multiple aggregations can be applied to uncover insights that may not be visible in the original dataset. Through several optimizations, the system is able to enumerate insights over all meaningful composite extractors, all subspaces, and all insight types to provide users with a comprehensive set of insights.

2. Techniques Implemented

To automatically extract top insights from the DBLP dataset, I implemented the *Insight Extraction* component of the original system, including **Algorithm 1** and **Algorithm 2**, as well as the **impact** and various **significance score** functions.

My implementation allows users to extract top-k insights from a dataset that has multiple dimension columns and a single measure column. The system supports both `COUNT` and `SUM` aggregate measures, and implements all four extractors featured in the paper: Δ_{prev}, Δ_{avg}, `RANK`, and `PCT`. Users can specify a depth $\tau \in \{1, 2\}$ to extract either simple aggregate or composite aggregate measures. The project README has instructions to fully reproduce the results obtained in this paper.

2.1. The `InsightExtractor` Class

The `InsightExtractor` class contains the logic for loading data and extracting insights. It can load in data from a file or by setting its `data` attribute manually. The underlying data is stored and manipulated as a `Pandas` DataFrame. When data is loaded, I precompute the sum of the measure for the dataset, since this same value is used to calculate each impact score and need not be recalculated.

Once data has been loaded, insights can be extracted using the InsightExtractor's `extract_insights(depth, k)` method, which is implemented as follows.

2.2. Algorithm 1: Extract Insights

Closely following the original authors' approach, I implement **Algorithm 1** by enumerating all composite extractors of depth τ, and then calling the recursive `enumerate_insight(subspace, dimension, composite_extractor)` for all combinations of composite extractors and dividing dimensions.

The `enumerate_insight` method consists of two phases: in the first phase, the composite extractor is applied to the sibling group to extract a result set, and the result set is assigned a significance score. In the second phase, additional insights are enumerated recursively by narrowing the subspace by fixing a value of the dividing dimension and choosing a new dividing dimension from the remaining unfixed dimensions.

2.2.1 Pruning by Upper Bound Score

I use the authors' Pruning by Upper Bound Score optimization by calculating the impact score of the subspace before considering the insight. If the impact score is below the top k^{th} insight score, the subspace and all child subspaces are too small to return a top-k insight score, so the function returns early without executing Phase 1 or 2.

2.2.2 Phase 1

In the first phase, the method `is_valid(subspace, dimension, composite_extractor)` checks the validity of the sibling group-composite extractor pair, ensuring that the ensuing result set has known values for the aggregate measures. Any invalid combination skips to Phase 2, since recursive calls may be valid.

When a valid combination is encountered, this method calls `extract_result_set(subspace, dimension, composite_extractor)` to create a result set, implemented as described in Section 2.3. Then, it will iterate over both *Shape* and *Point* insight types and determine which significance test to use based on the insight type, composite extractor, and dividing dimension used, given that different result sets will fit different distributions and require different null hypotheses. The approach and justification for choosing each null hypothesis is described in Section 2.4.

Finally, the insight score is calculated as the product of the significance and impact scores; if it exceeds the k^th insight score, an `Insight` object is created to wrap all information associated with the insight, and the insight is added to the `InsightExtractor`'s top-k min-heap.

2.2.3 Phase 2

In Phase 2, the function iterates over each child sibling group by fixing a value of the current dividing dimension and choosing a new dividing dimension from the set of unfixed dimensions. The base case of the recursion is when all dimensions of the subspace are fixed, meaning that there is no remaining dividing dimension to choose.

2.3. Algorithm 2: Extracting Result Sets

For each valid sibling group-composite extractor pair, a result set is computed by applying the extractor to the sibling group. In my implementation, the composite extractor is just an array of (string, string) tuples, each representing a (aggregate measure, analysis dimension) pair. Based on the values of this string, the method `extract_result_set(subspace, dividing_dimension, composite_extractor)` will compute the result set by applying different aggregate measures over the specified dividing dimension and analysis dimension. To do so, I first filter the dataset to contain the values specified by the *subspace*, employ various `Pandas groupby` aggregation functions to implement the Δ_{prev}, Δ_{avg}, `RANK`, and `PCT` extractors, and then return the resulting DataFrame.

2.4. Calculating Insight Scores

Following the authors' approach, I calculate insight scores as the product of a sibling group's impact score and a result set's significance score.

$$insight = imp(SG(S, D)) * sig(\Phi) \tag{1}$$

The impact score is calculated as the market share of the considered subspace, and is precomputed as described in Section 2.2. For a `COUNT` measure, the impact is the number of rows in the subspace divided by the total number of rows in the dataset; for a `SUM` measure, it is the sum of the measure of the subspace rows divided by the sum of the measure across the full dataset.

My approach to calculate significance scores closely follows the approach taken by the authors, but extends it to support additional null hypotheses (discussed in Section 2.4.3 and Section 2.4.4).

I implement both *Point* and *Shape* significance tests. Table 1 describes the four significance tests I use and the insights that can be derived from them. Each significance test returns a (string, float) tuple with the insight meaning and the significance score $\in [0.0, 1.0]$.

Each significance test is implemented in the `significance_tests.py` module, and takes as input a DataFrame with a measure column named `'M'`. The shape significance function requires a second ordinal column, since the shape insight reveals trends over an ordinal dimension.

2.4.1 Linear Shape Significance

As the authors show, it is often insightful to find positive and negative trends in measures and aggregate measures. The linear shape significance test is applied to all result sets with dividing dimension $year$, since that is the only dividing dimension in the datasets analyzed.

The linear shape significance test fits a linear model to the measure and ordinal column using least squares regression. In the original paper, the authors propose calculating the p value for the slope using a logistic regression $L(\mu, \lambda)$ with $\mu = 0.2, \lambda = 2$. Since I found the test sensitive to the parameters used in $L(\mu, \lambda)$, I instead opted to use the p value computed for the slope. Since the p value is based on the null hypothesis that the slope is zero, I found it fitting to use to determine the insight of a non-zero slope. As the authors do as well, I weigh the significance score by the goodness of fit of the linear model.

$$sig = r^2 * (1 - p) \tag{2}$$

2.4.2 Power Law Point Significance

Power law distributions are common in business domains in which a small portion of the data has high values of a measure (such as sales) while the vast majority have a low value, and values do not tend to deviate around a central mean. As shown in Section 3, the power law seems to be a good distribution to model several measures in the DBLP dataset, including the number of publications for each venue and author, and the number of coauthors per publication.

The power law describes data that fit the relation:

$$y = ax^b \tag{3}$$

with $b < -1$, and y representing the probability density. By taking the logarithm of each side, this can equivalently be expressed as a linear model with intercept $\log(a)$ and slope b:

$$\log y = \log(ax^b) \tag{4}$$

$$\log y = \log(a) + b \log x \tag{5}$$

The steps to computing a point significance score are:

1. Remove the maximum value from measure M
2. Compute the ascending rank of values
3. Compute the log of rank and measure M
4. Fit linear model where $x = \log rank$ and $y = \log M$

The significance score is computed based on how surprising the excluded maximum value is. I follow the approach of the authors to compute this:

5. Compute errors for each value y of measure M:

$$epsilon = \hat{y} - y \tag{6}$$

$$epsilon = 2 * *(\log(a) + b \log(x)) - y \tag{7}$$

6. Fit errors to a Gaussian distribution $N(\mu, \sigma)$
7. Compute ϵ_{ymax}:

$$\epsilon_{ymax} = \hat{y_{max}} - y_{max} \tag{8}$$

$$\epsilon_{ymax} = 2^{(\log(a) + b \log(1))} - y_{max} \tag{9}$$

$$\epsilon_{ymax} = a - y_{max} \tag{10}$$

8. Compute $p = P(\epsilon > \epsilon_{ymax} | N(\mu, \sigma))$
9. Compute $sig = 1 - p$

2.4.3 Gaussian Point Significance

While a power law distribution may be suitable for aggregate measures such as the count of papers per author, I found other measures to better fit a Gaussian distribution.

While the original authors do not describe how to calculate the significance score using a Gaussian distribution, I use a similar approach as used in the power law point significance score. The steps are as follows:

1. Exclude maximum value x_{max} from M
2. Fit a Gaussian distribution $N(\mu, \sigma)$ to $M \backslash x_{max}$
3. Compute $p = P(x > x_{max} | N(\mu, \sigma))$
4. Compute $\alpha = 1/|M|$
5. Compute $sig = max(0, 1 - p/\alpha)$

The intuition behind this approach is that it tests whether the maximum value belongs to the same normal distribution fit by the rest of the measure, or alternatively, whether it is an unlikely outlier. The Gaussian cumulative density function tells us that we would expect $1/|M|$ of data points to have values as extreme or more extreme than that with an $\alpha = 1/|M|$. If the measure M (including x_{max}) fit the Gaussian distribution perfectly, x_{max} should have a p value of $p = 1/|M|$, and therefore a significance score $sig = 1 - (1/|M|)/(1/|M|) = 0$. However, if the maximum value is surprisingly high and falls to the right of the distribution, the probability that it is drawn from $N(\mu, \sigma)$ is $p < \alpha$, causing a positive, non-zero significance score.

To test this approach, I applied the significance test to two randomly generated datasets. In Figure 1, I show that a random sample of 1000 points from a Gaussian distribution is given a small significance score of 0.103, since the maximum value falls closely to the expected value with an $\alpha = 1/1000$.

Second, I generate another series from a random Gaussian distribution but shift the maximum value to make it a right-tail outlier. Figure 2 shows that with this outlier, the series is appropriately given a high significance score, since the calculated $p = P(x > x_{max} | N(\mu, \sigma))$ is much smaller than the value $\alpha = 1/1000$.

Last, if x_{max} is smaller than the expected value of the top $1/|M|$ percentile, it is certainly not significant, and is given a significance score of zero since a negative significance score is meaningless.

2.4.4 Linear Point Significance

I further extend the author's work by adding a third point significance test. When exploring the data, I observed several linear trends: for example, a positive linear trend in the number of publications across years. While the linear shape significance tests capture insights about positive or negative

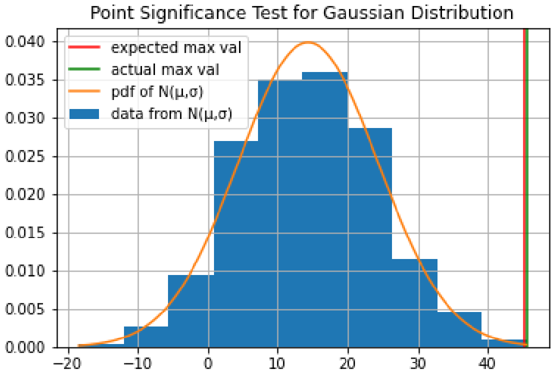

Figure 1: $[Sig(\Phi) = 0.103]$: Since the maximum point hardly exceeds the expected maximum value for $N(\mu, \sigma)$, the significance score is low.

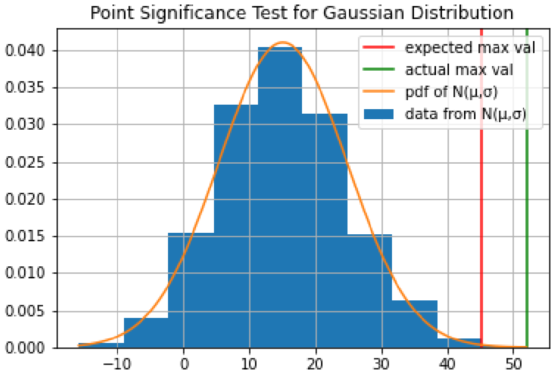

Figure 2: $[Sig(\Phi) = 0.926]$: A positive significance score indicates that the maximum point is surprisingly high, given $N(\mu, \sigma)$.

Type	**H0 Dist.**	**Possible Insights**
Shape	Linear	$\{$positive, negative$\}$ trend
Point	Power law	surprisingly high max value
Point	Gaussian	surprisingly $\{$high, low$\}$ value
Point	Linear	surprisingly $\{$high, low$\}$ value

Table 1: Shape and Point Significance Tests Used

trends, a linear point significance test captures whether a particular point stands out: for example, whether there was a surprising dip or peak in papers published among an otherwise steady linear trend. For example, Figure 6 shows that year 2014 had a surprisingly high count of publications given a positive linear trend.

The linear point significance test uses the same approach as the previously described tests:

1. Fit ordinal column x and measure y to the linear model $y = mx + b$ using least squares regression
2. Compute predicted errors ϵ:

$$\epsilon = \hat{y} - y \tag{11}$$

$$\epsilon = (mx + b) - y \tag{12}$$

3. Compute the Gaussian point significance test using errors ϵ described in Section 2.4.3; the significant point is the one with the maximum magnitude error
4. Compute the goodness of fit term r^2 by fitting X and measure y to a linear model using least squares regression, excluding potential outlier $y_{max-error}$
5. Weigh the significance score by the linear model's goodness of fit:

$$sig = r^2 * sig_{errors} \tag{13}$$

The intuition behind this approach is that linear regression assumes normality of residual errors. If a particular error falls improbably far beyond the expected distribution of errors, it is likely a point of interest. Last, weighing the significance score by the goodness of fit will discount insights with poorly fit linear trends; since the insight that "*point $y_{max-error}$ stands out from linear trend across ordinal dimension X*" supposes the relation of a linear trend of $Y \backslash y_{max-error}$, whether this trend holds should factor in to the insight score.

2.5. Data Preprocessing

In Section 3.1, I show the top insights extracted from two DBLP datasets: papers and collaborators. Each dataset is created by exporting a CSV file from relevant tables joined in Postgres.

To prepare the papers dataset, I join `Papers` and `Venues` in Postgres, selecting all relevant dimension columns and filtering between years 1990 and 2015. I chose to filter between 1990 and 2015 because almost 95% of papers are published in this range, and I was interested in insights during the years I have lived. I exclude year 2016 because the DBLP dataset is incomplete for this year. The resulting dataset contains 2,991,406 rows, each representing a paper.

To prepare the collaborators dataset, I joined the `Paperauths`, `Papers` and `Venues` relations keeping the paperid, authid, and year columns. As before, I filter between 1990 and 2015; the resulting dataset has 8,704,778 rows, representing 96.7% of the original dataset size.

3. Experimental Results

I use two approaches to analyze my implementation and the results I obtain. First, I explore the DBLP datasets to visually confirm top insights and inspect whether they are meaningful and valid. Second, I create artificial data distributions that either do or do not contain significant points or trends, and test whether my implementation gives them an appropriately high or low significance score. While Section 2.4.3 visualizes the tests for the Gaussian point significance functions, all four significance tests are tested in the unit test suite `tests/test_sigtests.py`.

3.1. Insights from the Papers Dataset

In this section, I analyze the top insights I extract from the DBLP papers and author collaborators dataset. Tables 2 through 5 show the list of top-10 insights from the papers dataset for depths 1 and 2, and Tables 6 through 9 show the top insights of depth 1 and 2 extracted from the collaborators dataset.

As intended, the top-10 insights of the papers dataset contain a mix of point and shape insights captured using different significance tests, suggesting that the implementation achieves a balance between the different insight types and significance tests.

In the following figures, I manually recreate a variety of the top insights on the papers dataset, highlighting the significant point or trend. Visually, these results verify that the top insights are, in fact, insightful, and that the implementation is working properly.

3.1.1 Selected Insights

Among the top insights, Figure 3 visualizes the point insight that the publication venue ‘CoRR’ has, by far, the greatest number of publications compared to the average number of publications per venue.

Figure 4 shows a clear linear trend in the rank of the total count of papers across years, meaning that the number of publications per year is increasing.

In Figure 5, the point insight highlighted shows that the vast majority of publications have dimension value = ”, meaning that they are non-school publication venues.

Figure 6 shows a point insight using the linear point significance test, which assumes a linear relation across an ordinal dimension and looks for surprisingly high or low points. As intended, it finds the point with the greatest deviation from the linear trend. Unsurprisingly, since the deviation is not large, this insight has a lower insight score ($\mathbb{S} = .23$) than the more apparent insights in the previous figures.

The top-10 depth-1 insights appeared to be accurate, but of limited insight. Of the top 10 insights, eight of the ten revealed the positive trend of the number of publications per year in different subspaces of the data. The top insight, however, was that the majority of publications belong to non-school publication venues (those with domain `school=''`). Confirming the findings of the author, depth-1 insights are of limited interest.

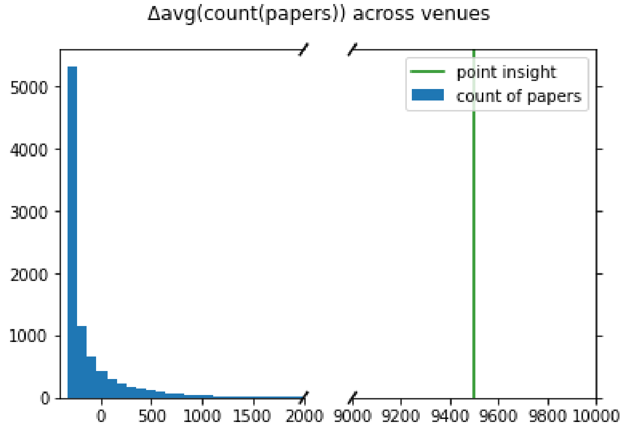

Figure 3: [$\mathbb{S} = 1.00$] Publication venue ‘CoRR’ has more papers than average when considering the full dataset.

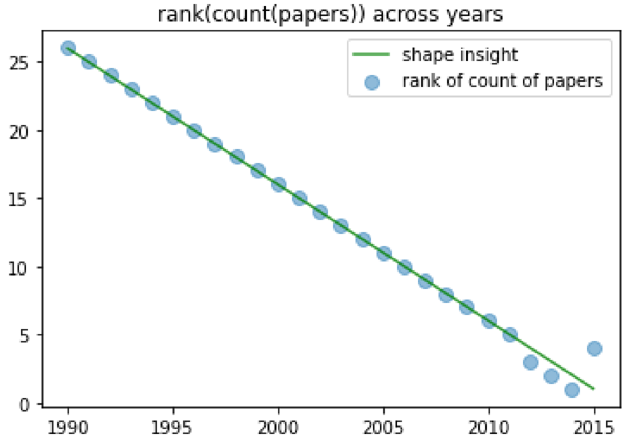

Figure 4: [$\mathbb{S} = .99$] The rank of count of papers between years has a downward trend across years.

3.2. Insights from the Collaborators Dataset

To evaluate the insights on the Collaborators dataset, I manually recreate selected top insights. The full lists of top-10 insights for depth-1 and 2 extraction are show in Tables 2 to 9.

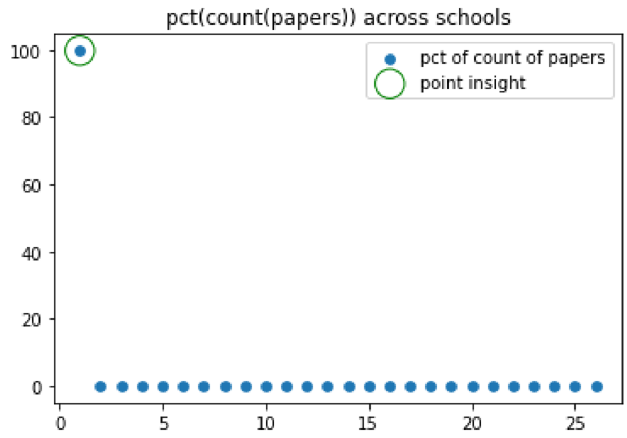

Figure 5: $[\mathbb{S} = 1.00]$ Non-school venues have the highest percent of publications compared to other schools.

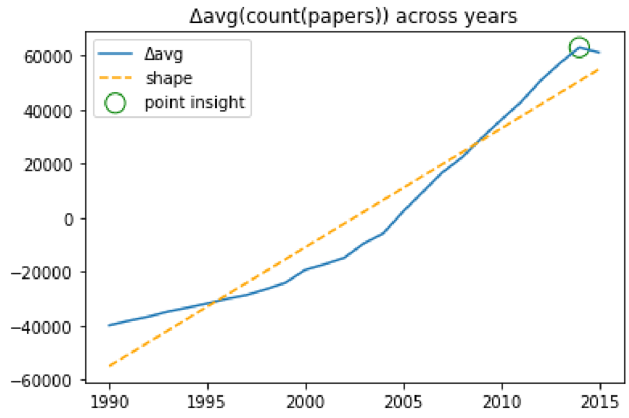

Figure 6: $[\mathbb{S} = 0.23]$ Non-school venues have the highest percent of publications compared to other schools, when considering just journal article venues ($venue\_type = 0$).

3.2.1 Selected Insights

Figure 7 shows the top insight found for the depth-2 insight extraction. Applying the composite extractor $[(count, paperauths), (\Delta_{avg}, authid)]$ over sibling group $SG(\{\}, paperid)$ creates a result set with the number of authors for each paper. So, the point insight found tells us which paper had the highest number of authors (119) exceeding the average (almost 3).

Figure 8 shows the second top insight found for the depth-2 insight extraction. Applying the composite extractor $[(count, paperauths), (\Delta_{avg}, paperid)]$ over sibling group $SG(\{\}, authid)$ provides a result set with the number of papers for each author. The highly significant point insight highlights the highest-producing author, who published 1156 more papers than the average author.

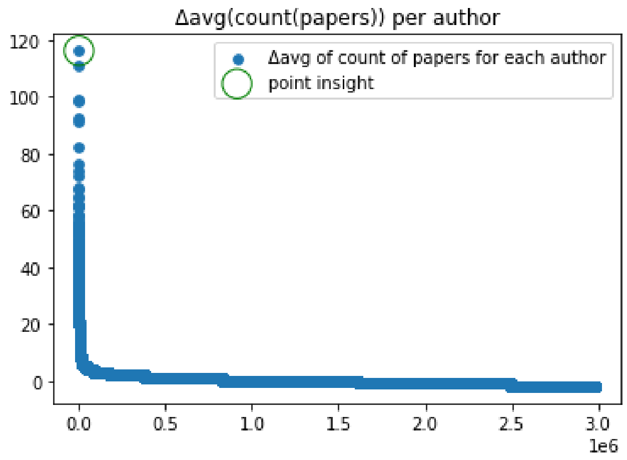

Figure 7: $[\mathbb{S} = 1.00]$ Paperid '1364275' had the highest number of coauthors, with 116.08 beyond the average.

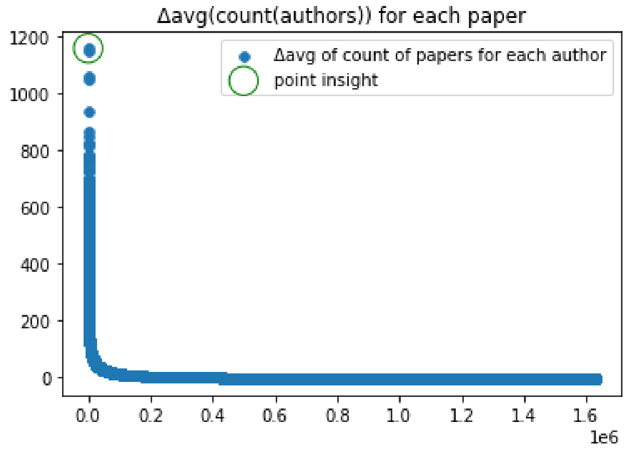

Figure 8: $[\mathbb{S} = 1.00]$ Authorid '1364275' had the highest number of publications, with 1155.66 beyond the average.

4. Summary

The system implemented by the authors of *Extracting Top-K Insights from Multidimensional Data* demonstrates how insights can be automatically extracted from datasets. The main contributions of the authors are:

1. proposing a meaningful insight score that captures the impact of an insight and its statistical significance
2. formulating significance scores to provide comparability between different insight types
3. implementing a system that efficiently enumerates subspaces, dimensions, and aggregate measures to uncover

trends that are not visible in the original dataset

In my implementation, I develop the core of the authors' system: the insight extraction engine, which consists of **Algorithm 1**, **Algorithm 2**, and the **impact** and **significance** scoring functions. I implement a couple of the authors' optimizations, sufficient to extract top depth 1 or depth 2 insights from the DBLP datasets described in Section 3 in a matter of minutes.

I develop two new significance scores to uncover point insights using the null hypothesis that data is normally distributed, or that it is distributed linearly across an ordinal dimension.

Last, I empirically test my implementation and significance tests in two ways. First, I generate samples from known distributions to ensure that the significance tests provide reasonable scores. Second, I recreate top insights from the papers and collaborators datasets, visually inspecting whether the automated insights are valid and insightful.

5. Team Contributions

This work was completed individually.

References

[1] B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang. Extracting top-k insights from multi-dimensional data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1509–1524, 2017. 1

id	rank	score	sig	impact	type	SG(S,D)	CE	insight	H0
1727	1	1.000	1.000	1.000	point	SG({},school)	[(’count’), (’pct’, ’school’)]	maximum point 99.77	powerlaw
1641	2	1.000	1.000	1.000	point	SG({},venue_name)	[(’count’), (’delta_avg’, ’venue_name’)]	maximum point 94634.48	normal
1731	3	1.000	1.000	1.000	point	SG({},school)	[(’count’), (’delta_avg’, ’school’)]	maximum point 2869567.92	normal
1679	4	0.998	1.000	0.998	point	SG({’school’: ”},venue_name)	[(’count’), (’delta_avg’, ’venue_name’)]	maximum point 94635.17	normal
1690	5	0.992	0.992	1.000	shape	SG({},year)	[(’count’), (’rank’, ’year’)]	negative slope -1.00	linear_shape
1692	6	0.990	0.992	0.998	shape	SG({’school’: ”},year)	[(’count’), (’rank’, ’year’)]	negative slope -1.00	linear_shape
1710	7	0.951	0.951	1.000	shape	SG({},year)	[(’count’), (’pct’, ’year’)]	positive slope 0.33	linear_shape
1720	8	0.951	0.951	1.000	shape	SG({},year)	[(’count’), (’delta_avg’, ’year’)]	positive slope 9907.30	linear_shape
1712	9	0.949	0.951	0.998	shape	SG({’school’: ”},year)	[(’count’), (’pct’, ’year’)]	positive slope 0.33	linear_shape
1722	10	0.949	0.951	0.998	shape	SG({’school’: ”},year)	[(’count’), (’delta_avg’, ’year’)]	positive slope 9890.16	linear_shape

Table 2: Top-10 depth-2 Insights on the DBLP Papers dataset

id	rank	interpretation
1727	1	Aggregating pct of school of count over dividing dimension school,maximum point 99.77 stood out using powerlaw test, considering only the subspace with dimensions {}
1641	2	Aggregating delta_avg of venue_name of count over dividing dimension venue_name,maximum point 94634.48 stood out using normal test, considering only the subspace with dimensions {}
1731	3	Aggregating delta_avg of school of count over dividing dimension school,maximum point 2869567.92 stood out using normal test, considering only the subspace with dimensions {}
1679	4	Aggregating delta_avg of venue_name of count over dividing dimension venue_name,maximum point 94635.17 stood out using normal test, considering only the subspace with dimensions {’school’: ”}
1690	5	Aggregating rank of year of count over dividing dimension year,negative slope -1.00 stood out using linear_shape test, considering only the subspace with dimensions {}
1692	6	Aggregating rank of year of count over dividing dimension year,negative slope -1.00 stood out using linear_shape test, considering only the subspace with dimensions {’school’: ”}
1710	7	Aggregating pct of year of count over dividing dimension year,positive slope 0.33 stood out using linear_shape test, considering only the subspace with dimensions {}
1720	8	Aggregating delta_avg of year of count over dividing dimension year,positive slope 9907.30 stood out using linear_shape test, considering only the subspace with dimensions {}
1712	9	Aggregating pct of year of count over dividing dimension year,positive slope 0.33 stood out using linear_shape test, considering only the subspace with dimensions {’school’: ”}
1722	10	Aggregating delta_avg of year of count over dividing dimension year,positive slope 9890.16 stood out using linear_shape test, considering only the subspace with dimensions {’school’: ”}

Table 3: Generated interpretations of the top-10 depth-2 Insights of the DBLP Papers dataset.

rank	score	sig	impact	type	SG(S,D)	CE	insight	H0
1	1.000	1.000	1.000	point	SG({},school)	[(’count’)]	maximum point 2984622.00	powerlaw
2	0.951	0.951	1.000	shape	SG({},year)	[(’count’)]	positive slope 9907.30	linear_shape
3	0.949	0.951	0.998	shape	SG({’school’: ”},year)	[(’count’)]	positive slope 9890.16	linear_shape
4	0.523	0.932	0.561	shape	SG({’venue_type’: 1},year)	[(’count’)]	positive slope 5486.67	linear_shape
5	0.523	0.932	0.561	shape	SG({’school’: ”, ’venue_type’: 1},year)	[(’count’)]	positive slope 5486.67	linear_shape
6	0.523	0.932	0.561	shape	SG({’venue_type’: 1, ’school’: ”},year)	[(’count’)]	positive slope 5486.67	linear_shape
7	0.409	0.938	0.436	shape	SG({’venue_type’: 0, ’school’: ”},year)	[(’count’)]	positive slope 4403.49	linear_shape
8	0.409	0.938	0.436	shape	SG({’school’: ”, ’venue_type’: 0},year)	[(’count’)]	positive slope 4403.49	linear_shape
9	0.409	0.938	0.436	shape	SG({’venue_type’: 0},year)	[(’count’)]	positive slope 4403.49	linear_shape
10	0.056	1.000	0.056	point	SG({’year’: 2007},school)	[(’count’)]	maximum point 167670.00	powerlaw

Table 4: Top-10 depth-1 Insights on the DBLP Papers dataset.

id	rank	interpretation
1131	1	Aggregating count over dividing dimension school,maximum point 2984622.00 stood out using powerlaw test, considering only the subspace with dimensions {}
1013	2	Aggregating count over dividing dimension year,positive slope 9907.30 stood out using linear_shape test, considering only the subspace with dimensions {}
1133	3	Aggregating count over dividing dimension year,positive slope 9890.16 stood out using linear_shape test, considering only the subspace with dimensions {’school’: ”}
1249	4	Aggregating count over dividing dimension year,positive slope 5486.67 stood out using linear_shape test, considering only the subspace with dimensions {’venue_type’: 1}
1198	5	Aggregating count over dividing dimension year,positive slope 5486.67 stood out using linear_shape test, considering only the subspace with dimensions {’school’: ”, ’venue_type’: 1}
1276	6	Aggregating count over dividing dimension year,positive slope 5486.67 stood out using linear_shape test, considering only the subspace with dimensions {’venue_type’: 1, ’school’: ”}
1236	7	Aggregating count over dividing dimension year,positive slope 4403.49 stood out using linear_shape test, considering only the subspace with dimensions {’venue_type’: 0, ’school’: ”}
1185	8	Aggregating count over dividing dimension year,positive slope 4403.49 stood out using linear_shape test, considering only the subspace with dimensions {’school’: ”, ’venue_type’: 0}
1213	9	Aggregating count over dividing dimension year,positive slope 4403.49 stood out using linear_shape test, considering only the subspace with dimensions {’venue_type’: 0}
1085	10	Aggregating count over dividing dimension school,maximum point 167670.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2007}

Table 5: Generated interpretations of the top-10 depth-1 Insights of the DBLP Papers dataset.

id	rank	score	sig	impact	type	SG(S,D)	CE	insight	H0
1045	1	1.000	1.000	1.000	point	SG({},paperid)	[('count'), ('delta_avg', 'paperid')]	maximum point 116.08	normal
1078	2	1.000	1.000	1.000	point	SG({},authid)	[('count'), ('delta_avg', 'authid')]	maximum point 1155.66	normal
1086	3	0.996	0.996	1.000	shape	SG({},year)	[('count'), ('rank', 'year')]	negative slope -1.00	linear_shape
1098	4	0.938	0.938	1.000	shape	SG({},year)	[('count'), ('pct', 'year')]	positive slope 0.38	linear_shape
1102	5	0.938	0.938	1.000	shape	SG({},year)	[('count'), ('delta_avg', 'year')]	positive slope 33116.28	linear_shape
1095	6	0.090	1.000	0.090	point	SG({'year': 2014},authid)	[('count'), ('delta_prev', 'year')]	maximum point 135.00	normal
1054	7	0.090	1.000	0.090	point	SG({'year': 2014},paperid)	[('count'), ('delta_avg', 'paperid')]	maximum point 97.70	normal
1083	8	0.090	1.000	0.090	point	SG({'year': 2014},authid)	[('count'), ('delta_avg', 'authid')]	maximum point 197.86	normal
1103	9	0.090	1.000	0.090	point	SG({'year': 2014},authid)	[('count'), ('delta_avg', 'year')]	maximum point 152.12	normal
1089	10	0.090	1.000	0.090	point	SG({'year': 2014},authid)	[('count'), ('rank', 'year')]	maximum point 26.00	normal

Table 6: Top-10 depth-2 Insights on the DBLP Collaborators dataset.

id	rank	interpretation
1045	1	Aggregating delta_avg of paperid of count over dividing dimension paperid,maximum point 116.08 stood out using normal test, considering only the subspace with dimensions {}
1078	2	Aggregating delta_avg of authid of count over dividing dimension authid,maximum point 1155.66 stood out using normal test, considering only the subspace with dimensions {}
1086	3	Aggregating rank of year of count over dividing dimension year,negative slope -1.00 stood out using linear_shape test, considering only the subspace with dimensions {}
1098	4	Aggregating pct of year of count over dividing dimension year,positive slope 0.38 stood out using linear_shape test, considering only the subspace with dimensions {}
1102	5	Aggregating delta_avg of year of count over dividing dimension year,positive slope 33116.28 stood out using linear_shape test, considering only the subspace with dimensions {}
1095	6	Aggregating delta_prev of year of count over dividing dimension authid,maximum point 135.00 stood out using normal test, considering only the subspace with dimensions {'year': 2014}
1054	7	Aggregating delta_avg of paperid of count over dividing dimension paperid,maximum point 97.70 stood out using normal test, considering only the subspace with dimensions {'year': 2014}
1083	8	Aggregating delta_avg of authid of count over dividing dimension authid,maximum point 197.86 stood out using normal test, considering only the subspace with dimensions {'year': 2014}
1103	9	Aggregating delta_avg of year of count over dividing dimension authid,maximum point 152.12 stood out using normal test, considering only the subspace with dimensions {'year': 2014}
1089	10	Aggregating rank of year of count over dividing dimension authid,maximum point 26.00 stood out using normal test, considering only the subspace with dimensions {'year': 2014}

Table 7: Generated interpretations of the top-10 depth-2 Insights of the DBLP Collaborators dataset.

id	rank	score	sig	impact	type	SG(S,D)	CE	insight	H0
1004	1	0.94	0.94	1.00	shape	SG({},year)	[(’count’)]	positive slope 33116.28	linear_shape
1001	2	0.00	0.00	1.00	point	SG({},paperid)	[(’count’)]	maximum point 119.00	powerlaw
1002	3	0.00	0.00	1.00	point	SG({},authid)	[(’count’)]	maximum point 1161.00	powerlaw
1003	4	0.00	0.00	1.00	point	SG({},year)	[(’count’)]	year {2014} surprisingly high at {782072.00}	linear_point
1008	5	0.00	0.00	0.08	point	SG({’year’: 2011},paperid)	[(’count’)]	maximum point 77.00	powerlaw
1005	6	0.00	0.00	0.02	point	SG({’year’: 2001},authid)	[(’count’)]	maximum point 52.00	powerlaw
1006	7	0.00	0.00	0.02	point	SG({’year’: 2001},paperid)	[(’count’)]	maximum point 52.00	powerlaw
1007	8	0.00	0.00	0.08	point	SG({’year’: 2011},authid)	[(’count’)]	maximum point 112.00	powerlaw
1009	9	0.00	0.00	0.05	point	SG({’year’: 2006},authid)	[(’count’)]	maximum point 92.00	powerlaw
1010	10	0.00	0.00	0.05	point	SG({’year’: 2006},paperid)	[(’count’)]	maximum point 119.00	powerlaw

Table 8: Top-10 depth-1 Insights on the DBLP Collaborators dataset.

id	rank	interpretation
1004	1	Aggregating count over dividing dimension year,positive slope 33116.28 stood out using linear_shape test, considering only the subspace with dimensions {}
1001	2	Aggregating count over dividing dimension paperid,maximum point 119.00 stood out using powerlaw test, considering only the subspace with dimensions {}
1002	3	Aggregating count over dividing dimension authid,maximum point 1161.00 stood out using powerlaw test, considering only the subspace with dimensions {}
1003	4	Aggregating count over dividing dimension year,year {2014} surprisingly high at {782072.00} stood out using linear_point test, considering only the subspace with dimensions {}
1008	5	Aggregating count over dividing dimension paperid,maximum point 77.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2011}
1005	6	Aggregating count over dividing dimension authid,maximum point 52.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2001}
1006	7	Aggregating count over dividing dimension paperid,maximum point 52.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2001}
1007	8	Aggregating count over dividing dimension authid,maximum point 112.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2011}
1009	9	Aggregating count over dividing dimension authid,maximum point 92.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2006}
1010	10	Aggregating count over dividing dimension paperid,maximum point 119.00 stood out using powerlaw test, considering only the subspace with dimensions {’year’: 2006}

Table 9: Generated interpretations of the top-10 depth-1 Insights of the DBLP Collaborators dataset.