Data Dictionary for the `cocit` Database: Supplemental Document for *Frequently Cocited Publications: Features and Kinetics*

James R. Bradley

July 7, 2020

Abstract

The `cocit MySQL` database contains the results of the numerical analysis described in Devarakonda, Bradley, Korobskiy, Warnow, and Chacko (2020). This document provides data dictionaries for tables in the `cocit` database.

1 Introduction

Power law and lognormal distribution were fit to the right tail of the frequency distribution for cocitations in Devarakonda et al. (2020). This accompanying document provides a data dictionary for a MySQL database named `cocit`, which contains the results of those statistical computations (Section 2).

2 Data Dictionary for Tables

Table 1 describes the fields in the `results_ln` table in the `cocit` MySQL database, which contains the results for a lognormal fit to various extremities of the right tail of the cocitation frequency distribution. The fields in the `results_pl` table are similar, except that they refer to the results for a power law distribution. Both of these tables use a unique integer identifier for the θ interval, which refers to a foreign key in the `cocit.t_o` table. In turn, Table 2 describes the fields of the `cocit.t_o` table, which define the lower and upper bounds on the θ intervals. The values in `cocit.t_o` are shown in Table 3.

Table results_ln		
Field	Data Type	Description
time_stamp	DATETIME	Time when distributional fit was computed.
dist	VARCHAR(45)	Indicates the distributional form fit to the data. This field is redundant due to revisions made to the database tables.
cutoff	INT	Right tail cutoff.
t_o_id	INT	Id for θ interval (foreign key for primary key t_o.id)
connected	VARCHAR(10)	Indicates whether the results are for connected articles (True), unconnected articles (False), or all articles (all).
num_pts	FLOAT	Number of data points analyzed in this computation.
obs_max	INT	The maximum frequency of co-citation.
mean	DOUBLE	Mean frequency of co-citation data.
std_dev	DOUBLE	Standard deviation of co-citation frequency data.
mean_fit	DOUBLE	Mean of the fit distribution.
std_dev_fit	DOUBLE	Standard deviation of the fit distribution.
mean_norm	DOUBLE	Mean of the normal distribution underlying the fit lognormal distribution.
std_dev_norm	DOUBLE	Standard deviation of the normal distribution underlying the fit lognormal distribution.
k_samp	DOUBLE	Maximum difference in cumulative distributions between data and fit distribution.
k90	DOUBLE	90th percentile of difference between cumulative distributions of data and the fit distribution (related to Kolmogorov-Smirnov test).
k95	DOUBLE	95th percentile of difference between cumulative distributions of data and the fit distribution (related to Kolmogorov-Smirnov test).
ks_p	DOUBLE	Kolmogorov-Smirnov p-value for distributional fit. Computed by simulating 100 maximum differences between two lognormal fit distributions.
kl1	DOUBLE	One of two (asymmetric) Kullback-Leibler Divergence computations.
kl2	DOUBLE	One of two (asymmetric) Kullback-Leibler Divergence computations.
chi2_10	DOUBLE	Chi-squared test p-value with data binned with a minimum expected number of observations of 10 co-citation instances.
chi2_20	DOUBLE	Same as above with a minimum expected number of co-citations per bin of 20.
chi2_50	DOUBLE	Same as above with a minimum expected number of co-citations per bin of 50.
chi2_70	DOUBLE	Same as above with a minimum expected number of co-citations per bin of .
chi2_100	DOUBLE	Not used, as indicated by a value of 99.999999999.

Table 1: Data Fields for Table results2ln, which contains results for the fitting of lognormal distributions (indicated by ln) to the data.

Table `t_o`		
Field	Data Type	Description
`id`	`INT`	Unique identifier for the interval
`lo`	`FLOAT`	lower bound of the interval
`hi`	`FLOAT`	upper bound of the interval

Table 2: Data Fields for Table `t_o`

Table `t_o`		
`id`	`lo`	`hi`
0	0.0	0.2
1	0.2	0.4
2	0.4	0.6
3	0.6	0.8
4	0.8	1.0

Table 3: Values for Table `t_o`

References

Devarakonda, S., Bradley, J. R., Korobskiy, D., Warnow, T., & Chacko, G. (2020). Frequently cocited publications: Features and kinetics. *Quantitative Social Science*, *Forthcoming*. Retrieved from https://www.mitpressjournals.org/doi/abs/10.1162/qss_a_00075 doi: 10.1162/qss_a_00075