LexisNexis (LN) Sample Appeal Opinion Text Analysis

Identification of Word Pair Relationships

Tom Balmat

January 8, 2019

The following is a brief study of text and word relationships in a sample of U.S. Courts of Appeals case opinions supplied to Duke University by LexisNexis. Methods presented might be useful in classifying cases by word patterns, identifying characteristic phrases, and measuring correlation of words in text.

1 Data

The complete text of 212 appeals case decisions were supplied, as a sample, by LexisNexis (LN) in XML format. Important XML data elements include case name, citations, court name, filing or decision date, judge panel, representation, case summary, concurring, dissenting, and majority opinion text, and various LN designed meta data tags. XML elements were parsed and uploaded into a SQL database for querying. Filing or decision dates span years 1929 to 2018. Case distribution by year of decision, if available, otherwise year of filing, appears in table 1. Case categories as supplied were incomplete, so alternatives were assigned by the team after a review of each case. Categories and frequency of cases supplied appear in table 2. Distribution of opinion text by concurring, dissenting, or majority appears in table 3.

Table 1: LexisNexis sample case distribution by year of decision date

Year	Cases	Year	Cases	Year	Cases	Year	Cases
1929	1	1965	6	1984	1	2001	1
1930	1	1966	4	1985	2	2002	2
1931	1	1967	9	1986	3	2003	1
1932	1	1968	10	1987	2	2004	2
1934	1	1969	4	1988	1	2005	2
1937	1	1970	15	1989	2	2006	3
1938	2	1971	12	1990	1	2007	3
1939	3	1972	14	1991	3	2009	1
1954	1	1973	3	1992	3	2010	1
1955	1	1974	2	1993	3	2011	3
1957	5	1975	3	1994	4	2012	1
1958	7	1976	2	1995	2	2013	1
1959	4	1977	1	1996	1	2014	3
1961	5	1978	1	1997	2	2015	1
1962	2	1981	1	1998	6	2016	2
1963	6	1982	1	1999	4	2017	3
1964	6	1983	1	2000	4	2018	1

Table 2: LexisNexis sample case distribution by category

Category	Cases
Commercial	45
Criminal/Punitive	95
Insurance/Injury	25
Property	31
Other	16

Table 3: LexisNexis sample case distribution by opinion type

Opinion Type	Opinions
Concur	16
Dissent	17
Majority	212

2 Opinion Text Analysis

2.1 Length of Text in Words

The opinion sections of the LN supplied case documents contain more text than other sections. Average word count for all opinions is 2,391.[1] Figure 1 shows the distribution of word count in opinion text by case category and opinion type. Each category exhibits a right skewed and bi-modal distribution, with a tendency toward opinions of length less than approximately 5,000 words and a distinct sub-population of opinions of length between 5,000 and 15,000 words. The colored regions within category represent the proportion of opinions by type (concurring, dissenting, majority). Proportions are stacked, so that the height of a graph at a given word count is the total proportion of opinions with that count in the corresponding category. Figure 2 shows the distribution of opinions within each case category, opinion type combination. In our sample, the mult-modal effect appears most pronounced in dissenting opinions. Distributions for some combinations do not appear (property, dissenting, for instance) due to having insufficient instances to compute density.

[1] A small number of header labels, such as I., A., and b. are included in word counts.

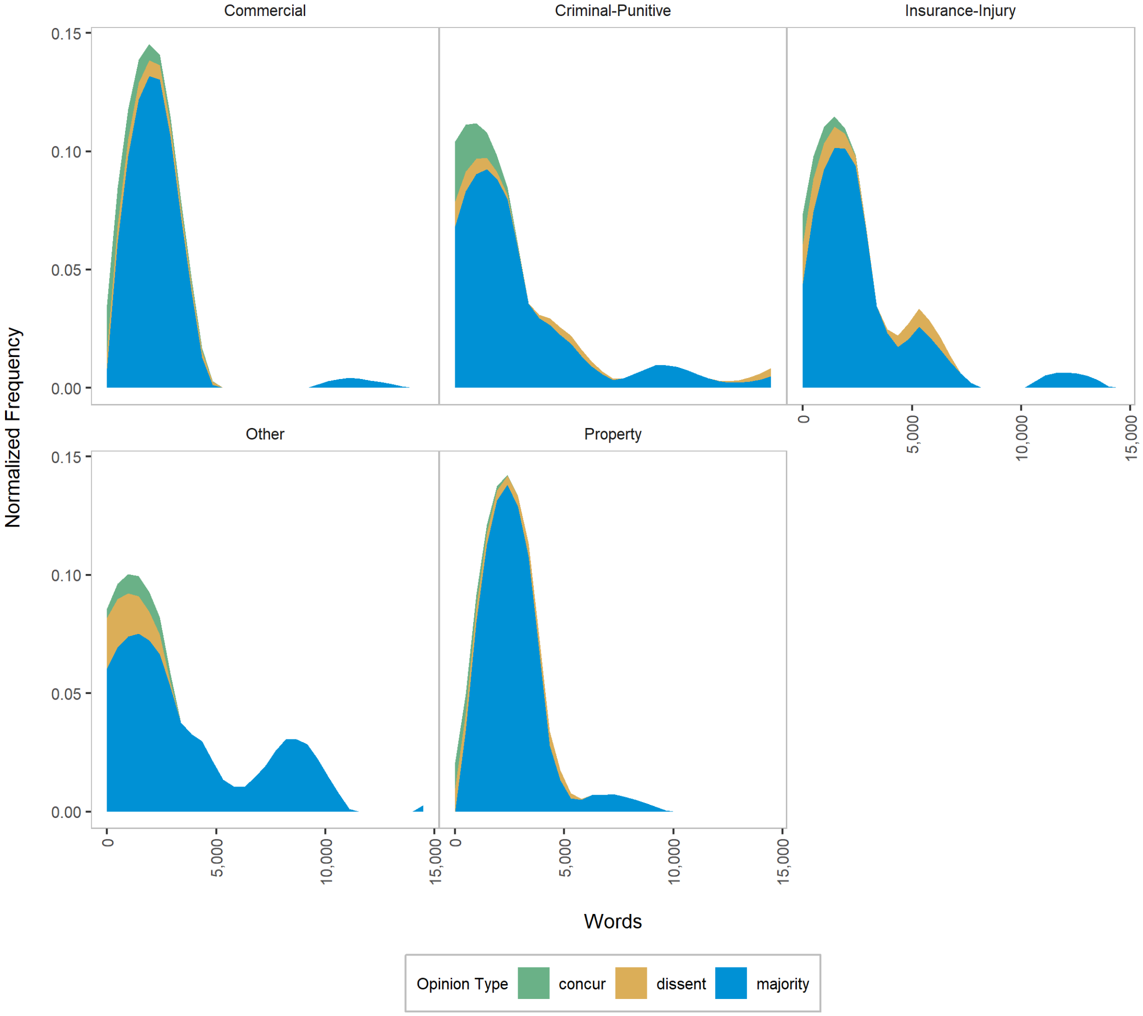


Figure 1: Distribution of word count by case category and opinion type

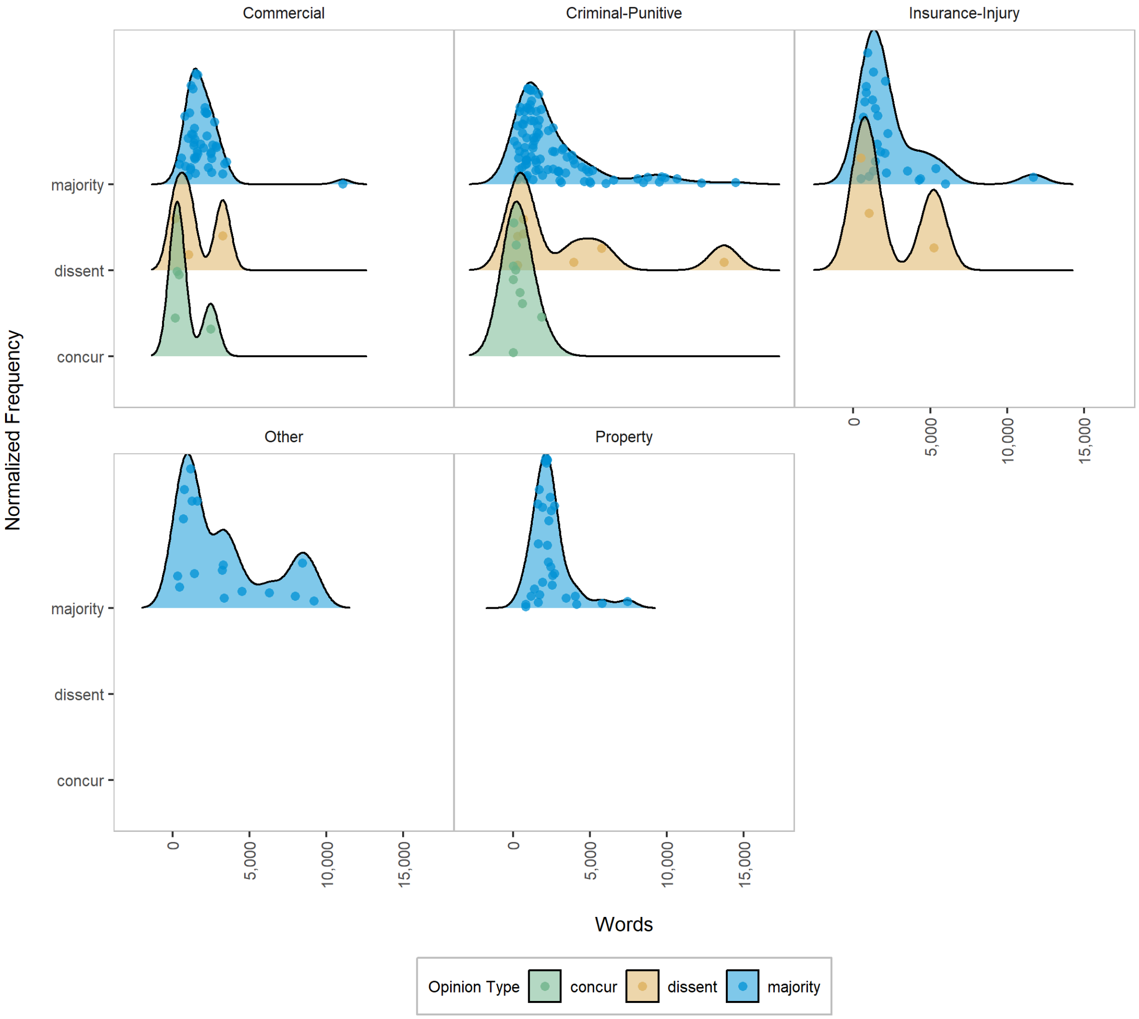


Figure 2: Density of word count by case category and opinion type

2.2 High Frequency Words

It is assumed that words appearing in opinion text are chosen to characterize aspects of an appeal, its merits, legal basis, and practical concerns. Given this, the proportion of occurrence of particular words might characterize a class of opinions or cases. Figures 3 and 4 plot proportions of high frequency words by case category and opinion type, respectively.[2] They identify high-use words that might aid in choosing one particular classification over another. Figures 5 through 10 compare word occurrence proportions between pairs of case classes and pairs of opinion types. Words with points near the reference line of slope 1 appear with similar frequency in both subsets, points distant from the reference line indicate significantly different rates of occurrence in the two subsets. These conditions may aid in identifying case classes with similar or dissimilar word usage. Example words with similar proportions are (from figure 5) the words *case* and *law*. Example dissimilar words (from the same plot) are *business*, *company*, *counsel*, and *jury*. Perhaps of interest is an apparent closer agreement in proportions for words appearing in property and commercial cases (figure 7) than in other combinations. An extension of this method would be to compute and plot proportion appearance of key multi-word phrases.

[2] Pronouns, conjunctive words, names of people, cities, and states (to the degree possible), and several uninformative words (such as all, this, may, etc.) have been omitted from the text prior to analysis.

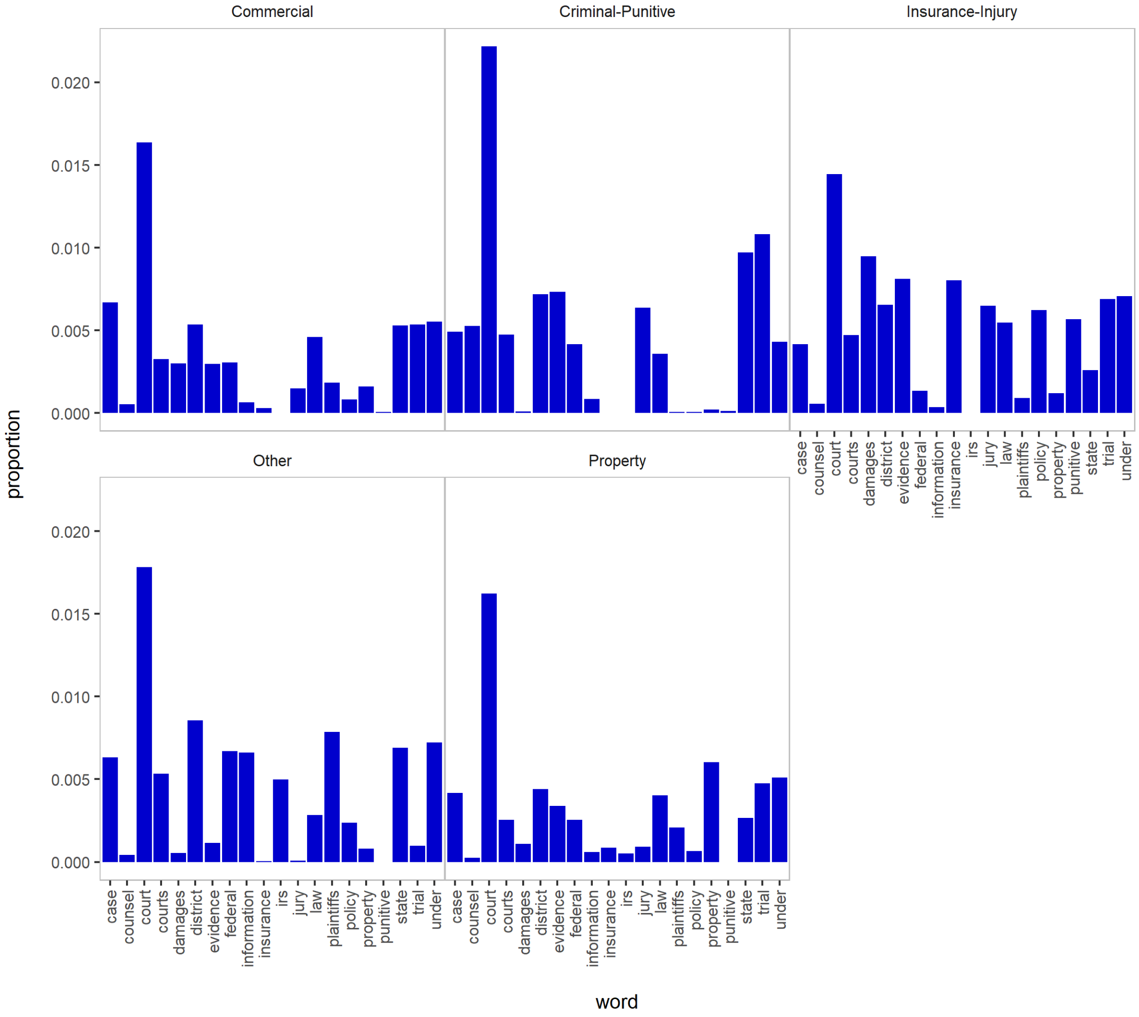


Figure 3: Distribution of twenty most frequent appearing words by case class

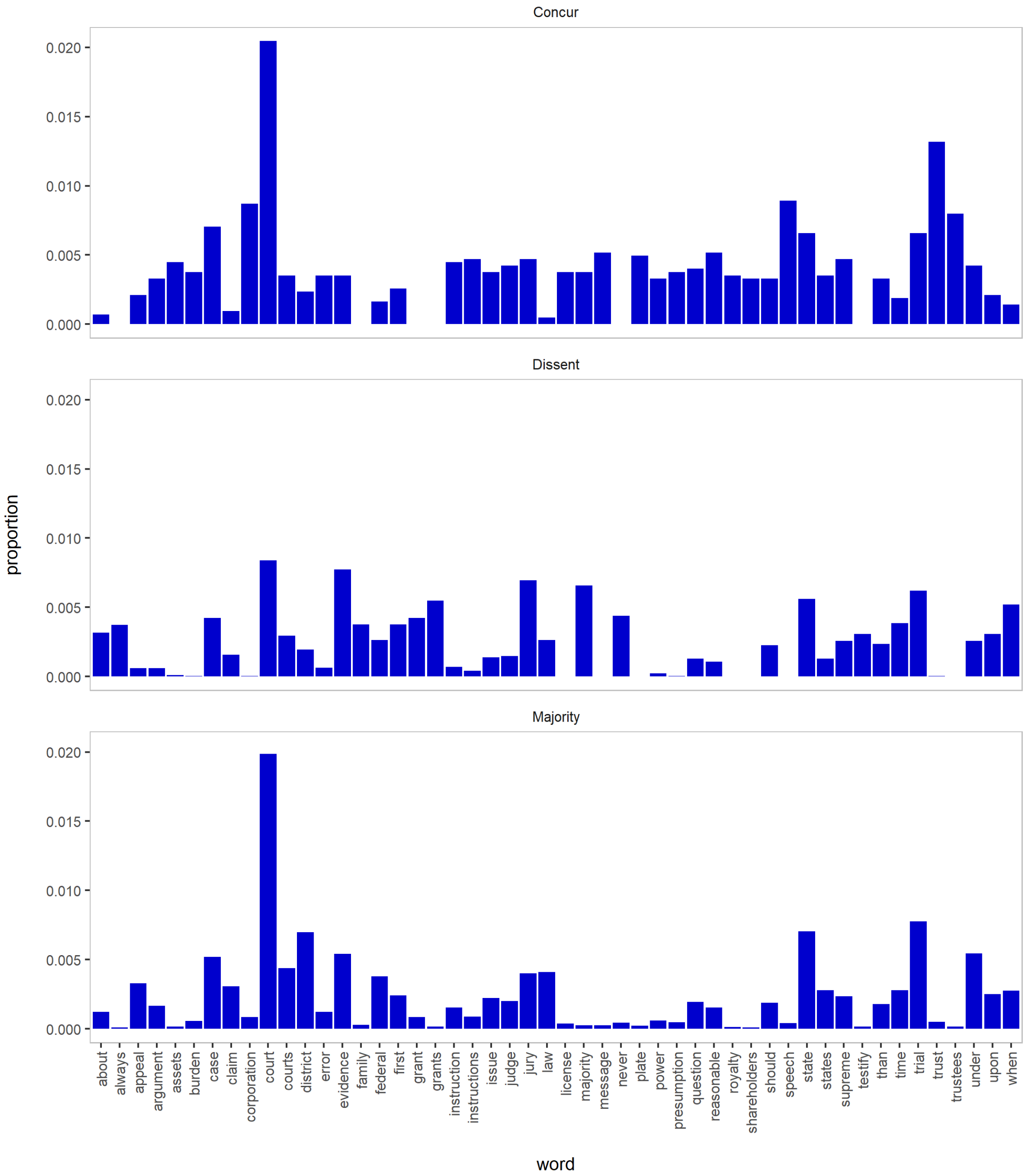


Figure 4: Distribution of fifty most frequent appearing words by opinion type

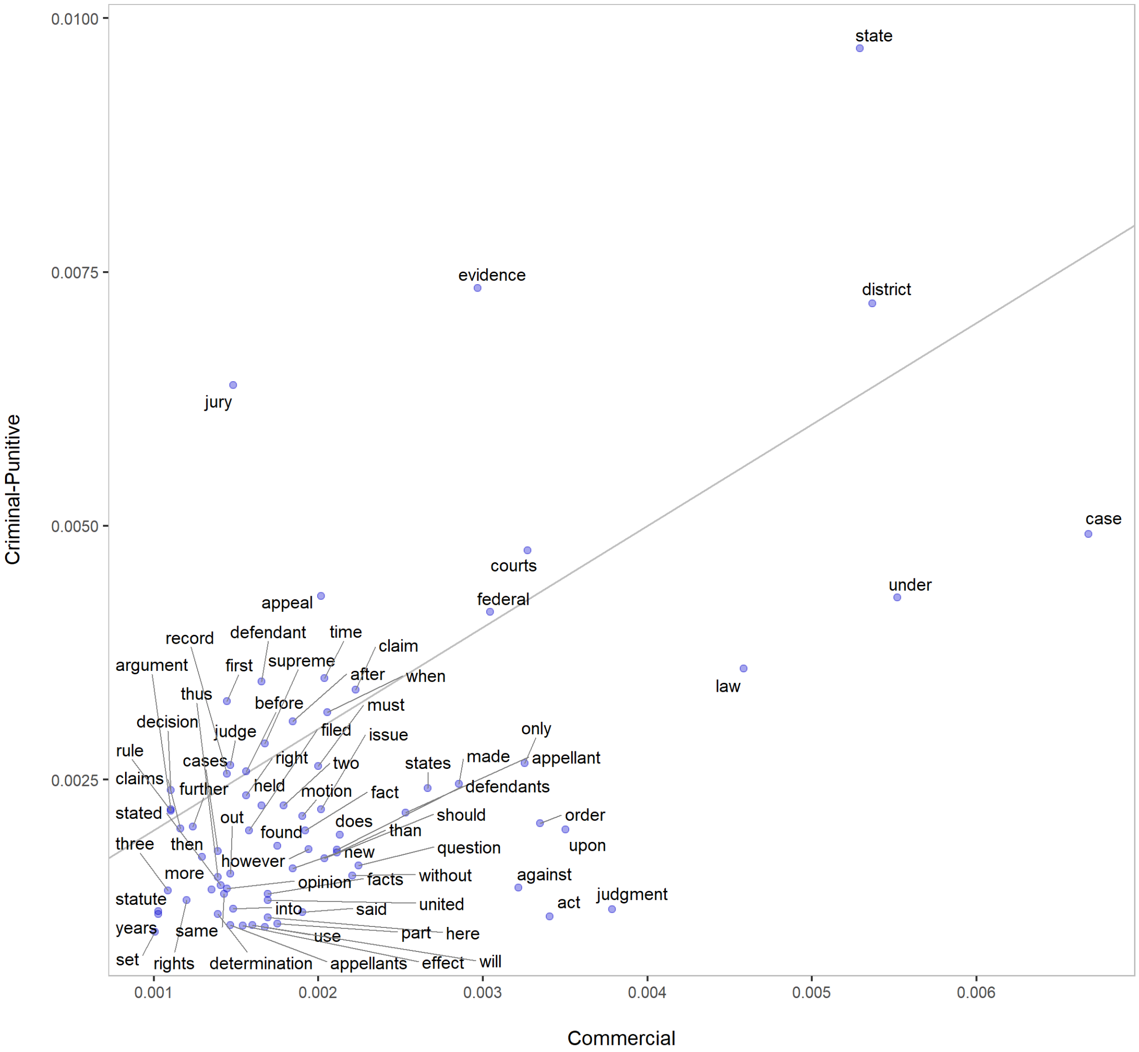


Figure 5: Word occurrence proportions, criminal-punitive vs. commercial cases

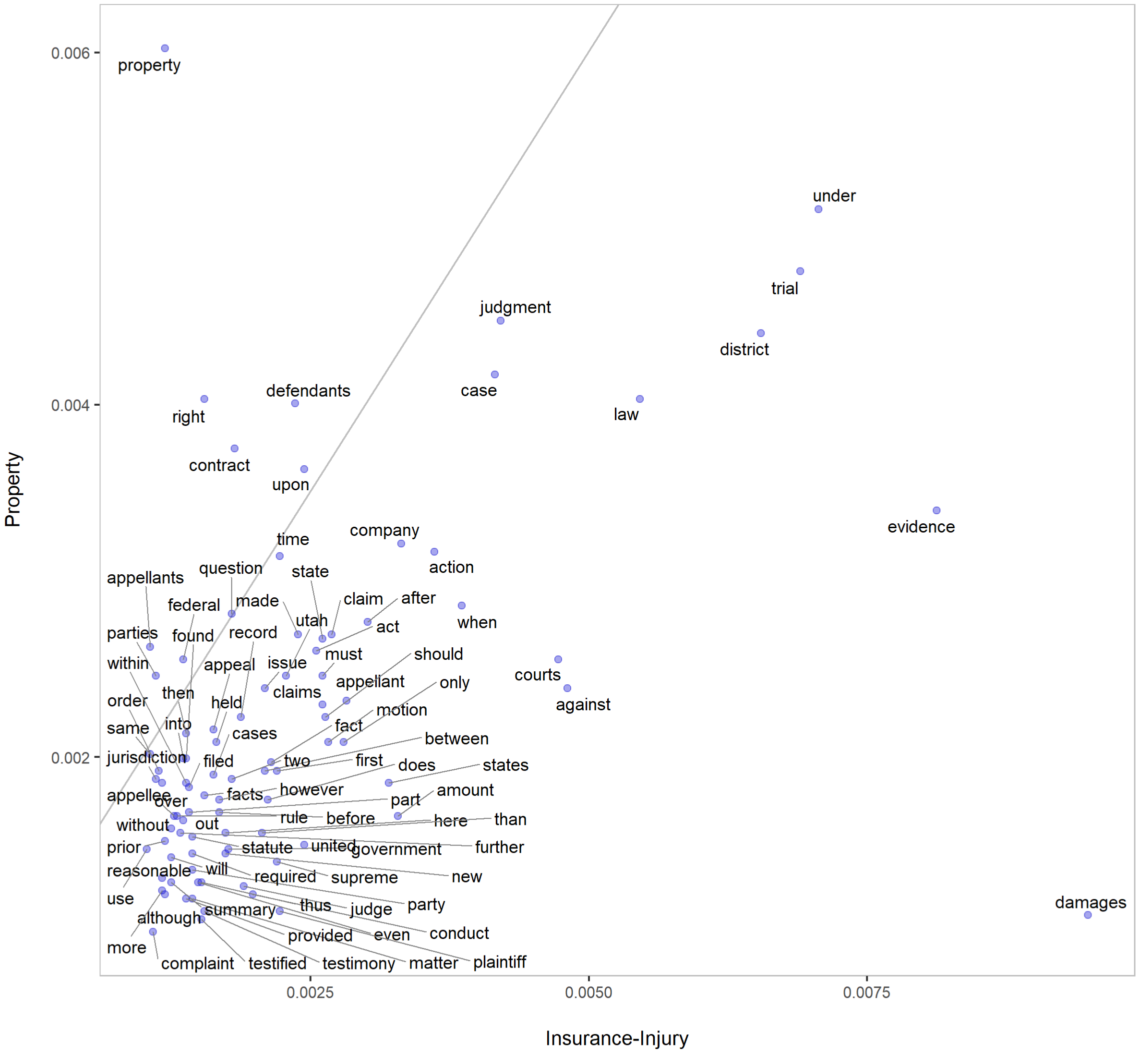


Figure 6: Word occurrence proportions, property vs. insurance-injury cases

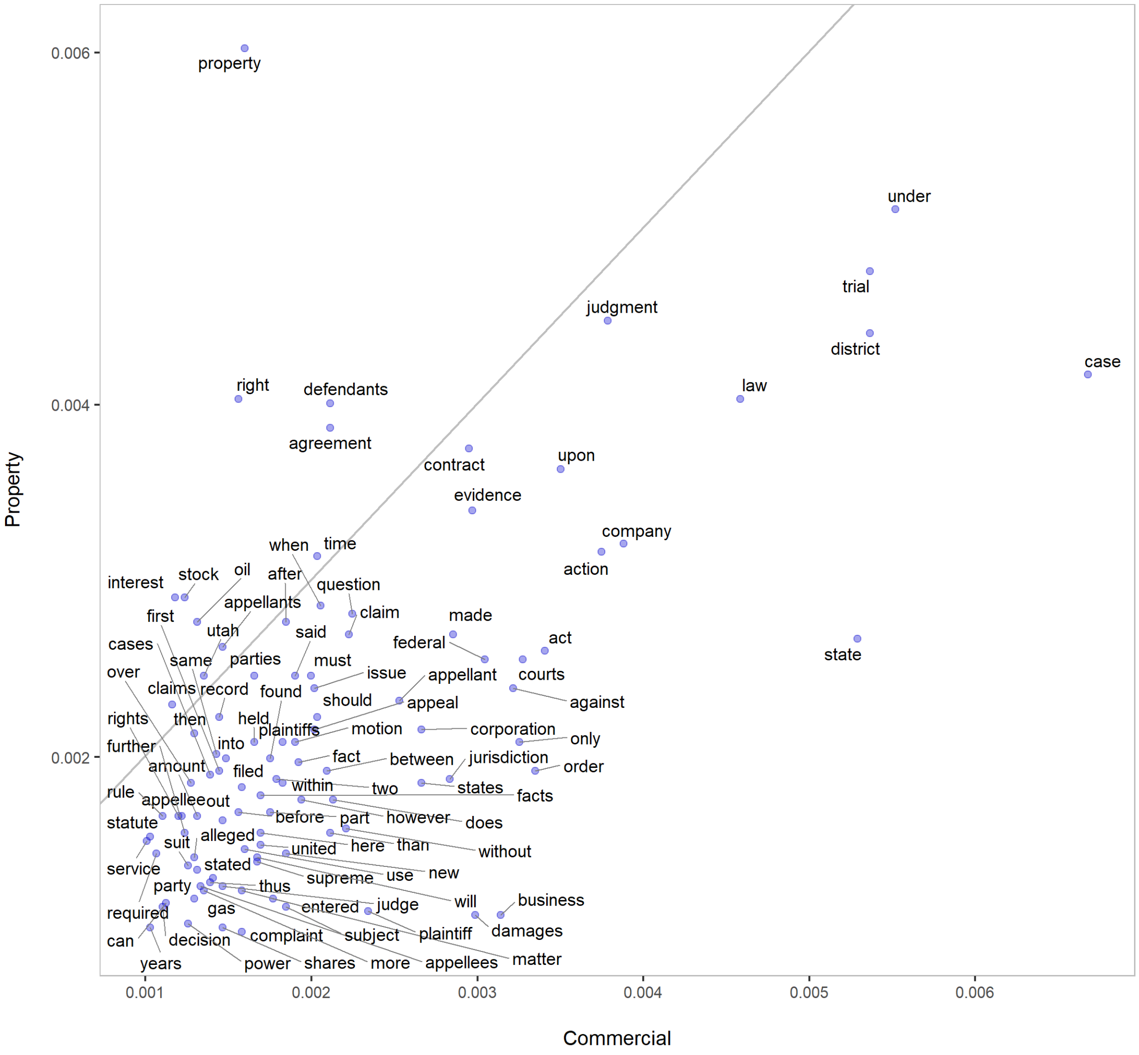


Figure 7: Word occurrence proportions, property vs. commercial cases

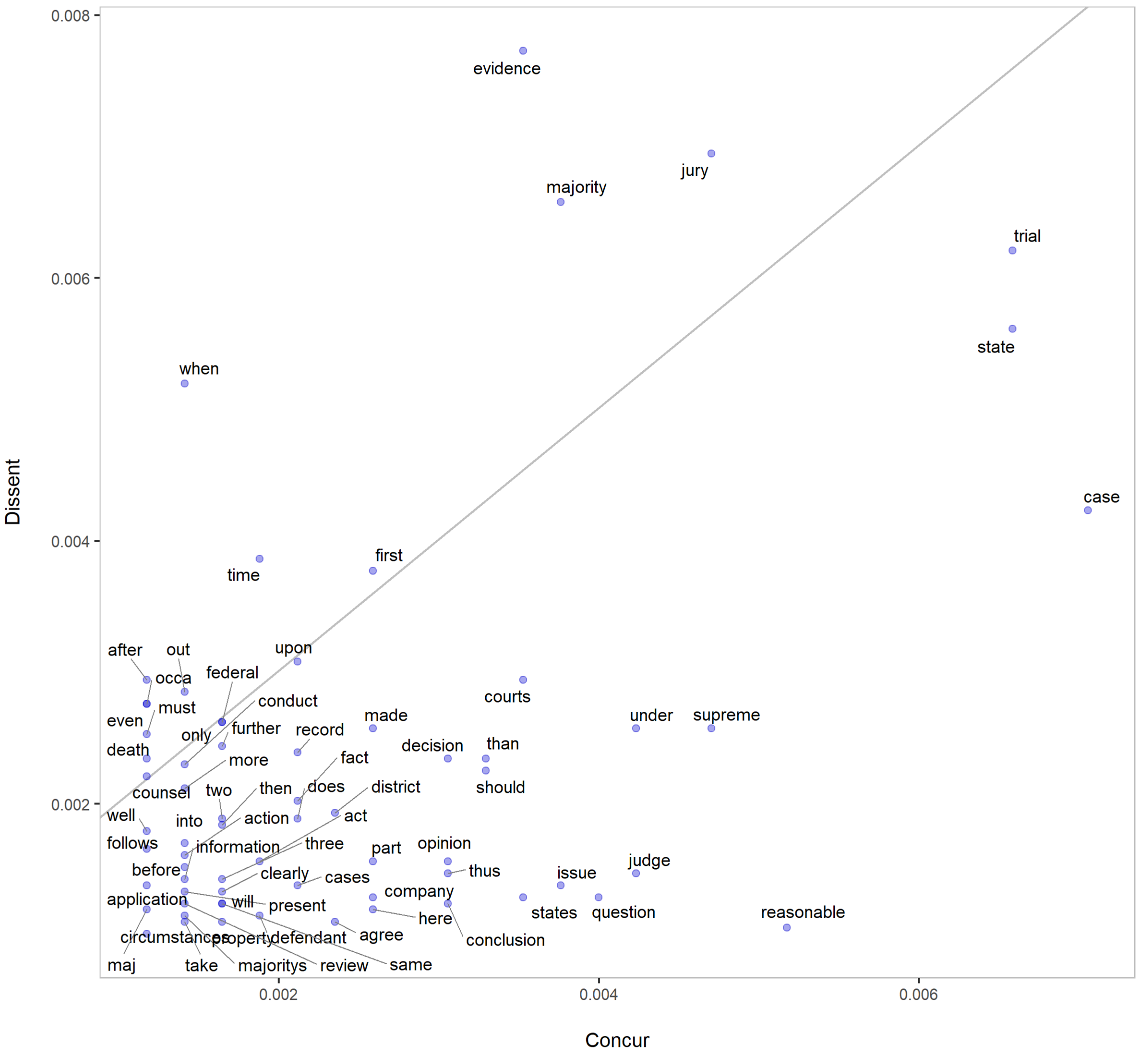


Figure 8: Word occurrence proportions, dissenting vs. concurring opinions

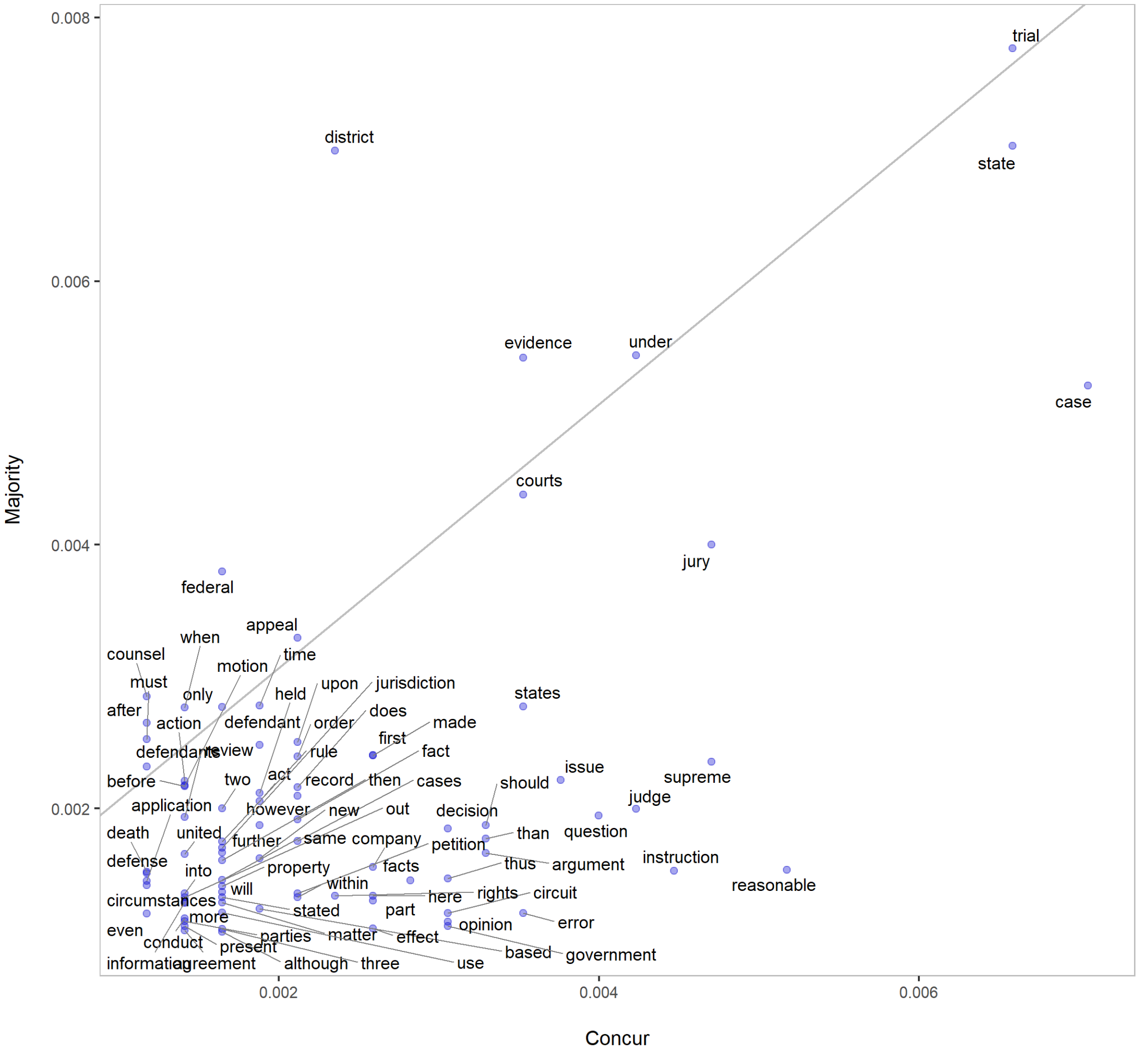


Figure 9: Word occurrence proportions, majority vs. concurring opinions

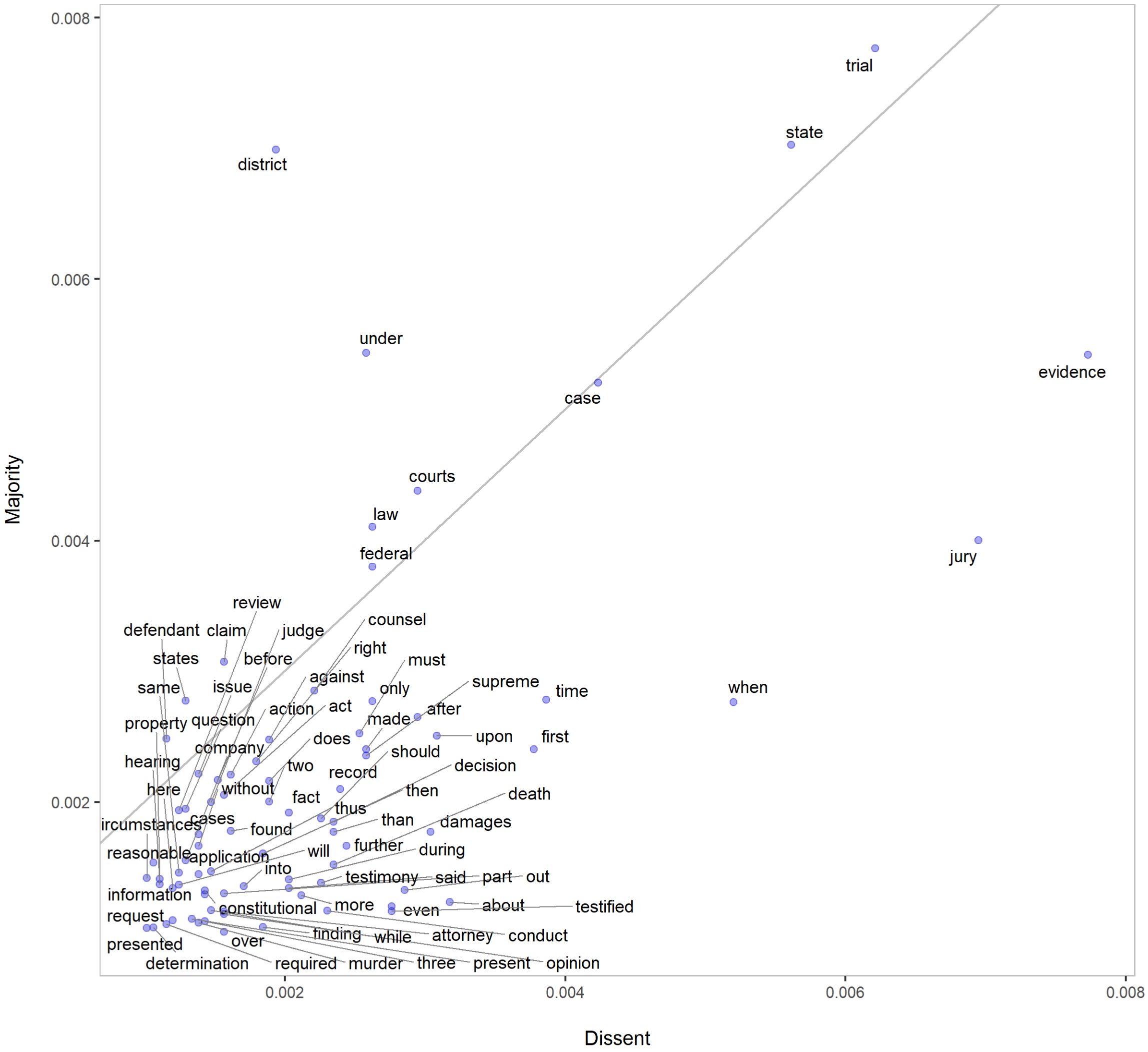


Figure 10: Word occurrence proportions, majority vs. dissenting opinions

2.3 Word Associations

In addition to individual word frequency and proportion, we want to identify words that tend to appear together. An example of this is leading and trailing word pair correlation, where appearance of a leading word is associated with greater than random probability of appearance a set of trailing words. Figures 11 through 22 use network graphs to represent associations between leading (out edge) and trailing (in edge) words. Each graph represents word pairs from opinions in a single case category, opinion type combination.[3] Graphs are paired with the first in each pair (solid, colored nodes with gray edges) using color to represent degree of in and out frequency of words and the second (circular nodes with colored edges) using node radius to represent in and out frequency and edge color to represent word pair correlation.

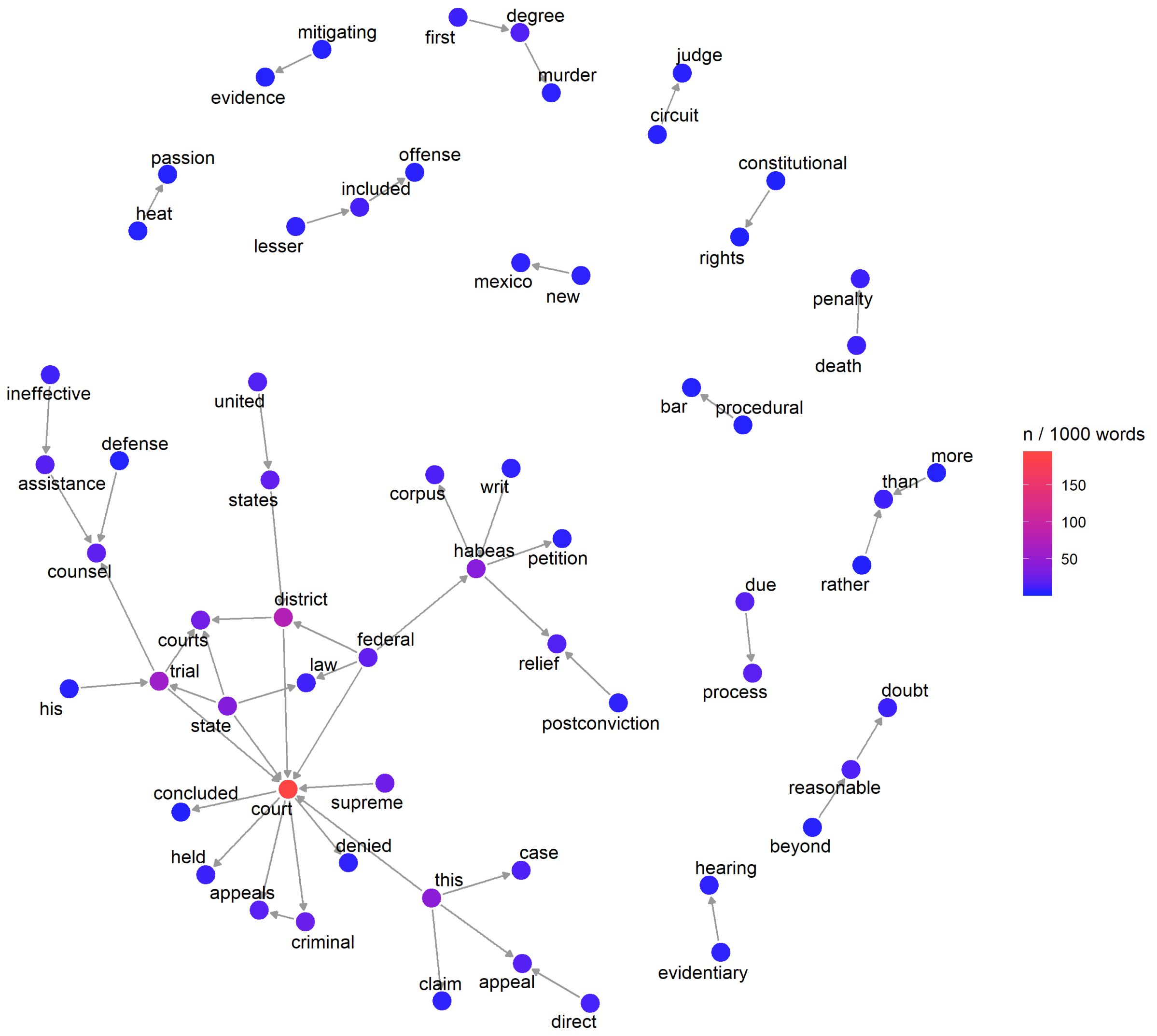


Figure 11: Criminal majority opinion leading and trailing word pair frequency network. Word pairs that occur at a rate of at least 0.4 per 1,000 pairs.

[3] Case categories and opinion types are identified in captions

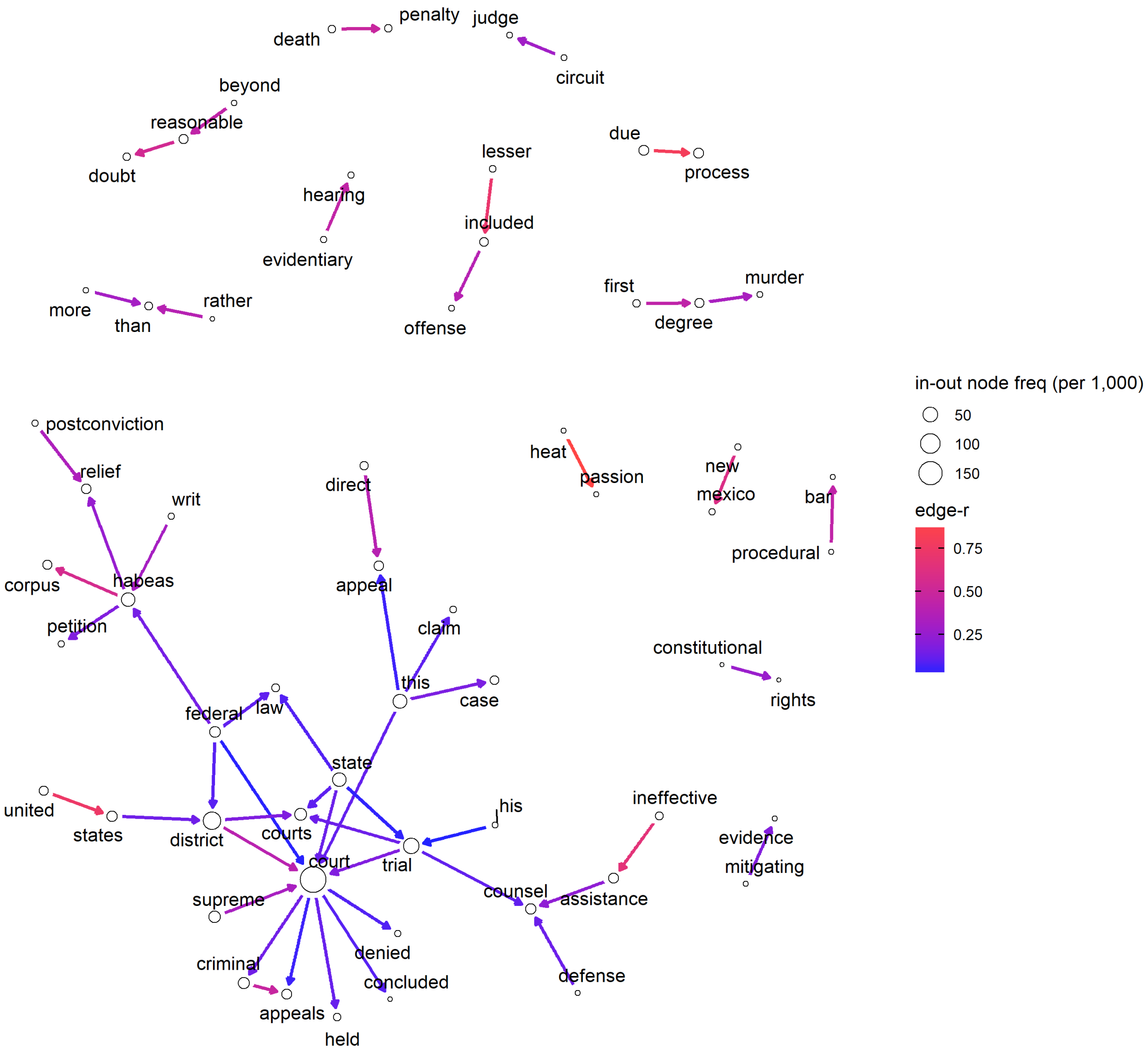


Figure 12: Criminal majority opinion leading and trailing word pair correlation network. Word pairs that occur at a rate of at least 0.4 per 1,000 pairs.

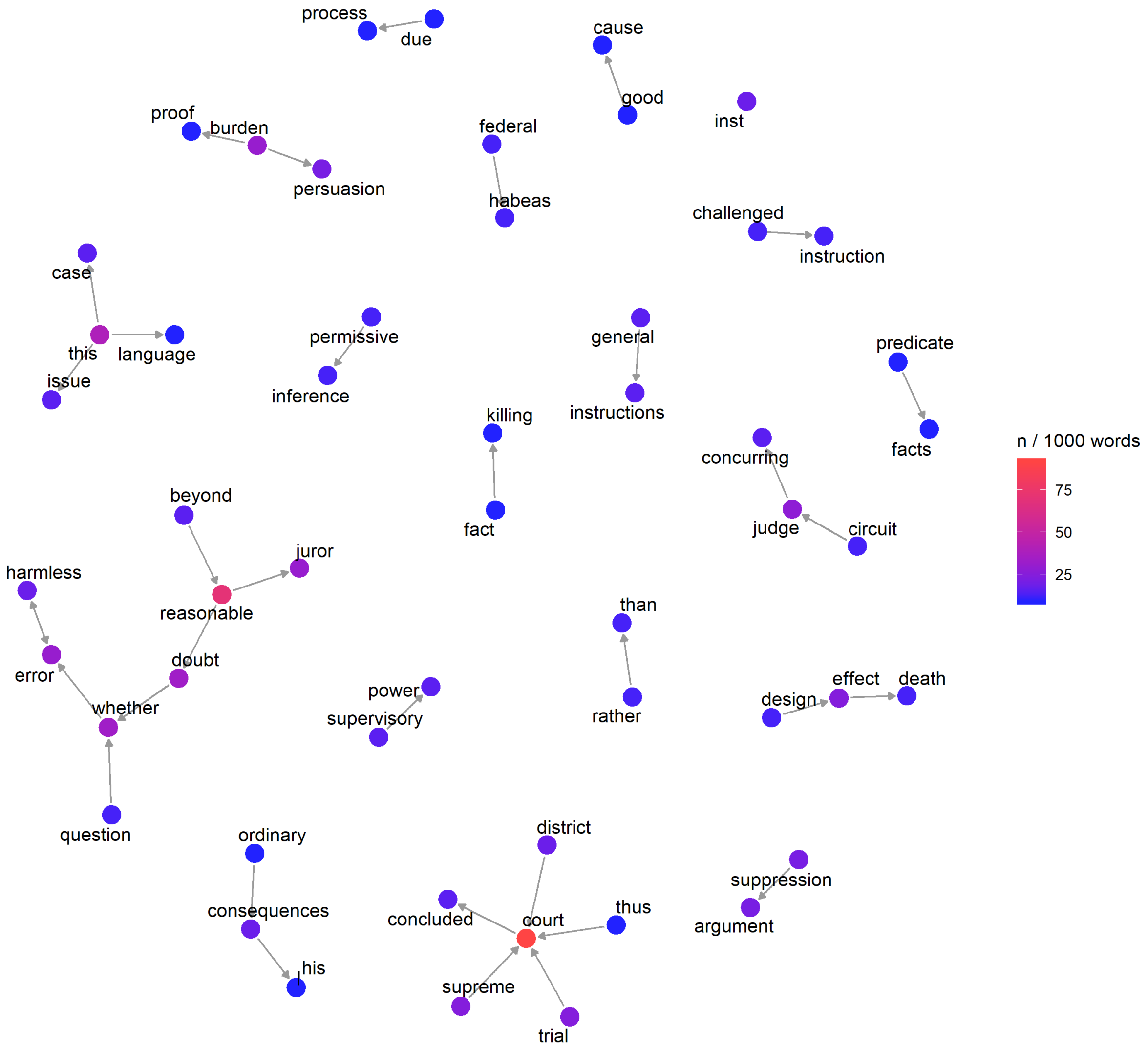


Figure 13: Criminal concurring opinion leading and trailing word pair frequency network. Word pairs that occur at a rate of at least 1.2 per 1,000 pairs.

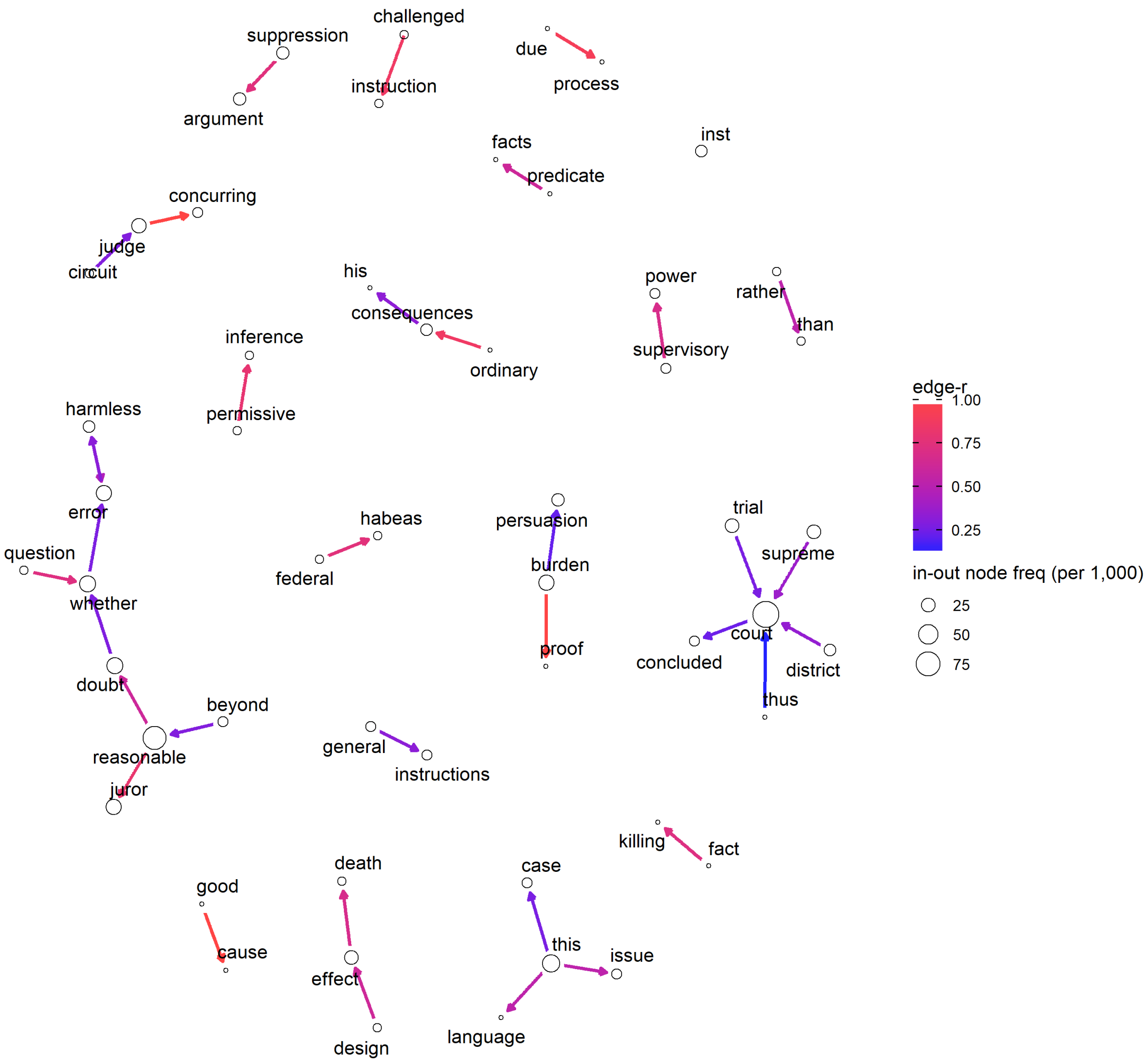


Figure 14: Criminal concurring opinion leading and trailing word pair correlation network. Word pairs that occur at a rate of at least 1.2 per 1,000 pairs.

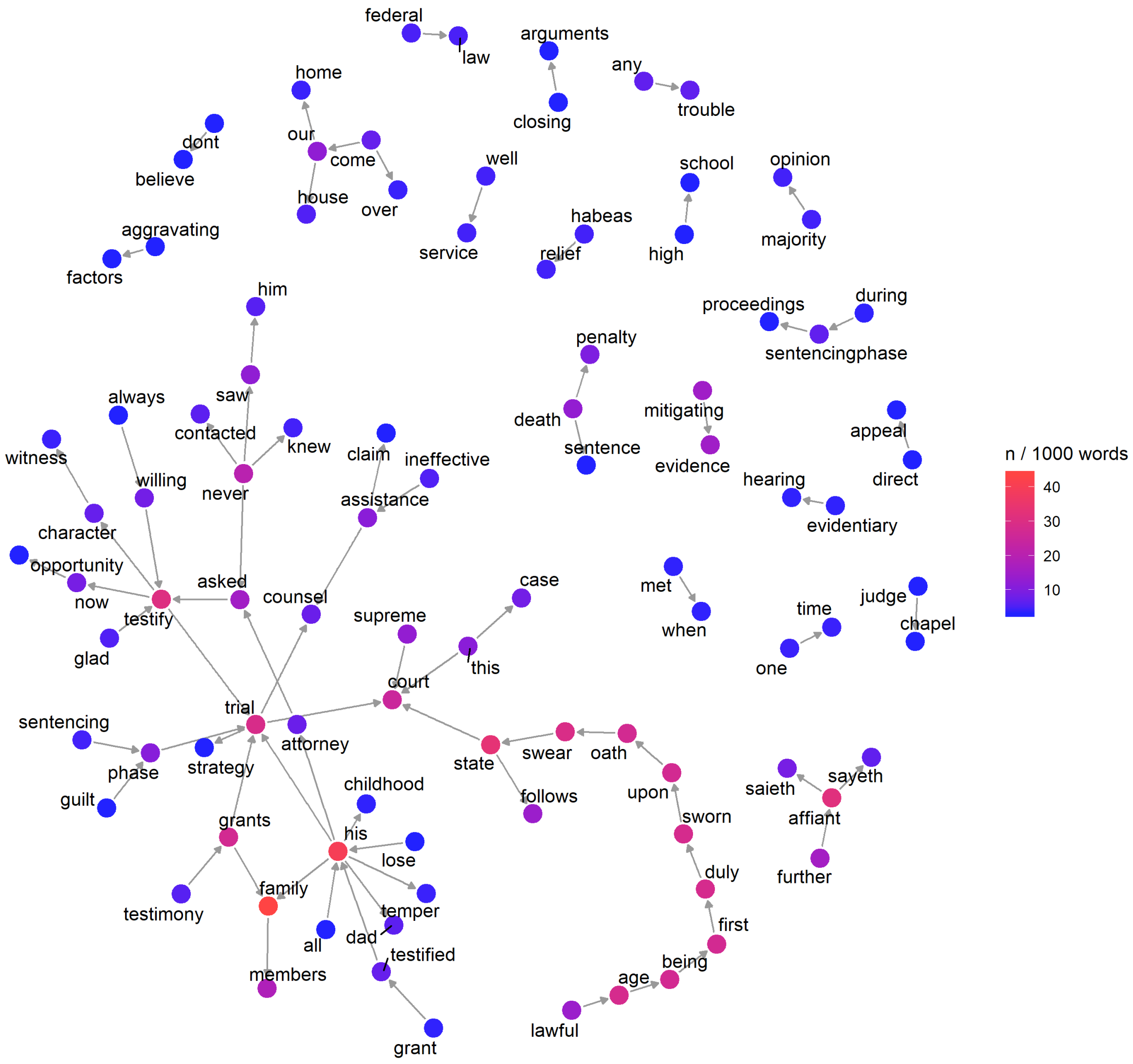


Figure 15: Criminal dissenting opinion leading and trailing word pair frequency network. Word pairs that occur at a rate of at least 0.5 per 1,000 pairs.

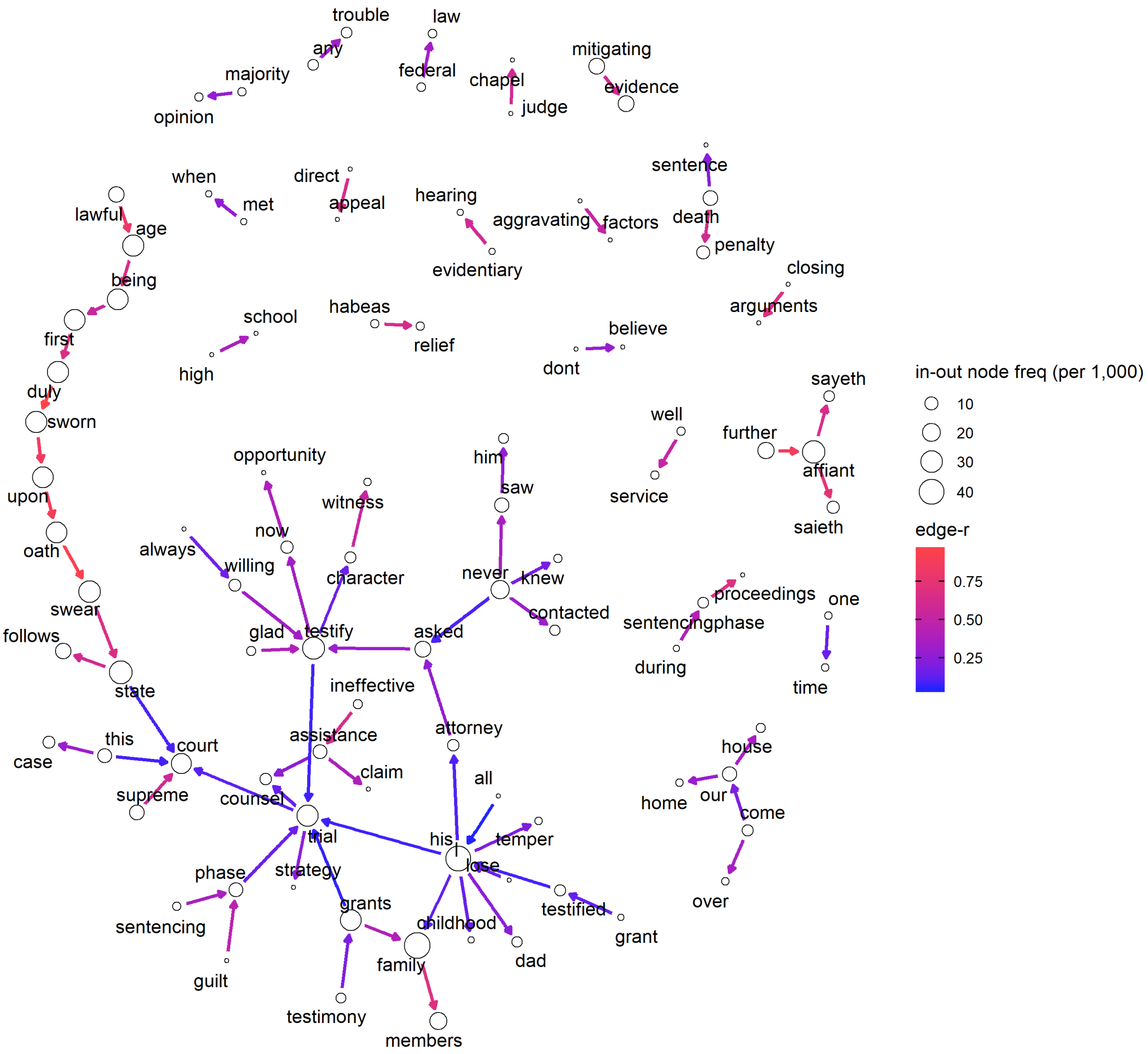


Figure 16: Criminal dissenting opinion leading and trailing word pair correlation network. Word pairs that occur at a rate of at least 0.5 per 1,000 pairs.

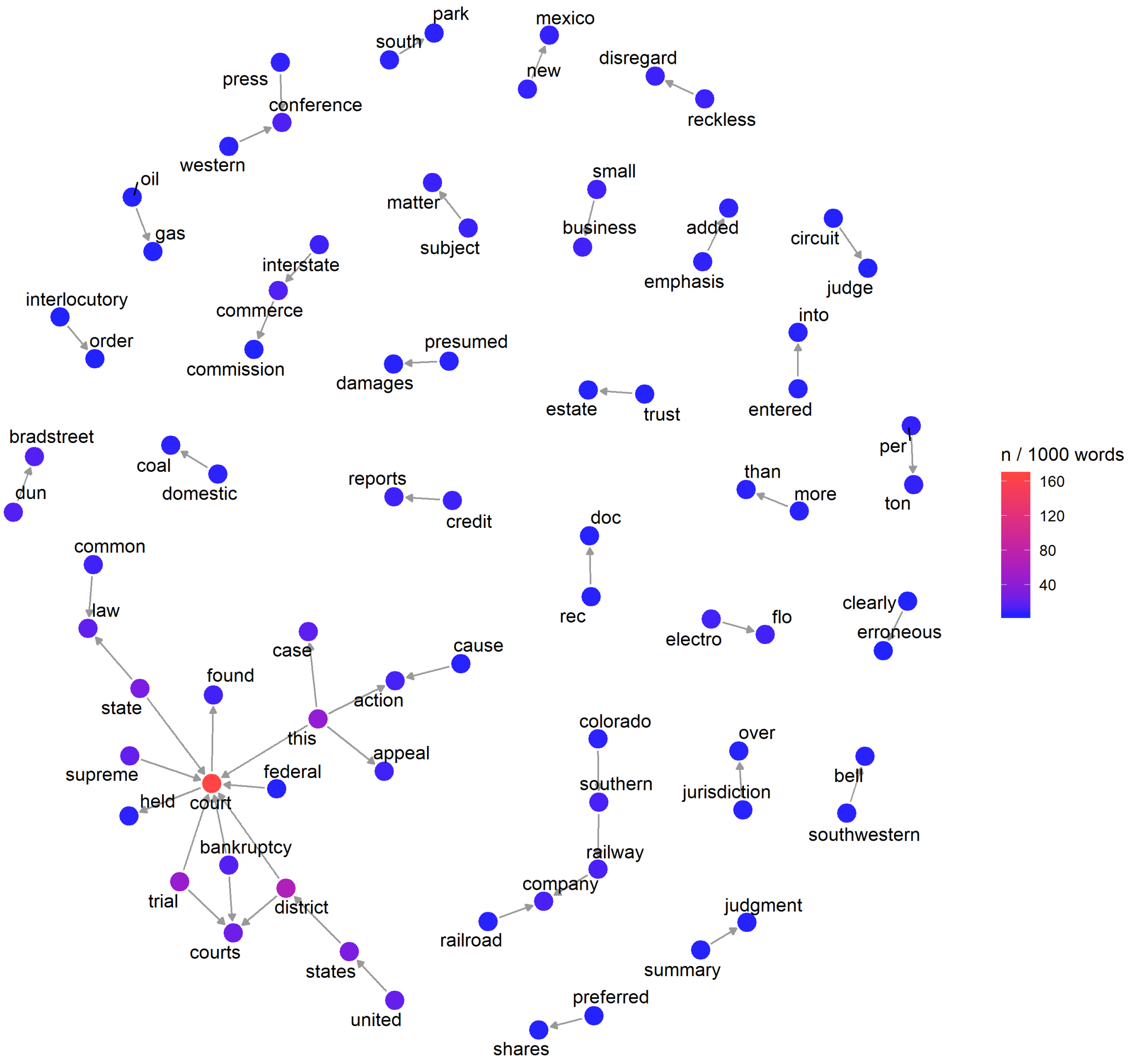


Figure 17: Commercial majority opinion leading and trailing word pair frequency network. Word pairs that occur at a rate of at least 0.4 per 1,000 pairs.

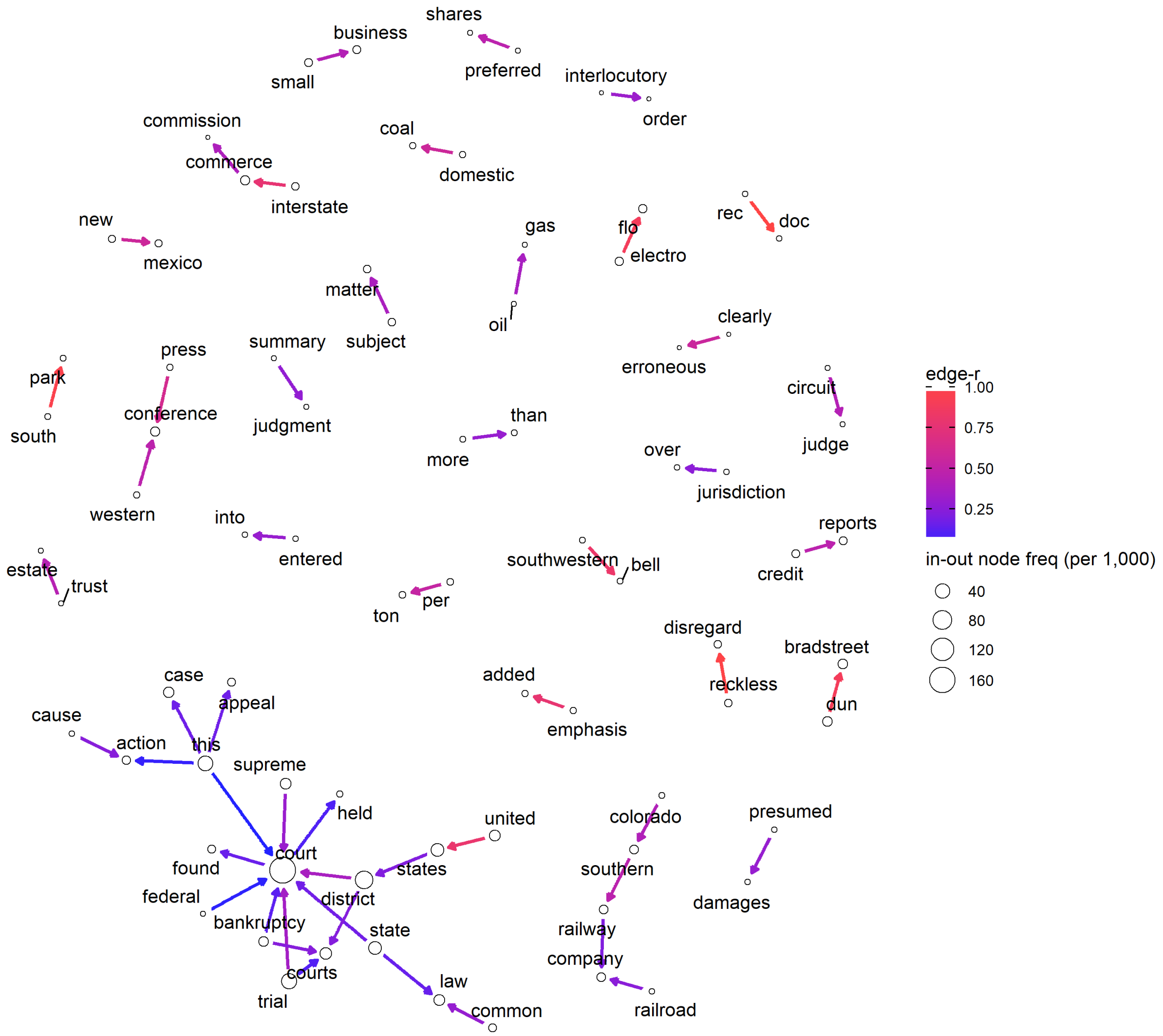


Figure 18: Commercial majority opinion leading and trailing word pair correlation network. Word pairs that occur at a rate of at least 0.4 per 1,000 pairs.

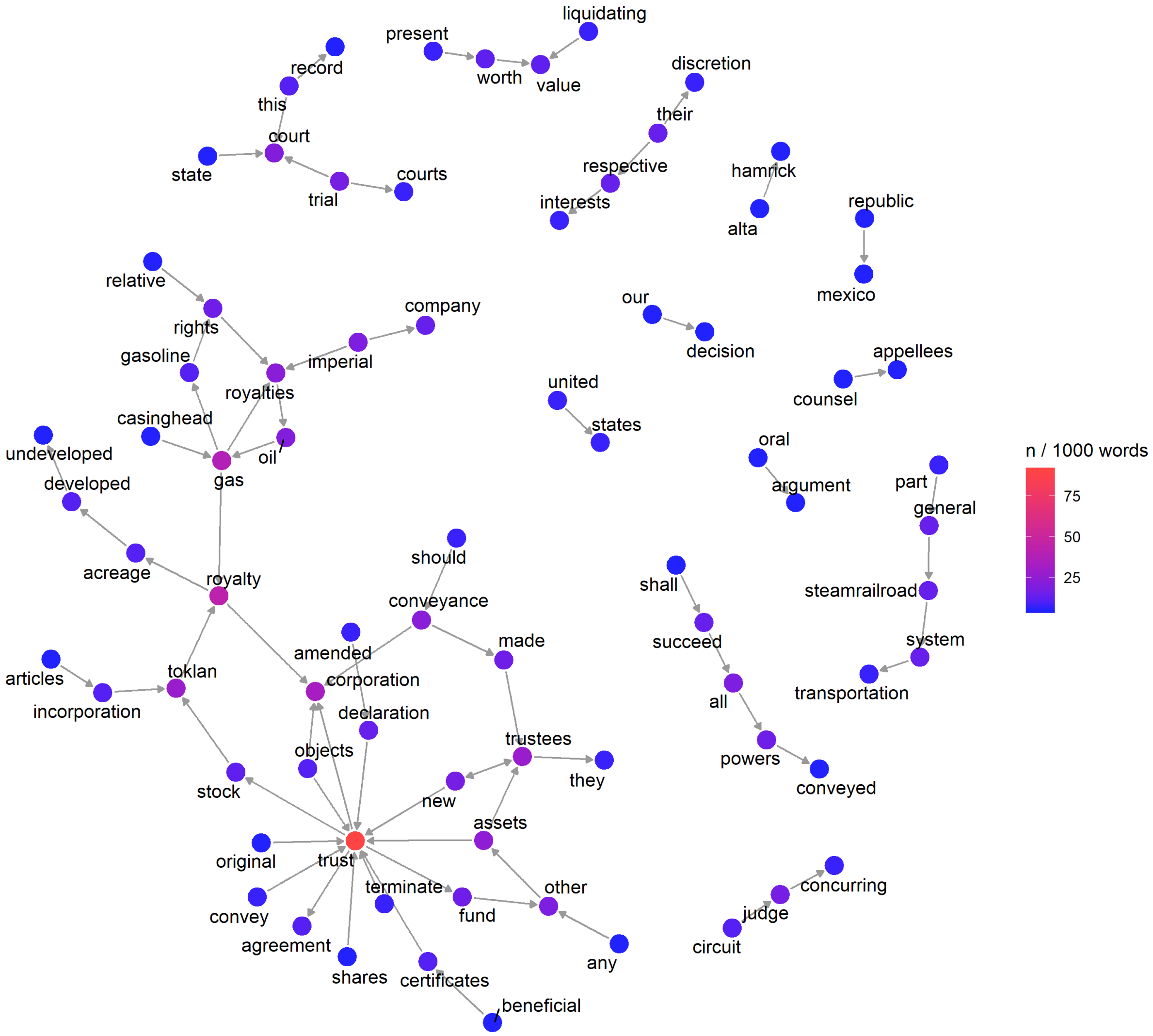


Figure 19: Commercial concurring opinion leading and trailing word pair frequency network. Word pairs that occur at a rate of at least 1.2 per 1,000 pairs.

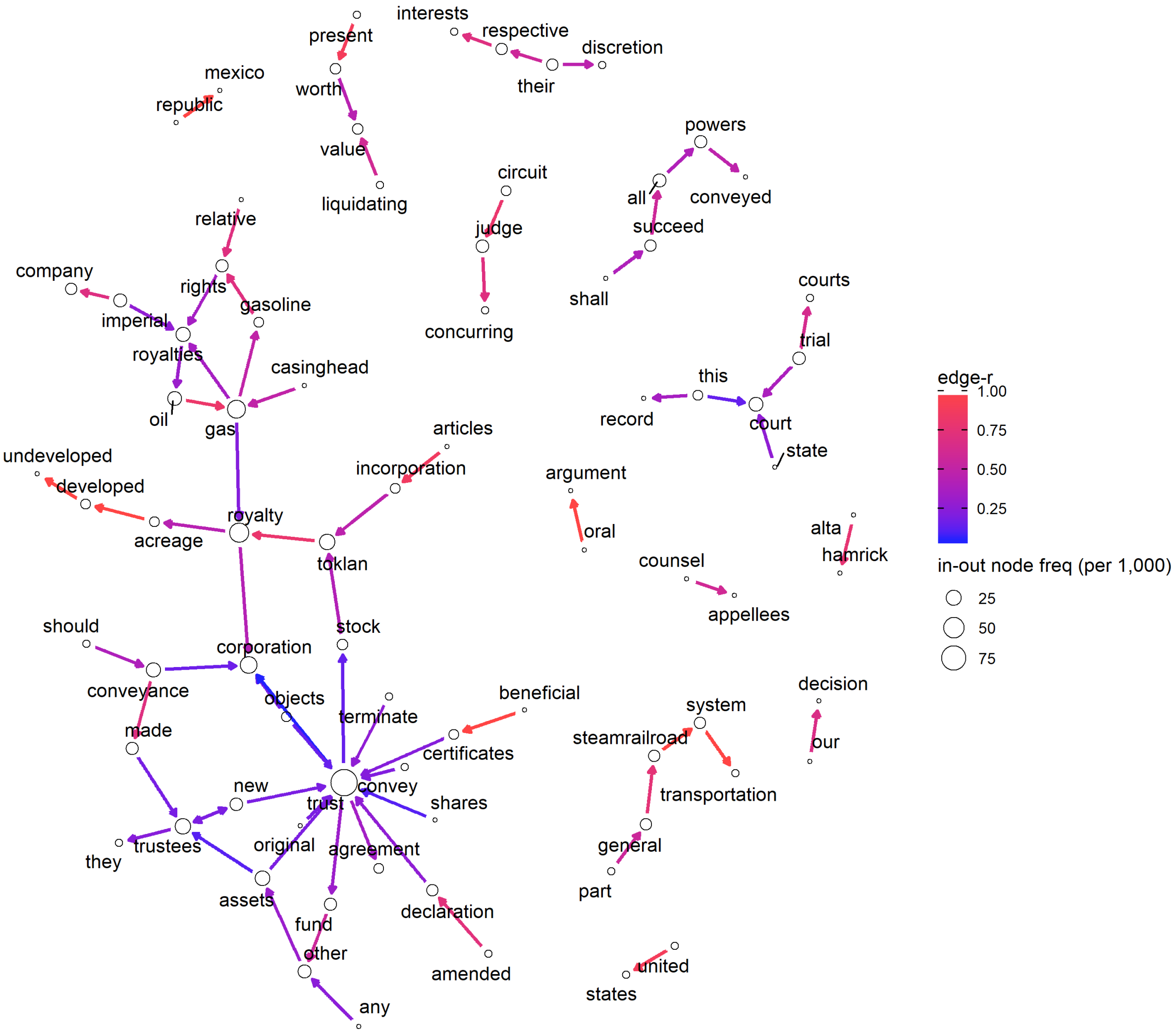


Figure 20: Commercial concurring opinion leading and trailing word pair correlation network. Word pairs that occur at a rate of at least 1.2 per 1,000 pairs.

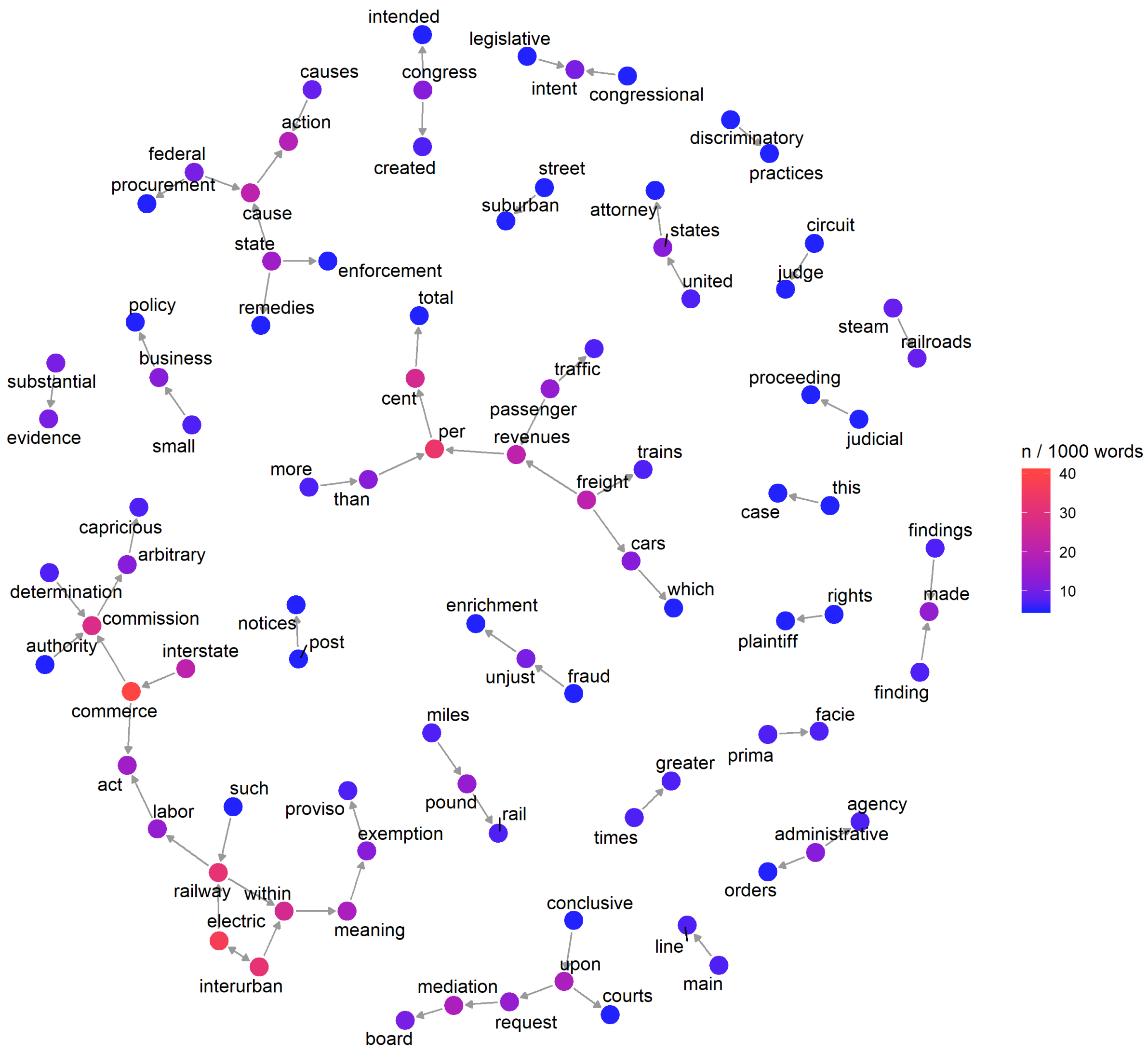


Figure 21: Commercial dissenting opinion leading and trailing word pair frequency network. Word pairs that occur at a rate of at least 0.8 per 1,000 pairs.

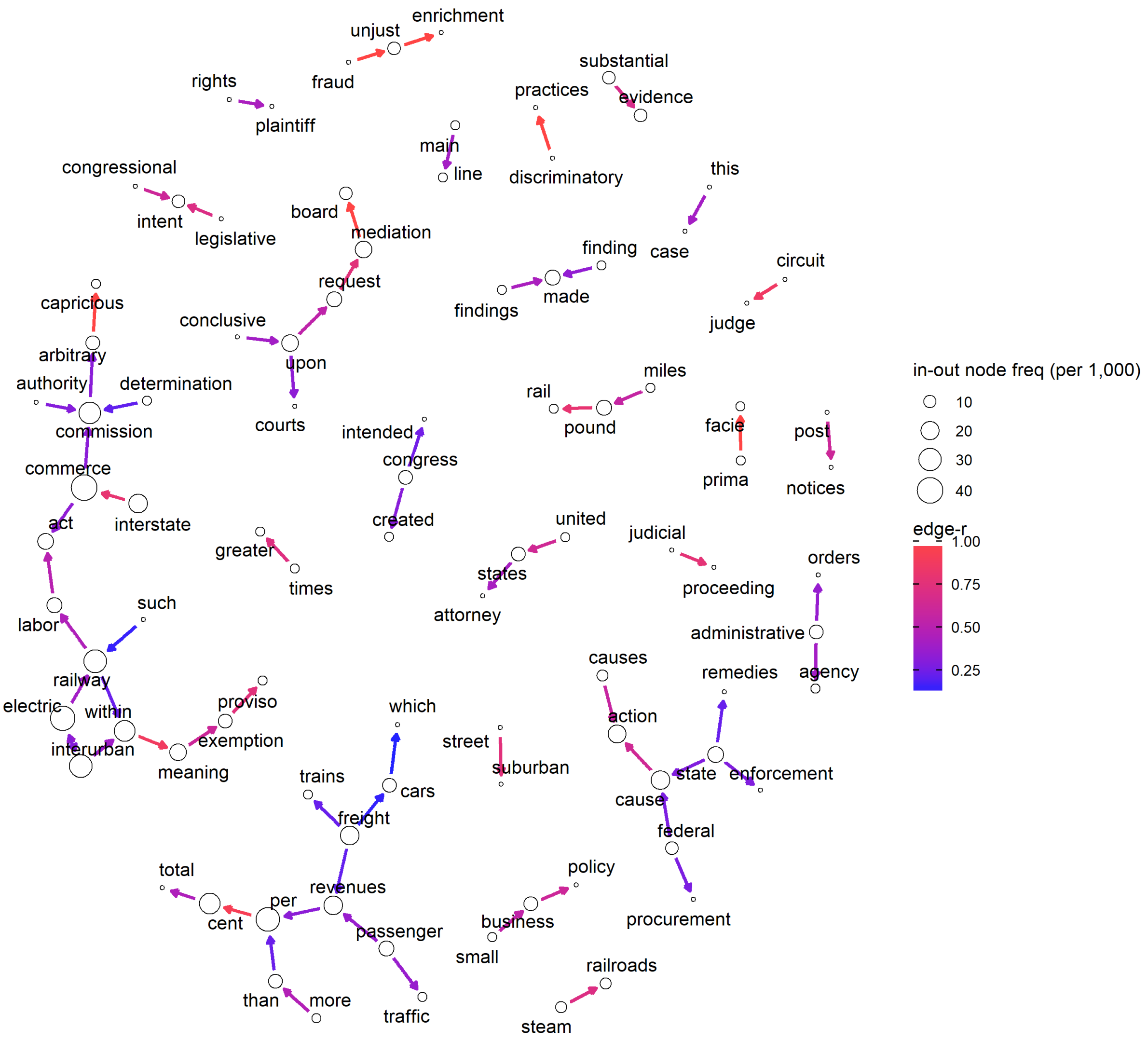


Figure 22: Commercial dissenting opinion leading and trailing word pair correlation network. Word pairs that occur at a rate of at least 0.8 per 1,000 pairs.

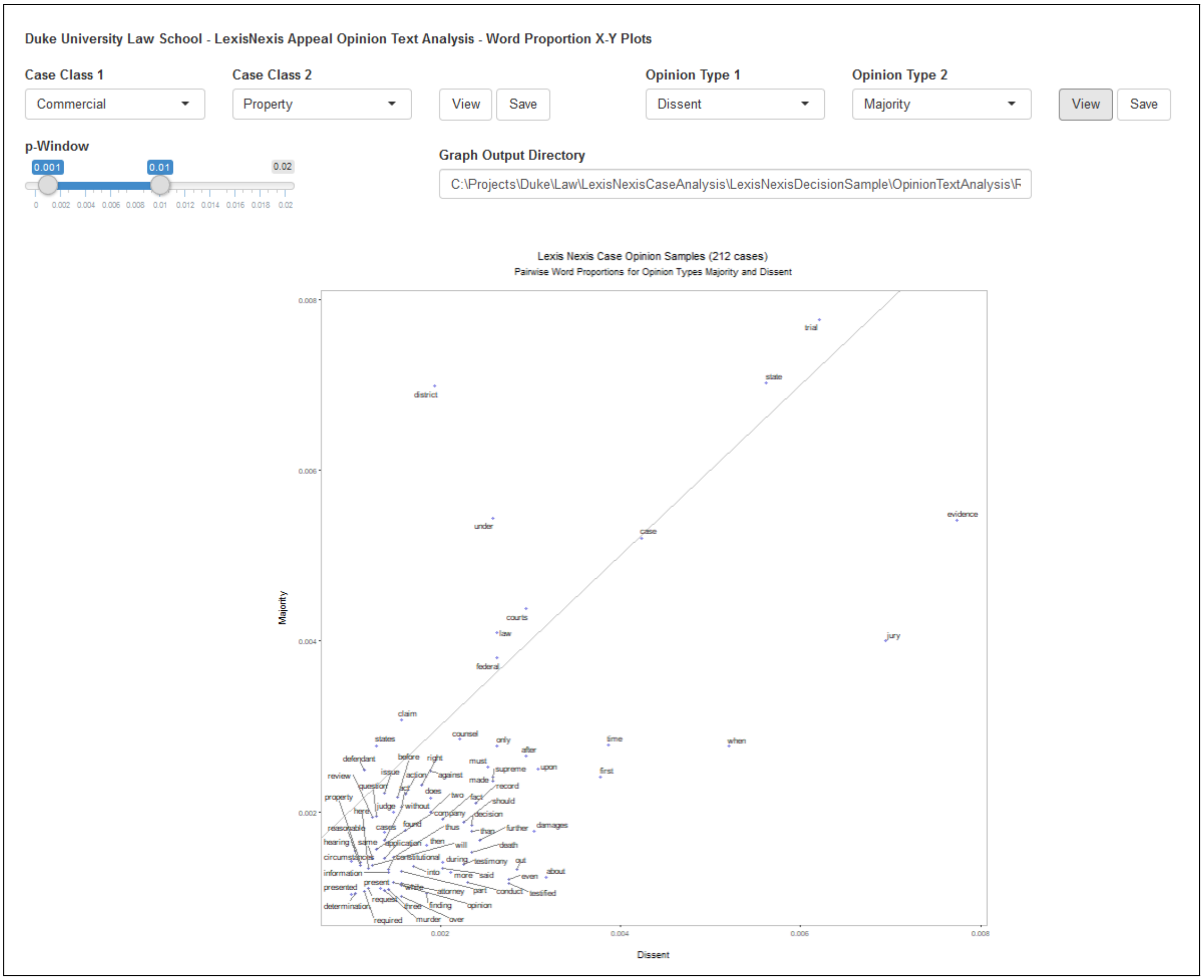
3 R-Shiny Applications

Two R-Shiny applications have been developed to aid in visualizing word relationships within opinion text. The first application generates plots of word proportions within text of one case category against another. Figures 5 through 10 were produced with this app. Figure 23 is a sample screen shot of the user interface. Prompts and controls include:

- Selection fields for two case categories (classes), one for the x-axis (Case Class 1) and one for the y-axis (Case Class 2)
- Selection fields for two case opinion types, one for the x-axis (Opinion Type 1) and one for the y-axis (Opinion Type 2)
- Proportion range filter (p-Range) to limit selected words (nodes) to those with appearance proportion in the specified range, which is useful in identifying contrast or agreement of high proportion words
- View and save buttons for both category and opinion type plots
- Directory location in which to save png versions of plots for use in presentations and Latex documents

The second application produces network graphs of leading and trailing word pair proportions and correlation. Figures 11 through 22 were produced with this app. Figure 24 is a sample screen-shot of the user interface. Prompts and controls include:

- Selection fields for case category (class) and opinion type
- Edge frequency threshold, which limits the graph to edges with a specified minimum frequency (per 1,000 edges)
- View and save buttons
- Directory location in which to save png versions of the graph



Duke University Law School - LexisNexis Appeal Opinion Text Analysis - Word Proportion X-Y Plots

Case Class 1: Commercial

Case Class 2: Property

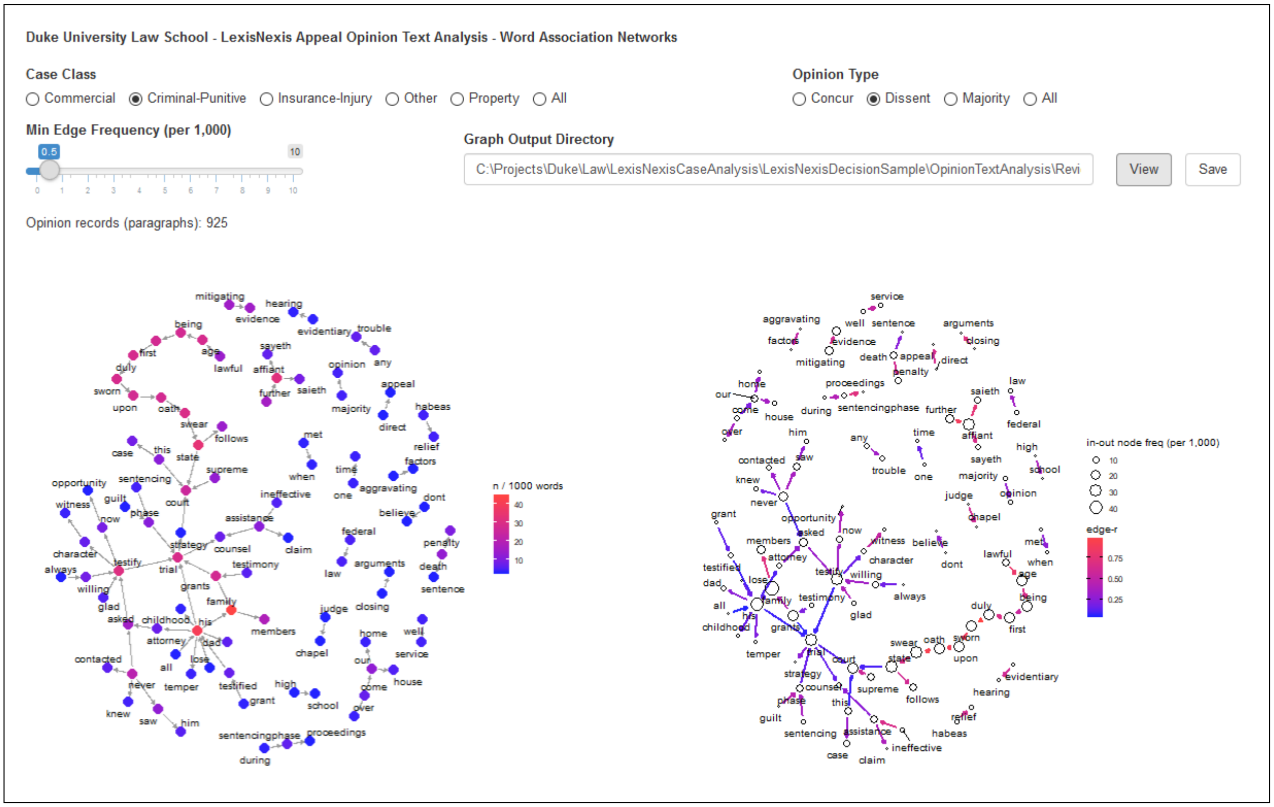
Opinion Type 1: Dissent

Opinion Type 2: Majority

p-Window: 0.001 – 0.01

Graph Output Directory: C:\Projects\Duke\Law\LexisNexisCaseAnalysis\LexisNexisDecisionSample\OpinionTextAnalysis\F

Figure 23: Screenshot of two-way opinion word proportion appearance R-Shiny application



Duke University Law School - LexisNexis Appeal Opinion Text Analysis - Word Association Networks

Case Class

○ Commercial ● Criminal-Punitive ○ Insurance-Injury ○ Other ○ Property ○ All

Opinion Type

○ Concur ● Dissent ○ Majority ○ All

Min Edge Frequency (per 1,000)

0.5 10

Graph Output Directory

C:\Projects\Duke\Law\LexisNexisCaseAnalysis\LexisNexisDecisionSample\OpinionTextAnalysis\Revi

Opinion records (paragraphs): 925

Figure 24: Screenshot of opinion text word pair frequency and correlation analysis R-Shiny application

4 Further Analysis

Additional activities to identify relationships of words contained in text include:

- Identify high frequency associations of a given word to words within k neighboring words (a k-neighborhood)
- Measure pairwise correlation of words to members of their k-neighborhood, with weight inversely proportional to distance
- Measure frequency of entire key phrases
- Measure correlation of key phrase pairs
- Identify patterns of key phrase appearance (does appearance in a given paragraph predict appearance in preceding or subsequent paragraphs?)
- Develop additional R-Shiny apps to facilitate exploration and identification of opinion text relationships