Supplementary Material for “Friendships, Rivalries, and Trysts: Characterizing Relations between Ideas in Texts”

Chenhao Tan[*] **Dallas Card**[†] **Noah A. Smith**[*]
[*]Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA 98195, USA
[†]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
chenhao@chenhaot.com dcard@cmu.edu nasmith@cs.washington.edu

1 Procedure to Collect News Articles

The news articles for this paper were obtained from LexisNexis, using the Academic Search tool. For each issue, we downloaded all articles published between January 1, 1980 and December 31, 2016 matching a set of search criteria (Newspaper, Subject, and Geography). The newspapers searched were the same as those given in the Supplementary material of Card et al. (2015). We adapted the code from https://github.com/dallascard/media_frames_corpus. The search requirements were that articles possessed the tag “Geographic:United States”, and at least one relevant subject terms with ≥ 90% confidence. The search terms for each issue are given in Table 1.

Issue	Search terms
Abortion (27,549)	ABORTION; ABORTION LAWS
Immigration (51,898)	IMMIGRATION, CITIZENSHIP DISPLACEMENT; IMMIGRATION LAW; FOREIGN LABOR; IMMIGRATION; ILLEGAL IMMIGRANTS; IMMIGRANT DETENTION CENTERS; ALIEN SMUGGLING; INADMISSIBILITY OF IMMIGRANTS; US STATE IMMIGRATION LAW
Same-sex marriage (64,881)	DOMESTIC PARTNERSHIPS; SAME SEX MARRIAGE & UNIONS; SAME SEX MARRIAGE LAWS; GAYS & LESBIANS
Smoking (94,358)	TOBACCO; SMOKING; SMOKING CESSATION; TOBACCO HEALTH; TOBACCO PRODUCTS; TOBACCO FARMING; TOBACCO MFG; SMOKING BANS
Terrorism (29916)	TERRORISM

Table 1: LexisNexis subject terms for finding articles, with the number of articles shown in brackets.

2 Overall Distributions using Keywords to Represent Ideas

Table 2 shows the top 10 distinguishing words in each corpus. Fig. 1 shows the same plot as Fig. 3 in the main paper except that keywords are used to represent ideas. Distributions of cooccurrence are still unimodal, while distributions of prevalence correlation look different for news articles and research papers.

Terrorism	attack, terrorist, official, security, terrorism, war, military, iraq, bush, killed
Immigration	immigration, immigrant, illegal, worker, border, mexico, reform, visa, mexican, country
Abortion	abortion, woman, clinic, doctor, right, procedure, supreme court, pregnancy, life, bill
Same-sex marriage	marriage, gay, couple, church, married, same-sex, love, lesbian, right, partner
Smoking	company, industry, tobacco, smoking, health, cigarette, smoke, ban, product, tax
ACL	word, sentence, language, proceeding, corpus, translation, text, relation, table, feature
NIPS	function, network, algorithm, learning, image, matrix, distribution, problem, sample, point

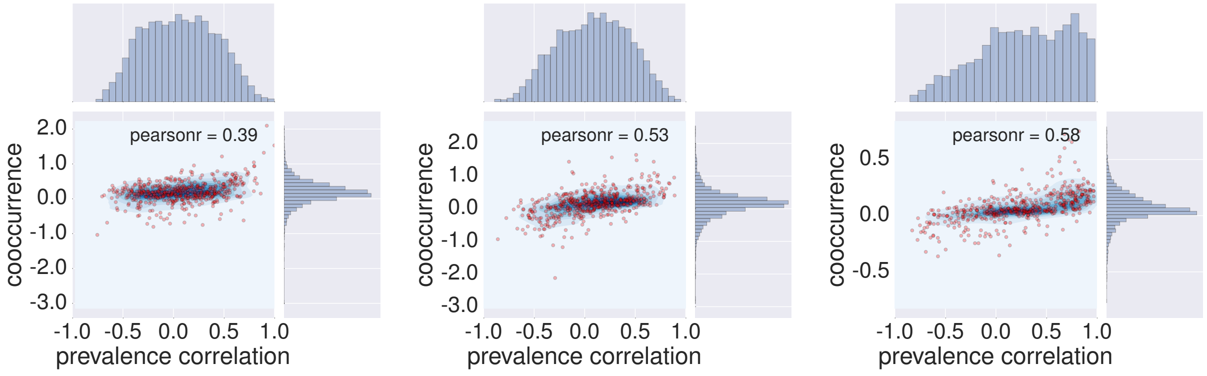
Table 2: Top 10 keywords using an informative Dirichelet prior model (Monroe et al., 2008).

References

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The Media Frames Corpus: Annotations of frames across issues. In *Proceedings of ACL*.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4):372–403.

David W. Scott. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.



(a) Terrorism keywords

(b) Immigration keywords

(c) ACL keywords

Figure 1: Overall distributions of cooccurrence and prevalence correlation. In the main plot, each point represents a pair of ideas: color density shows the kernel density estimation of the joint distribution (Scott, 2015). The plots along the axes show the marginal distribution of the corresponding dimension. In each plot, we give the Pearson correlation, and all Pearson correlations' p-values are less than 10^{-40}. In all plots, we use keywords to represent ideas.