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
chenhao@chenhaot.com nasmith@cs.washington.edu
†School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
dcard@cmu.edu

Abstract

Understanding how ideas relate to each other is a fundamental question in many domains, ranging from intellectual history to public communication. Because ideas are naturally embedded in texts, we propose the first framework to systematically characterize the relations between ideas based on their occurrence in a corpus of documents, independent of how these ideas are represented. Combining two statistics—*cooccurrence within documents* and *prevalence correlation over time*—our approach reveals a number of different ways in which ideas can cooperate and compete. For instance, two ideas can closely track each other's prevalence over time, and yet rarely cooccur, almost like a "cold war" scenario. We observe that pairwise cooccurrence and prevalence correlation exhibit different distributions. We further demonstrate that our approach is able to uncover intriguing relations between ideas through in-depth case studies on news articles and research papers.

1 Introduction

Ideas exist in the mind, but are made manifest in language, where they compete with each other for the scarce resource of human attention. Milton (1644) used the "marketplace of ideas" metaphor to argue that the truth will win out when ideas freely compete; Dawkins (1976) similarly likened the evolution of ideas to natural selection of genes. We propose a framework to quantitatively characterize competition and cooperation between ideas in texts, independent of how they might be represented.

By "ideas", we mean any discrete conceptual units that can be identified as being present or absent in a document. In this work, we consider representing ideas using keywords and topics obtained in an unsupervised fashion, but our way of characterizing the *relations* between ideas could be applied to many other types of textual representations, such as frames (Card et al., 2015) and hashtags.

What does it mean for two ideas to compete in texts, quantitatively? Consider, for example, the issue of immigration. There are two strongly competing narratives about the roughly 11 million people[1] who are residing in the United States without permission. One is "illegal aliens", who "steal" jobs and deny opportunities to legal immigrants; the other is "undocumented immigrants", who are already part of the fabric of society and deserve a path to citizenship (Merolla et al., 2013).

Although prior knowledge suggests that these two narratives compete, it is not immediately obvious what measures might reveal this competition in a corpus of writing about immigration. One question is whether or not these two ideas cooccur in the same documents. In the example above, these narratives are used by distinct groups of people with different ideologies. The fact that they don't cooccur is one clue that they may be in competition with each other.

However, cooccurrence is insufficient to express the selection process of ideas, i.e., some ideas fade out over time, while others rise in popularity, analogous to the populations of species in nature. Of the two narratives on immigration, we may expect one to win out at the expense of another as public opinion shifts. Alternatively, we might expect to see these narratives reinforcing each other, as both sides intensify their messaging in response to growing opposition, much like the U.S.S.R. and

[1] As of 2014, according to the most recent numbers from the Center for Migration Studies (Warren, 2016).

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 773–783
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-1072

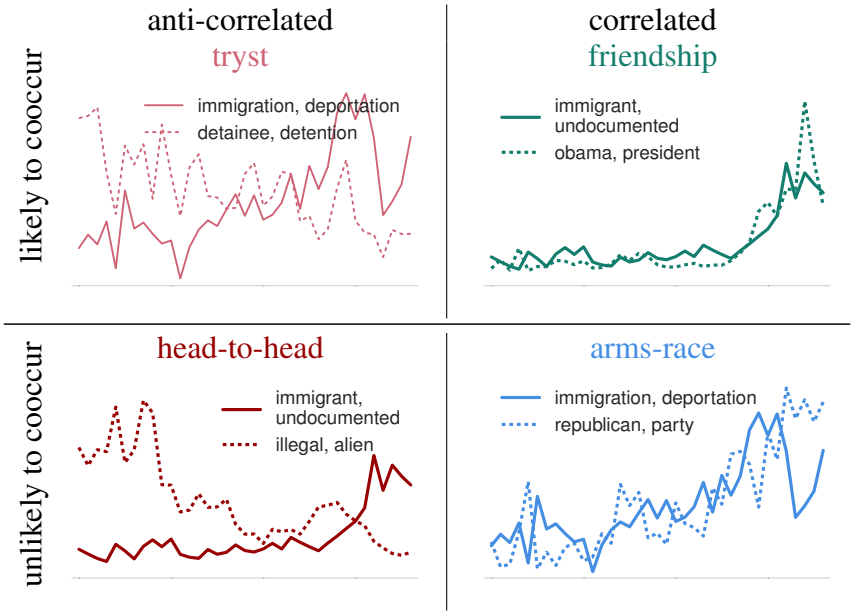


Figure 1: Relations between ideas in the space of cooccurrence and prevalence correlation (prevalence correlation is shown explicitly and cooccurrence is encoded in row captions). We use topics from LDA (Blei et al., 2003) to represent ideas. Each topic is named with a pair of words that are most strongly associated with the topic in LDA. Subplots show examples of relations between topics found in U.S. newspaper articles on immigration from 1980 to 2016, color coded to match the description in text. The y-axis represents the proportion of news articles in a year (in our corpus) that contain the corresponding topic. All examples are among the top 3 strongest relations in each type except ("immigrant, undocumented", "illegal, alien"), which corresponds to the two competing narratives. We explain the formal definition of strength in §2.

the U.S. during the cold war. To capture these possibilities, we use prevalence correlation over time.

Building on these insights, we propose a framework that combines cooccurrence within documents and prevalence correlation over time. This framework gives rise to four possible types of relation that correspond to the four quadrants in Fig. 1. We explain each type using examples from news articles in U.S. newspapers on immigration from 1980 to 2016. Here, we have used LDA to identify ideas in the form of topics, and we denote each idea with a pair of words most strongly associated with the corresponding topic.

Friendship (correlated over time, likely to cooccur). The "immigrant, undocumented" topic tends to cooccur with "obama, president" and both topics have been rising during the period of our dataset, likely because the "undocumented immigrants" narrative was an important part of Obama's framing of the immigration issue (Haynes et al., 2016).

Head-to-head (anti-correlated over time, unlikely to cooccur). "immigrant, undocumented" and "illegal, alien" are in a head-to-head competition: these two topics rarely cooccur, and "immigrant, undocumented" has been growing in prevalence, while the usage of "illegal, alien" in newspapers has been declining. This observation agrees with a report from Pew Research Center (Guskin, 2013).

Tryst (anti-correlated over time, likely to cooccur). The two off-diagonal examples use topics related to law enforcement. Overall, "immigration, deportation" and "detention, jail" often cooccur but "detention, jail" has been declining, while "immigration, deportation" has been rising. This possibly relates to the promises to overhaul the immigration detention system (Kalhan, 2010).[2]

Arms-race (correlated over time, unlikely to cooccur). One of the above law enforcement topics ("immigration, deportation") and a topic on the Republican party ("republican, party") hold an arms-race relation: they are both growing in prevalence over time, but rarely cooccur, perhaps suggesting an underlying common cause.

[2]The tryst relation is the least intuitive, yet is nevertheless observed. The pattern of being anti-correlated yet likely to cooccur is typically found when two ideas exhibit a friendship pattern (cooccurring and correlated), but only briefly, and then diverge.

Note that our terminology describes the relations between ideas in texts, not necessarily between the entities to which the ideas refer. For example, we find that the relation between "Israel" and "Palestine" is "friendship" in news articles on terrorism, based on their prevalence correlation and cooccurrence in that corpus.

We introduce the formal definition of our framework in §2 and apply it to news articles on five issues and research papers from ACL Anthology and NIPS as testbeds. We operationalize ideas using topics (Blei et al., 2003) and keywords (Monroe et al., 2008).

To explore whether the four relation types exist and how strong these relations are, we first examine the marginal and joint distributions of cooccurrence and prevalence correlation (§3). We find that cooccurrence shows a unimodal normal-shaped distribution but prevalence correlation demonstrates more diverse distributions across corpora. As we would expect, there are, in general, more and stronger friendship and head-to-head relations than arms-race and tryst relations.

Second, we demonstrate the effectiveness of our framework through in-depth case studies (§4). We not only validate existing knowledge about some news issues and research areas, but also identify hypotheses that require further investigation. For example, using keywords to represent ideas, a top pair with the tryst relation in news articles on terrorism is "arab" and "islam"; they are likely to cooccur, but "islam" is rising in relative prevalence while "arab" is declining. This suggests a conjecture that the news media have increasingly linked terrorism to a religious group rather than an ethnic group. We also show relations between topics in ACL that center around machine translation.

Our work is a first step towards understanding relations between ideas from text corpora, a complex and important research question. We provide some concluding thoughts in §6.

2 Computational Framework

The aim of our computational framework is to explore *relations* between ideas. We thus assume that the set of relevant ideas has been identified, and those expressed in each document have been tabulated. Our open-source implementation is at `https://github.com/Noahs-ARK/idea_relations/`. In the following, we introduce our formal definitions and datasets.

$$\forall x, y \in \mathcal{I}, \widehat{\mathrm{PMI}}(x,y) = \log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)}$$
$$= C + \log \frac{1+\sum_t \sum_k \mathbf{1}\{x \in d_{t_k}\} \cdot \mathbf{1}\{y \in d_{t_k}\}}{\left(1+\sum_t \sum_k \mathbf{1}\{x \in d_{t_k}\}\right) \cdot \left(1+\sum_t \sum_k \mathbf{1}\{y \in d_{t_k}\}\right)} \quad (1)$$

$$\hat{r}(x,y) = \frac{\sum_t \left(\hat{P}(x|t) - \overline{\hat{P}(x|t)}\right)\left(\hat{P}(y|t) - \overline{\hat{P}(y|t)}\right)}{\sqrt{\sum_t \left(\hat{P}(x|t) - \overline{\hat{P}(x|t)}\right)^2} \sqrt{\sum_t \left(\hat{P}(y|t) - \overline{\hat{P}(y|t)}\right)^2}} \quad (2)$$

Figure 2: Eq. 1 is the empirical pointwise mutual information for two ideas, our measure of cooccurrence of ideas; note that we use add-one smoothing in estimating PMI. Eq. 2 is the Pearson correlation between two ideas' prevalence over time.

2.1 Cooccurrence and Prevalence Correlation

As discussed in the introduction, we focus on two dimensions to quantify relations between ideas:

1. cooccurrence reveals to what extent two ideas tend to occur in the same contexts;

2. similarity between the relative prevalence of ideas over time reveals how two ideas relate in terms of popularity or coverage.

Our input is a collection of documents, each represented by a set of ideas and indexed by time. We denote a *static* set of ideas as $\mathcal{I}$ and a text corpus that consists of these ideas as $C = \{D_1, \ldots, D_T\}$, where $D_t = \{d_{t_1}, \ldots, d_{t_{N_t}}\}$ gives the collection of documents at timestep t, and each document, d_{t_k}, is represented as a subset of ideas in $\mathcal{I}$. Here T is the total number of timesteps, and N_t is the number of documents at timestep t. It follows that the total number of documents $N = \sum_{t=1}^{T} N_t$.

In order to formally capture the two dimensions above, we employ two commonly-used statistics. First, we use empirical pointwise mutual information (PMI) to capture the cooccurrence of ideas within the same document (Church and Hanks, 1990); see Eq. 1 in Fig. 2. Positive $\widehat{\mathrm{PMI}}$ indicates that ideas occur together more frequently than would be expected if they were independent, while negative $\widehat{\mathrm{PMI}}$ indicates the opposite.

Second, we compute the correlation between normalized document frequency of ideas to capture the relation between the relative prevalence of ideas across documents over time; see Eq. 2 in Fig. 2. Positive $\hat{r}$ indicates that two ideas have similar prevalence over time, while negative $\hat{r}$ sug-

gests two anti-correlated ideas (i.e., when one goes up, the other goes down).

The four types of relations in the introduction can now be obtained using $\widehat{\text{PMI}}$ and $\hat{r}$, which capture cooccurrence and prevalence correlation respectively. We further define the *strength* of the relation between two ideas as the absolute value of the product of their $\widehat{\text{PMI}}$ and $\hat{r}$ scores:

$$\forall x, y \in \mathcal{I}, \text{strength}(x, y) = |\widehat{\text{PMI}}(x, y) \times \hat{r}(x, y)|. \quad (3)$$

2.2 Datasets and Representation of Ideas

We use two types of datasets to validate our framework: news articles and research papers. We choose these two domains because competition between ideas has received significant interest in history of science (Kuhn, 1996) and research on framing (Chong and Druckman, 2007; Entman, 1993; Gitlin, 1980; Lakoff, 2014). Furthermore, interesting differences may exist in these two domains as news evolves with external events and scientific research progresses through innovations.

- News articles. We follow the strategy in Card et al. (2015) to obtain news articles from LexisNexis on five issues: abortion, immigration, same-sex marriage, smoking, and terrorism. We search for relevant articles using LexisNexis subject terms in U.S. newspapers from 1980 to 2016. Each of these corpora contains more than 25,000 articles. Please refer to the supplementary material for details.
- Research papers. We consider full texts of papers from two communities: our own ACL community captured by papers from ACL, NAACL, EMNLP, and TACL from 1980 to 2014 (Radev et al., 2009); and the NIPS community from 1987 to 2016.[3] There are 4.8K papers from the ACL community and 6.6K papers from the NIPS community. The processed datasets are available at `https://chenhaot.com/pages/idea-relations.html`.

In order to operationalize ideas in a text corpus, we consider two ways to represent ideas.

- Topics. We extract topics from each document by running LDA (Blei et al., 2003) on each corpus C. In all datasets, we set the number of topics to 50.[4] Formally, $\mathcal{I}$ is the 50 topics learned from the corpus, and each document is represented as the set of topics that are present with greater than 0.01 probability in the topic distribution for that document.
- Keywords. We identify a list of distinguishing keywords for each corpus by comparing its word frequencies to the background frequencies found in other corpora using the informative Dirichlet prior model in Monroe et al. (2008). We set the number of keywords to 100 for all corpora. For news articles, the background corpus for each issue is comprised of all articles from the other four issues. For research papers, we use NIPS as the background corpus for ACL and vice versa to identify what are the core concepts for each of these research areas. Formally, $\mathcal{I}$ is the 100 top distinguishing keywords in the corpus and each document is represented as the set of keywords within $\mathcal{I}$ that are present in the document. Refer to the supplementary material for a list of example keywords in each corpus.

In both procedures, we lemmatize all words and add common bigram phrases to the vocabulary following Mikolov et al. (2013). Note that in our analysis, ideas are only present or absent in a document, and a document can in principle be mapped to any subset of ideas in $\mathcal{I}$. In our experiments 90% of documents are marked as containing between 7 and 14 ideas using topics, 8 and 33 ideas using keywords.

3 Characterizing the Space of Relations

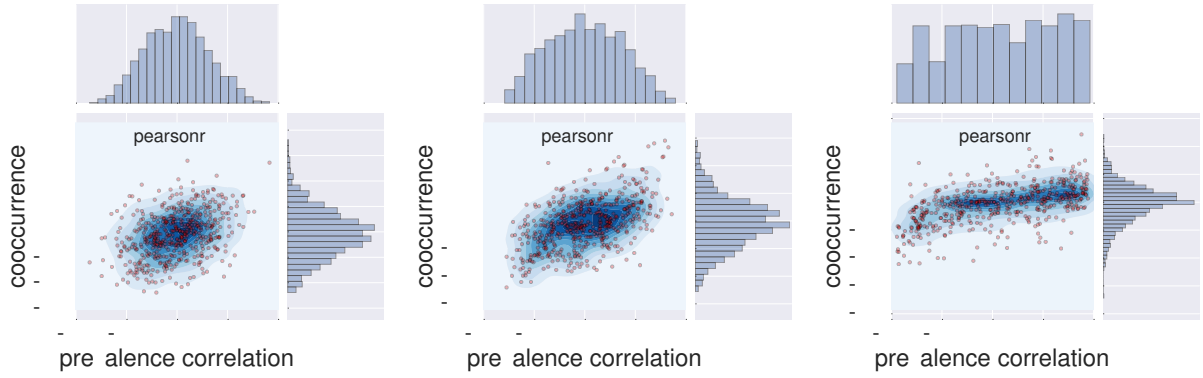
To provide an overview of the four relation types in Fig. 1, we first examine the empirical distributions of the two statistics of interest across pairs of ideas. In most exploratory studies, however, we are most interested in pairs that exemplify each type of relation, i.e., the most extreme points in each quadrant. We thus look at these pairs in each corpus to observe how the four types differ in salience across datasets.

3.1 Empirical Distribution Properties

To the best of our knowledge, the distributions of pairwise cooccurrence and prevalence correlation have not been examined in previous literature. We thus first investigate the marginal distributions of cooccurrence and prevalence correlation and then

[3] `http://papers.nips.cc/`.

[4] We chose 50 topics based on past experience, though this could be tuned for particular applications. Recall that our framework is to analyze *relations* between ideas, so this choice is not essential in this work.



(a) Terrorism topics (b) Immigration topics (c) ACL topics

Figure 3: 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 these plots, we use topics to represent ideas.

their joint distribution. Fig. 3 shows three examples: two from news articles and one from research papers. We will also focus our case studies on these three corpora in §4. The corresponding plots for keywords have been relegated to supplementary material due to space limitations.

Cooccurrence tends to be unimodal but not normal. In all of our datasets, pairwise cooccurrence ($\widehat{\text{PMI}}$) presents a unimodal distribution that somewhat resembles a normal distribution, but it is rarely precisely normal. We cannot reject the hypothesis that it is unimodal for any dataset (using topics or keywords) using the dip test (Hartigan and Hartigan, 1985), though D'Agostino's K^2 test (D'Agostino et al., 1990) rejects normality in almost all cases.

Prevalence correlation exhibits diverse distributions. Pairwise prevalence correlation follows different distributions in news articles compared to research papers: they are unimodal in news articles, but not in ACL or NIPS. The dip test only rejects the unimodality hypothesis in NIPS. None follow normal distributions based on D'Agostino's K^2 test.

Cooccurrence is positively correlated with prevalence correlation. In all of our datasets, cooccurrence is positively correlated with prevalence correlation whether we use topics or keywords to represent ideas, although the Pearson correlation coefficients vary. This suggests that there are more friendship and head-to-head relations than tryst and arms-race relations. Based on the results of kernel density estimation, we also observe that this correlation is often loose, e.g., in ACL topics, cooccurrence spreads widely at each mode of prevalence correlation.

3.2 Relative Strength of Extreme Pairs

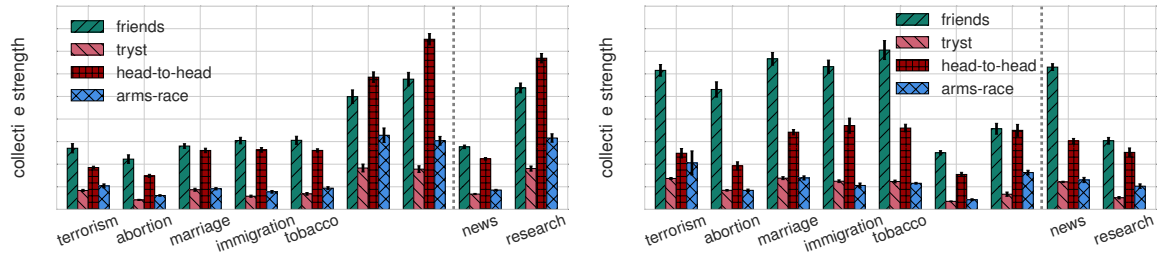
We are interested in how our framework can identify intriguing relations between ideas. These potentially interesting pairs likely correspond to the extreme points in each quadrant instead of the ones around the origin, where PMI and prevalence correlation are both close to zero. Here we compare the relative strength of extreme pairs in each dataset. We will discuss how these extreme pairs confirm existing knowledge and suggest new hypotheses via case studies in §4.

For each relation type, we average the strengths of the 25 pairs with the strongest relations in that type, with strength defined in Eq. 3. This heuristic (henceforth *collective strength*) allows us to collectively compare the strengths of the most prominent friendship, tryst, arms-race, and head-to-head relations. The results are not sensitive to the choice of 25.

Fig. 4 shows the collective strength of the four types in all of our datasets. The most common ordering is:

$$\text{friendship} > \text{head-to-head} > \text{arms-race} > \text{tryst}.$$

The fact that friendship and head-to-head relations are strong is consistent with the positive correlation between cooccurrence and prevalence correlation. In news, friendship is the strongest relation type, but head-to-head is the strongest in ACL topics and NIPS topics. This suggests, unsurprisingly, that there are stronger head-to-head competitions



(a) Topics

(b) Keywords

Figure 4: Collective strength of the four relation types in each dataset (*news* is the average of the news corpora and *research* is for ACL and NIPS). Fig. 4a uses topics to represent ideas, while Fig. 4b uses keywords to represent ideas. Each bar presents the average strength of the top 25 pairs in a relation type in the corresponding dataset. Error bars represent standard errors calculated in the usual way, but note that since the top 25 pairs are not random samples, they cannot be interpreted in the usual way.

(i.e., one idea takes over another) between ideas in scientific research than in news. We also see that topics show greater strength in our scientific article collections, while keywords dominate in news, especially in friendship. We conjecture that terms in scientific literature are often overloaded (e.g., a *tree* could be a parse tree or a decision tree), necessitating some abstraction when representing ideas. In contrast, news stories are more self-contained and seek to employ consistent usage.

4 Exploratory Studies

We present case studies based on strongly related pairs of ideas in the four types of relation. Throughout this section, "rank" refers to the rank of the relation strength between a pair of ideas in its corresponding relation type.

4.1 International Relations in Terrorism

Following a decade of declining violence in the 90s, the events of September 11, 2001 precipitated a dramatic increase in concern about terrorism, and a major shift in how it was framed (Kern et al., 2003). As a showcase, we consider a topic which encompasses much of the U.S. government's response to terrorism: "federal, state".[5] We observe two topics engaging in an "arms race" with this one: "afghanistan, taliban" and "pakistan, india". These correspond to two geopolitical regions closely linked to the U.S. government's concern with terrorism, and both were sites of U.S. military action during the period of our dataset. Events abroad and the U.S. government's responses follow the arms-race pattern, each holding increasing

[5]As in §1, we summarize each topic using a pair of strongly associated words, instead of assigning a name.

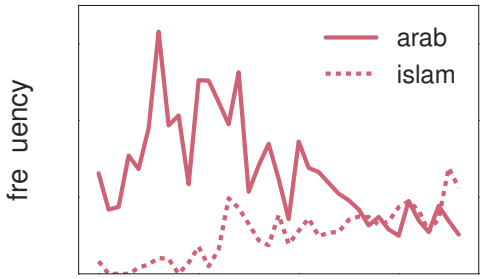
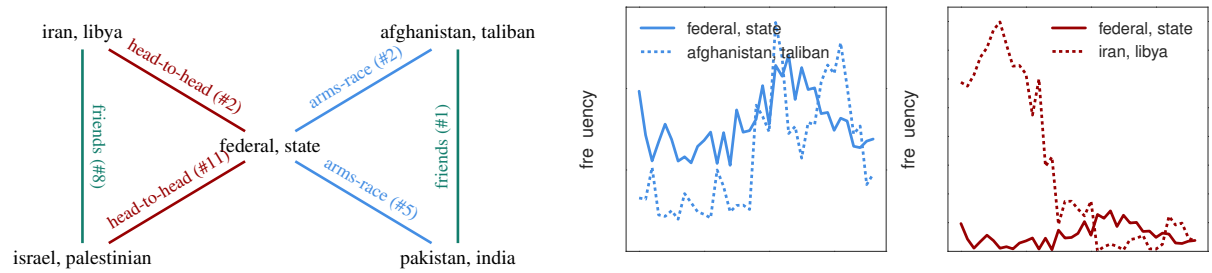


Figure 6: Tryst relation between arab and islam using keywords to represent ideas (#2 in tryst): these two words tend to cooccur but are anti-correlated in prevalence over time. In particular, islam was rarely used in coverage of terrorism in the 1980s.

attention with the other, likely because they share the same underlying cause.

We also observe two head-to-head rivals to the "federal, state" topic: "iran, libya" and "israel, palestinian". While these topics correspond to regions that are hotly debated in the U.S., their coverage in news tends not to correlate temporally with the U.S. government's responses to terrorism, at least during the time period of our corpus. Discussion of these regions was more prevalent in the 80s and 90s, with declining media coverage since then (Kern et al., 2003).

The relations between these topics are consistent with structural balance theory (Cartwright and Harary, 1956; Heider, 1946), which suggests that the enemy of an enemy is a friend. The "afghanistan, taliban" topic has the strongest friendship relation with the "pakistan, india" topic, i.e., they are likely to cooccur and are positively correlated in prevalence. Similarly, the "iran, libya" topic is a close "friend" with the "israel, palestinian" topic (ranked 8th in friendship).



(a) Relations between a United States topic and international topics.

(b) ("federal, state", "afghanistan, taliban")

(c) ("federal, state", "iran, libya")

Figure 5: Fig. 5a shows the relations between the "federal, state" topic and four international topics. Edge colors indicate relation types and the number in an edge label presents the ranking of its strength in the corresponding relation type. Fig. 5b and Fig. 5c represent concrete examples in Fig. 5a: "federal, state" and "afghanistan, taliban" follow similar trends, although "afghanistan, taliban" fluctuates over time due to significant events such as the September 11 attacks in 2001 and the death of Bin Laden in 2011; while "iran, lybia" is negatively correlated with "federal, state". In fact, more than 70% of terrorism news in the 80s contained the "iran, lybia" topic.

When using keywords to represent ideas, we observe similar relations between the term homeland security and terms related to the above foreign countries. In addition, we highlight an interesting but unexpected tryst relation between arab and islam (Fig. 6). It is not surprising that these two words tend to cooccur in the same news articles, but the usage of islam in the news is increasing while arab is declining. The increasing prevalence of islam and decreasing prevalence of arab over this time period can also be seen, for example, using Google's n-gram viewer, but it of course provides no information about cooccurrence.

This trend has not been previously noted to the best of our knowledge, although an article in the *Huffington Post* called for news editors to distinguish Muslim from Arab.[6] Our observation suggests a conjecture that the news media have increasingly linked terrorism to a religious group rather than an ethnic group, perhaps in part due to the tie between the events of 9/11 and Afghanistan, which is not an Arab or Arabic-speaking country. We leave it to further investigation to confirm or reject this hypothesis.

To further demonstrate the effectiveness of our approach, we compare a pair's rank using only cooccurrence or prevalence correlation with its rank in our framework. Table 1 shows the results for three pairs above. If we had looked at only cooccurrence or prevalence correlation, we would probably have missed these interesting pairs.

[6] http://www.huffingtonpost.com/haroon-moghul/even-the-new-york-times-d_b_766658.html

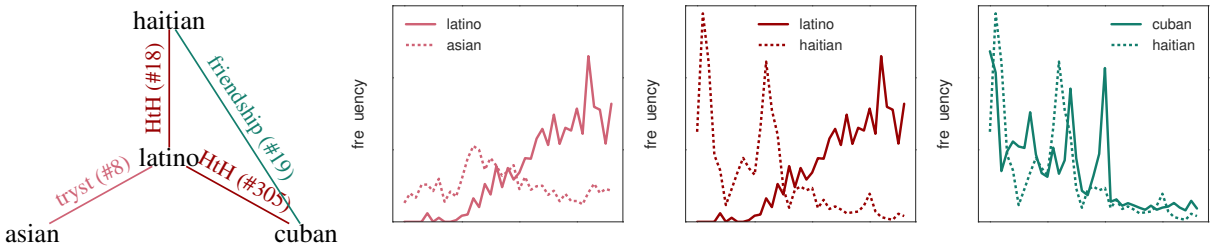
	PMI	Corr
"federal, state", "afghanistan, taliban" (#2 in arms-race)	43	99
"federal, state", "iran, lybia" (#2 in head-to-head)	36	56
arab, islam (#2 in tryst)	106	1,494

Table 1: Ranks of pairs by using the absolute value of only cooccurrence or prevalence correlation.

4.2 Ethnicity Keywords in Immigration

In addition to results on topics in §1, we observe unexpected patterns about ethnicity keywords in immigration news. Our observation starts with a top tryst relation between latino and asian. Although these words are likely to cooccur, their prevalence trajectories differ, with the discussion of Asian immigrants in the 1990s giving way to a focus on the word latino from 2000 onward. Possible theories to explain this observation include that undocumented immigrants are generally perceived as a Latino issue, or that Latino voters are increasingly influential in U.S. elections.

Furthermore, latino holds head-to-head relations with two subgroups of Latin American immigrants: haitian and cuban. In particular, the strength of the relation with haitian is ranked #18 in head-to-head relations. Meanwhile, haitian and cuban have a friendship relation, which is again consistent with structural balance theory. The decreasing prevalence of haitian and cuban perhaps speaks to the shifting geographical focus of recent immigration to the U.S., and issues of the Latino pan-ethnicity. In fact, a majority of Latinos prefer to identify with their national origin relative to the



(a) Relations graph. (b) (latino, asian) (c) (latino, haitian) (d) (cuban, haitian)

Figure 7: Relations between ethnicity keywords in immigration news (HtH for head-to-head): latino holds a tryst relation with asian and head-to-head relations with two subgroups from Latin America, haitian and cuban. We do not show the relations between asian and haitian, cuban, because their strength is close to 0.

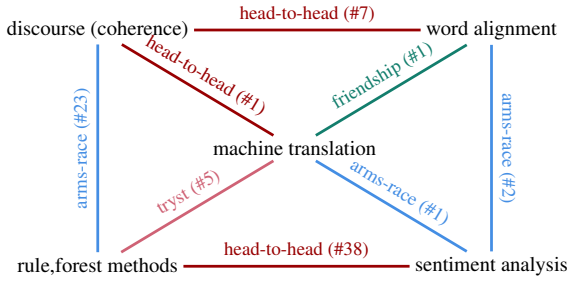


Figure 8: Top relations between the topics in ACL Anthology. The top 10 words for the rule, forest methods topic are *rule, grammar, derivation, span, algorithm, forest, parsing, figure, set, string*.

pan-ethnic terms (Taylor et al., 2012). However, we should also note that much of this coverage relates to a set of specific refugee crises, temporarily elevating the political importance of these nations in the U.S. Nevertheless, the underlying social and political reasons behind these head-to-head relations are worth further investigation.

4.3 Relations between Topics in ACL

Finally, we analyze relations between topics in the ACL Anthology. It turns out that "machine translation" is at a central position among top ranked relations in all the four types (Fig. 8).[7] It is part of the strongest relation in all four types except tryst (ranked #5).

The full relation graph presents further patterns. Friendship demonstrates transitivity: both "machine translation" and "word alignment" have similar relations with other topics. But such transitivity does not hold for tryst: although the prevalence of "rule, forest methods" is anti-correlated with both "machine translation" and "sentiment analysis", "sentiment analysis" seldom cooccurs with "rule, forest methods" because "sentiment analysis" is seldom built on parsing algorithms. Similarly, "rule, forest methods" and "discourse (coherence)" hold an arms-race relation: they do not tend to cooccur and both decline in relative prevalence as "machine translation" rises.

The prevalence of each of these ideas in comparison to machine translation is shown in in Fig. 9, which reveals additional detail.

[7] In the ranking, we filtered a topic where the top words are *ion, ing, system, process, language, one, input, natural language, processing, grammar*. For the purposes of this corpus, this is effectively a stopword topic.

5 Related Work

We present two strands of related studies in addition to what we have discussed.

Trends in ideas. Most studies have so far examined the trends of ideas individually (Michel et al., 2011; Hall et al., 2008; Rule et al., 2015). For instance, Hall et al. (2008) present various trends in our own computational linguistics community, including the rise of statistical machine translation. More recently, rhetorical framing has been used to predict these sorts of patterns (Prabhakaran et al., 2016). An exception is that Shi et al. (2010) use prevalence correlation to analyze lag relations between topics in publications and research grants. Anecdotally, Grudin (2009) observes a "head-to-head" relation between artificial intelligence and human-computer interaction in research funding. However, to our knowledge, our work is the first study to systematically characterize relations *between* ideas.

Representation of ideas. In addition to topics and keywords, studies have also sought to operationalize the "memes" metaphor using quotes and text reuse in the media (Leskovec et al., 2009; Niculae et al., 2015; Smith et al., 2013; Wei et al., 2013). In topic modeling literature, Blei and Lafferty (2006) also point out that topics do not cooccur independently and explicitly model the cooccurrence within documents.

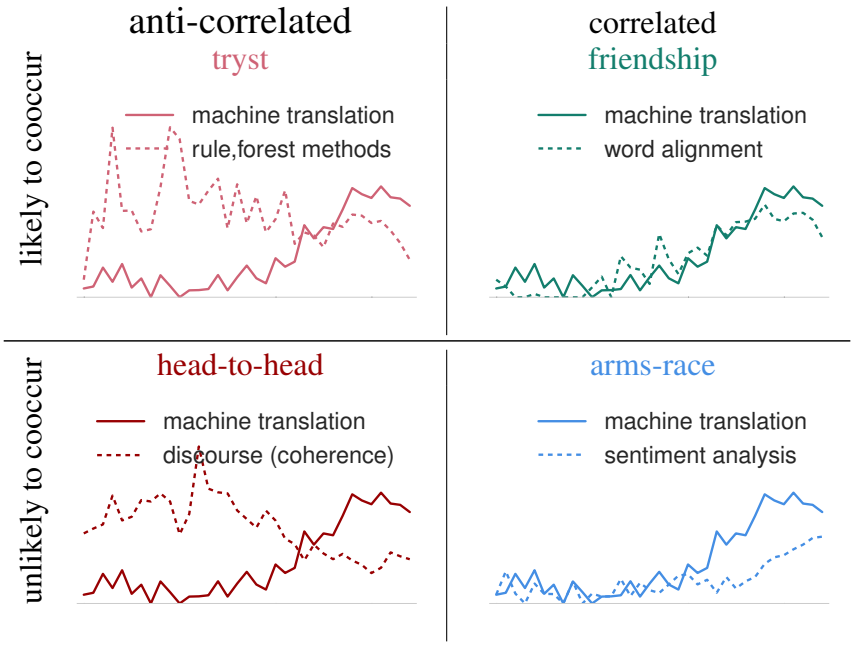


Figure 9: Relations between topics in ACL Anthology in the space of cooccurrence and prevalence correlation (prevalence correlation is shown explicitly and cooccurrence is encoded in row captions), color coded to match the text. The y-axis represents the relative proportion of papers in a year that contain the corresponding topic. The top 10 words for the rule, forest methods topic are *rule, grammar, derivation, span, algorithm, forest, parsing, figure, set, string*.

6 Concluding Discussion

We proposed a method to characterize relations between ideas in texts through the lens of cooccurrence within documents and prevalence correlation over time. For the first time, we observe that the distribution of pairwise cooccurrence is unimodal, while the distribution of pairwise prevalence correlation is not always unimodal, and show that they are positively correlated. This combination suggests four types of relations between ideas, and these four types are all found to varying extents in our experiments.

We illustrate our computational method by exploratory studies on news corpora and scientific research papers. We not only confirm existing knowledge but also suggest hypotheses around the usage of arab and islam in terrorism and latino and asian in immigration.

It is important to note that the relations found using our approach depend on the nature of the representation of ideas and the source of texts. For instance, we cannot expect relations found in news articles to reflect shifts in public opinion if news articles do not effectively track public opinion.

Our method is entirely observational. It remains as a further stage of analysis to understand the underlying reasons that lead to these relations between ideas. In scientific research, for example, it could simply be the progress of science, i.e., newer ideas overtake older ones deemed less valuable at a given time; on the other hand, history suggests that it is not always the correct ideas that are most expressed, and many other factors may be important. Similarly, in news coverage, underlying sociological and political situations have significant impact on which ideas are presented, and how.

There are many potential directions to improve our method to account for complex relations between ideas. For instance, we assume that both ideas and relations are statically grounded in keywords or topics. In reality, ideas and relations both evolve over time: a tryst relation might appear as friendship if we focus on a narrower time period. Similarly, new ideas show up and even the same idea may change over time and be represented by different words.

Acknowledgments. We thank Amber Boydstun, Justin Gross, Lillian Lee, anonymous reviewers, and all members of Noah's ARK for helpful comments and discussions. This research was made possible by a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship (to D.C.) and a University of Washington Innovation Award.

References

David M. Blei and John Lafferty. 2006. Correlated topic models. In *NIPS*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

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*.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: A generalization of Heider's theory. *Psychological Review* 63(5):277.

Dennis Chong and James N. Druckman. 2007. A theory of framing and opinion formation in competitive elite environments. *Journal of Communication* 57(1):99–118.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Ralph B. D'Agostino, Albert Belanger, and Ralph B. D'Agostino Jr. 1990. A suggestion for using powerful and informative tests of normality. *The American Statistician* 44(4):316–321.

Richard Dawkins. 1976. *The Selfish Gene*. Oxford University Press.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4):51–58.

Todd Gitlin. 1980. *The Whole World is Watching: Mass Media in the Making and Unmaking of the New Left*. Berkeley: University of California Press.

Jonathan Grudin. 2009. AI and HCI: Two fields divided by a common focus. *AI Magazine* 30(4):48.

Emily Guskin. 2013. 'Illegal', 'undocumented', 'unauthorized': News media shift language on immigration. Pew Research Center.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of EMNLP*.

John A. Hartigan and P. M. Hartigan. 1985. The dip test of unimodality. *The Annals of Statistics* pages 70–84.

Chris Haynes, Jennifer L. Merolla, and S. Karthick Ramakrishnan. 2016. *Framing Immigrants: News Coverage, Public Opinion, and Policy*. Russell Sage Foundation.

Fritz Heider. 1946. Attitudes and cognitive organization. *The Journal of Psychology* 21(1):107–112.

Anil Kalhan. 2010. Rethinking immigration detention. *Columbia Law Review Sidebar* 110:42.

Montague Kern, Marion Just, and Pippa Norris. 2003. The lessons of framing terrorism. In Pippa Norris, Montague Kern, and Marion Just, editors, *Framing Terrorism: The News Media, the Government and the Public*, Routledge.

Thomas S. Kuhn. 1996. *The Structure of Scientific Revolutions*. University of Chicago Press.

George Lakoff. 2014. *The All New Don't Think of an Elephant!: Know your Values and Frame the Debate*. Chelsea Green Publishing.

Jure Leskovec, Lars Backstrom, and Jon M. Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of KDD*.

Jennifer Merolla, S. Karthick Ramakrishnan, and Chris Haynes. 2013. "Illegal,", "undocumented," or "unauthorized": Equivalency frames, issue frames, and public opinion on immigration. *Perspectives on Politics* 11(03):789–807.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

John Milton. 1644. *Areopagitica, A speech of Mr. John Milton for the Liberty of Unlicenc'd Printing to the Parliament of England*.

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.

Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of WWW*.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of ACL*.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL anthology network corpus. In *Proceedings of ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. 2015. Lexical shifts, substantive changes, and

continuity in state of the union discourse, 1790-2014. *Proceedings of the National Academy of Sciences* 112(35):10837–10844.

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

Xiaolin Shi, Ramesh Nallapati, Jure Leskovec, Dan McFarland, and Dan Jurafsky. 2010. Who leads whom: Topical lead-lag analysis across corpora. In *Proceedings of NIPS Workshop on Computational Social Science*.

David A. Smith, Ryan Cordell, and Elizabeth M. Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *Proceedings of the Workshop on Big Humanities*.

Paul Taylor, Mark H. Lopez, Jessica Martínez, and Gabriel Velasco. 2012. When labels don't fit: Hispanics and their views of identity. *Washington, DC: Pew Hispanic Center* .

Robert Warren. 2016. US undocumented population drops below 11 million in 2014, with continued declines in the Mexican undocumented population. *Journal on Migration and Human Security* 4(1):1–15.

Xuetao Wei, Nicholas Valler, B. Aditya Prakash, Iulian Neamtiu, Michalis Faloutsos, and Christos Faloutsos. 2013. Competing memes propagation on networks: A network science perspective. *IEEE Journal on Selected Areas in Communications* 31:1049–1060.