LexisNexis Appeals Data Review
Legal Topics

Version 2.0, April 7, 2020

Tom Balmat, Duke University

To date, two primary data sets have been received from LexisNexis (LN): set A (3/13/2019) and set B (12/3/2019). The following summarizes results of several tests of integrity of legal topics assigned to case records in these data sets. Tests include:

- Verification of topic ID format
- Identification of topic IDs that do not appear in the master topic table
- Verification that legal topics are not duplicated on any case
- Comparison, between data sets A and B, of the distribution of legal topics by year
- Comparison, between data sets A and B, of the distribution of legal topics by ID
- Evaluation of topic IDs with a difference, between data sets A and B, in cases assigned
- Evaluation of cases with a difference, between data sets A and B, in legal topics assigned

Legal topic ID format

Legal topic ID codes are Global Unique Identifiers (GUID) assigned by LN. A valid GUID is a 128 bit binary integer with character representation as a string of 32 hexadecimal digits. Valid hexadecimal digits are 0-9, A-F. An example topic ID in hexadecimal is 0DFE2750D1294DBA936B6AAECB79C77F. Observations:

- There are 44,899 topics in the master table, each a valid hexadecimal GUID
- 16,134 distinct legal topic IDs are assigned to cases in data set A, each a valid hexadecimal GUID
- 14,755 distinct legal topic IDs are assigned to cases in data set B, each a valid hexadecimal GUID

Topic IDs that do not appear in the master topic table

Table 1 lists topic IDs that are assigned to cases, but do not appear in the master topics table. Note that, although frequencies differ, the set of unrecognized IDs is the same for both data sets, with the exception of ID C15FD5D77534430BBE65B9AE681A5282 appearing in data set B only.

Table 1: Data sets A and B legal topic IDs that do not appear in the master topic table

Topic ID	Length	n(Set A)	n(Set B)
024C354AF46A4A2FA751B8986DD10471	32	20	22
1FD8BB9677D9467C9BADEB256362B109	32	1	1
25236D06A3164C7E8E6A590AA2BE6D48	32	23	29
4F449B05E21948F09AF3AF140AAF1B87	32	1	5
82FA08A9928B4CF0893760370FF80E2E	32	1	6
8DDF4DB4FE4C487A97DCEED7E7F9D696	32	1	1
C0BE07E8370E4C76AB1D1976E848D4F2	32	3	25
C15FD5D77534430BBE65B9AE681A5282	32	0	2
DDF6E78E4E724C78865E235054136851	32	3	15
DFBA247750FF42F4AA2AC464EBCCE7DE	32	1	3

Duplicate topic IDs by case

Duplicate case LNI, legal topic ID combinations have been observed in the records provided by LN. Unique combinations of LNI and legal topic ID were imported into the SQL database.

Distribution of legal topics assigned by year

Figure 1 shows the distribution, by year, of the total number of topics assigned to all cases. There is an apparent significant difference in the number of topics assigned between sets A and B for years 1974 through 2004.

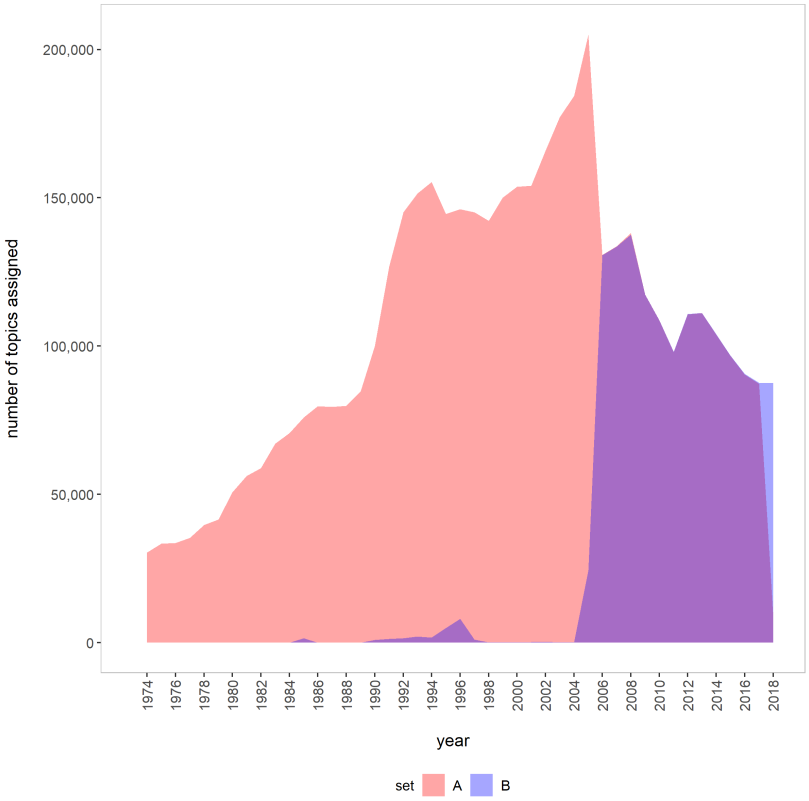

Figure 1: Distribution of total number of topics assigned by data set and year

Distribution of legal topics assigned by case

Figure 2 shows the distribution of cases by the number of topics assigned. There is an apparent difference in assignment frequency, with cases in data set A generally having more topics assigned than those in data set B.

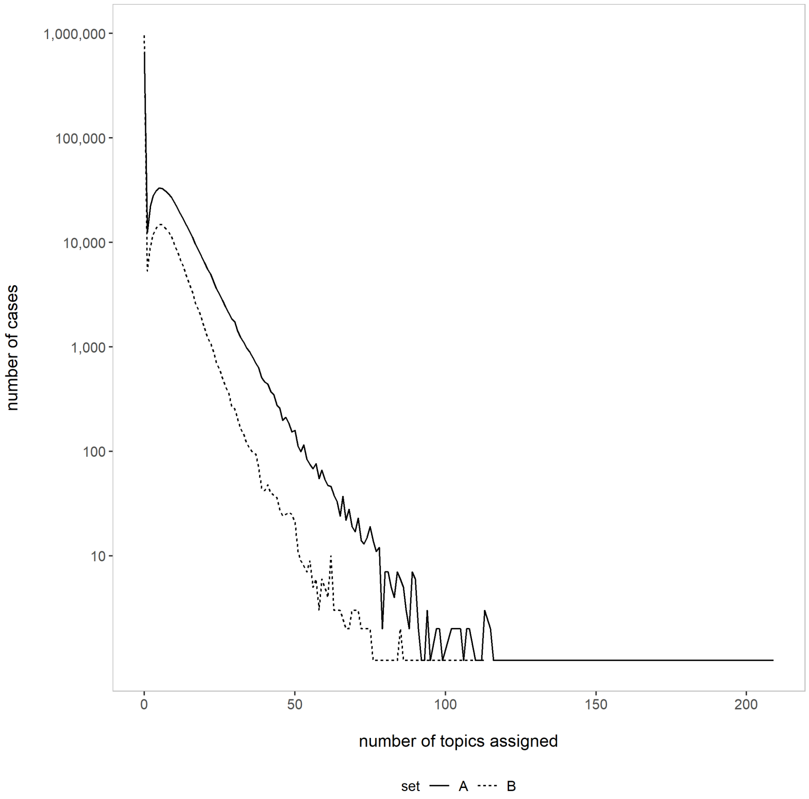

Figure 2: Distribution of cases by data set and number of topics assigned

Topics with a difference in cases assigned between data sets B and A

Figure 3 shows the distribution of the difference, between data sets B and A, in cases assigned by topic. There is an apparent significant difference in the number of cases assigned a given topic between the two data sets. Differences are number assigned in set B minus number assigned in set A, indicating that topics are typically assigned with lesser frequency in set B than they are in set A. Note that topic "US Legal Taxonomy>Civil Procedure>Appeals>Standards of Review>De Novo Review" has maximum assignment frequency in both data sets, with 34,670 and 86,582 case assignments, respectively (B and A). The x-axis was truncated at -2,500. All case frequencies corresponding to topic assignment with differences less than -2,500 are near 0.

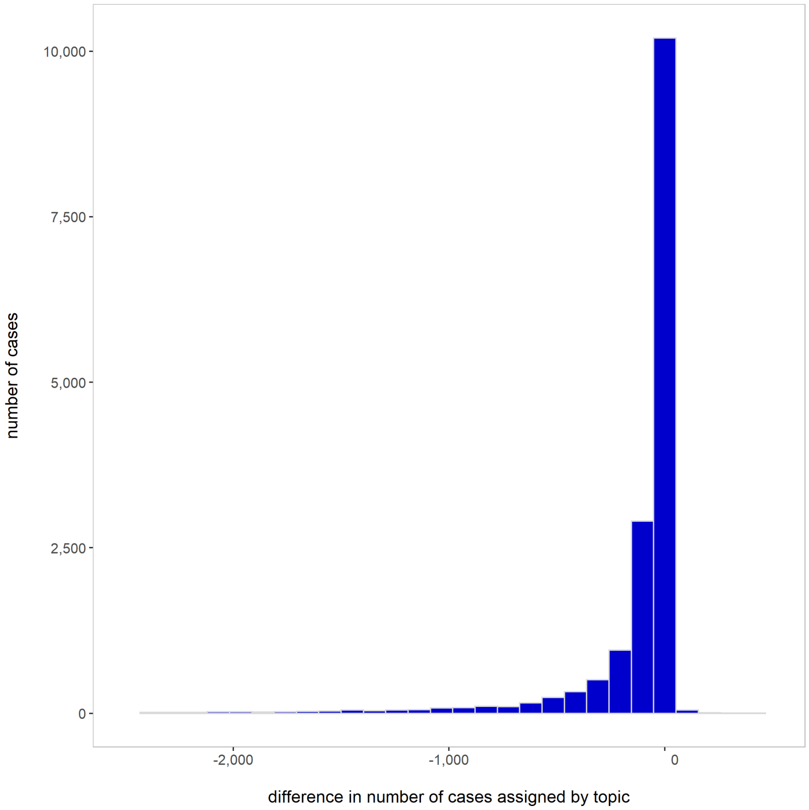

Figure 3: Distribution of difference, between data sets B and A, in cases assigned by topic

Cases with a difference in number of topics assigned between data sets B and A

Figure 4 shows the distribution of the difference, between data sets B and A, in number of topics assigned by case. Differences are number of topics assigned in set B minus the number assigned in set A, indicating that cases typically have fewer topics assigned in set B than they do in set A.

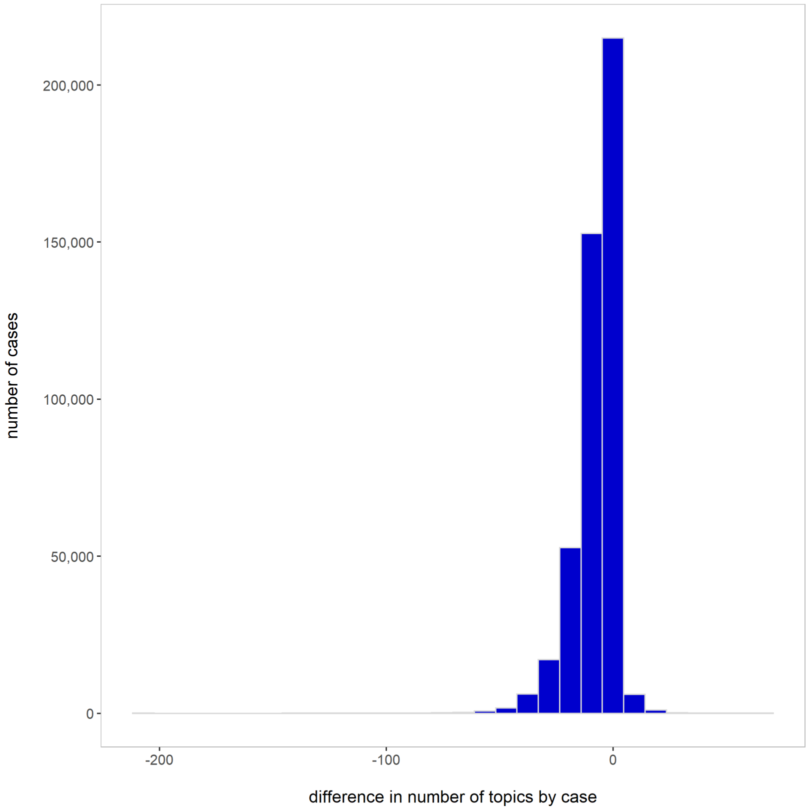

Figure 4: Distribution of difference, between data sets B and A, in cases assigned by case