iCoRe: The GDELT Interface for the Advancement of Communication Research

Preprint · May 2019

CITATIONS
0

READS
53

4 authors:

Frederic R. Hopp
University of California, Santa Barbara
17 PUBLICATIONS **51** CITATIONS
SEE PROFILE

James Schaffer
Sysco Corporation
25 PUBLICATIONS **85** CITATIONS
SEE PROFILE

Jacob T Fisher
University of California, Santa Barbara
14 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Rene Weber
University of California, Santa Barbara
116 PUBLICATIONS **1,650** CITATIONS
SEE PROFILE

iCoRe: The GDELT Interface for the Advancement of Communication Research

Frederic R. Hopp[1]

James Schaffer[2]

Jacob T. Fisher[1]

Rene Weber[1]*

[1]University of California, Santa Barbara – Department of Communication

Media Neuroscience Lab

[2]Sysco Labs, Sysco Corporation

* Please address correspondence to René Weber (renew@comm.ucsb.edu); University of California Santa Barbara, Department of Communication - Media Neuroscience Lab, Santa Barbara, CA 93106-4020 (renew@comm.ucsb.edu)

Abstract

This article introduces the interface for communication research (iCoRe) to access, explore, and analyze the Global Database of Events, Language and Tone (GDELT; Leetaru & Schrodt, 2013). GDELT provides a vast, open source, and continuously updated repository of online news and event metadata collected from tens of thousands of news outlets around the world. Despite GDELT's promise for advancing communication science, its massive scale and complex data structures have hindered efforts of communication scholars aiming to access and analyze GDELT. We thus developed iCoRe, an easy-to-use web interface that (a) provides fast access to the data available in GDELT, (b) shapes and processes GDELT for theory-driven applications within communication research, and (c) enables replicability through transparent query and analysis protocols. After providing an overview of how GDELT's data pertain to addressing communication research questions, we provide a tutorial of utilizing iCoRe across three theory-driven case studies. We conclude this article with a discussion and outlook of iCoRe's future potential for advancing communication research.

Keywords: online news, event data, big data, global database of events language and tone, computational social science, open science

iCoRe: The GDELT Interface for the Advancement of Communication Research

The study of news has enjoyed a long and fruitful history within communication scholarship. Many of the most notable theories to emerge from communication research within the last half century have been largely based on the study of news and journalism. These include framing (Entman, 1993; Scheufele, 1999), agenda-setting (McCombs & Shaw, 1972; McCombs, 2005), the spiral of silence (Noelle-Neumann, 1974), and many others. The advent of online journalism has brought with it an explosion in the amount of data that are available to communication scholars. At the same time, the increasing quantity and ephemerality of news data creates challenges for researchers using sampling methods and analysis techniques developed for use with more traditional media sources (Hester & Dougall, 2007; Hopkins & King, 2010; Grimmer & Stewart, 2013). For this reason, communication scholars are increasingly turning to large scale, open source datasets to assist in filtering, document selection, and analysis of news data (M. Weber, 2018).

The increasing size and diversity of these news datasets presents numerous challenges and opportunities for computational communication research (CCR; Van Atteveldt & Peng 2018). As communication scholars navigate through this novel data landscape, they are frequently required to determine the value of a particular dataset for advancing communication research. This endeavor is often compromised by several factors: (a) lack of documentation regarding how data in the dataset are collected or stored (b) the presence of complex, unprocessed data structures and noise, and (c) lack of access to or knowledge of the tools that are needed to ingest, parse, and organize the data for scientific work. Furthermore, method sections describing how particular large-scale datasets were acquired, preprocessed, and analyzed are

frequently not comprehensive enough to allow for scientific replication (Peng, 2011). Without clearer documentation, open-access interfaces, and collaborative research efforts, CCR is at risk to be constrained to a few privileged researchers (Huberman, 2012) and may be subject to increased replicability problems in the future (Wallach, 2016).

With these issues in mind, we introduce the Global Database of Events, Language, and Tone (GDELT; Leetaru & Schrodt, 2013) in conjunction with our newly developed GDELT interface for Communication Research (iCoRe). GDELT is an open-source, large-scale repository of global online news and events metadata collected from diverse media formats and updated in near real-time. GDELT has sparked considerable interest within cognate social science disciplines, for example, to forecast and understand the latent structure of political unrest and violence in East Asia (Qiao et al., 2017) and the Middle East (Smith, Smith, Legg & Francis, 2017). Despite its clear utility for addressing these critical questions, GDELT has been used by only a handful of communication scholars for conducting theory-driven research. These studies are primarily aimed at harnessing GDELT as a data source to aid the selection and analysis of news articles for subsequent human annotation (R. Weber. et al., 2018, study 6). Furthermore, GDELT has proven useful for advancing research on agenda-setting (Guo & Vargo, 2017; Vargo & Guo, 2017) and fake-news (Guo & Vargo, 2018; Vargo, Guo, & Amazeen, 2018).

We argue that GDELT's limited use within communication research can at least partially be attributed to several pitfalls common to many large-scale datasets. These shortcomings have increasingly been voiced within the broader computational social science literature (e.g., boyd & Crawford, 2012; Lazer et al., 2009; Van Atteveldt & Peng 2018; Wallach, 2016): First, although GDELT's datasets are *de facto* "open-access," the sheer size, complex structure, and opaque documentation of these datasets pose significant challenges to communication researchers trying

to query, parse, and explore GDELT. Because of this, GDELT is a clear case in which *availability* of data does not assure *accessibility*. For example, parsing GDELT's raw, unstructured Global Content Analysis Measures (GCAM) requires significant amounts of time, advanced computational programming and pattern matching skills, along with the technical knowledge of how to store and wrangle these massive amounts of data. Thus, an interface is required in order to grant communication researchers without a strong computational background access to GDELT, thereby flattening the hierarchy around "who can read the numbers" in large datasets (boyd & Crawford, 2012).

Second, scripts and algorithms utilized to query and parse GDELT have to this point largely remained closed "black boxes" solely available and interpretable to the user or research group that develops these tools. This hinders endeavors to comprehend and replicate studies that have drawn on GDELT's datasets. Hence, making GDELT accessible to communication scholarship necessitates transparent guidelines on *how* its data are being accessed and analyzed, replacing opaque scripts and algorithms with replicable, comprehensive data-analytic pipelines (Trilling & Jonkman, 2018).

Third, similar to other large-scale datasets, GDELT was not initially developed and tailored with the purpose of advancing communication scholarship, but rather to serve as a "passive crowdsourcing" platform (Leetaru, 2011) to monitor global societal changes. Thus, data processing pipelines are needed to shape GDELT's datasets in a manner that meets the requirements of communication scholars aiming to address theory-driven research questions. In fact, choosing the operations that are employed in a preprocessing pipeline for large-scale web archives is often a critical research question in its own right, with its own theoretical considerations (M. Weber, 2018). In the same vein, an increasing number of large-scale news

and event datasets have recently become available. Prominent platforms include MediaCloud (https://mediacloud.org/), Crimson Hexagon (https://www.crimsonhexagon.com/), Phoenix (http://eventdata.utdallas.edu/), and the Integrated Crisis Early Warning System (ICEWS; https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html). While an extensive review of these datasets is beyond the confines of this paper, we argue that GDELT has several advantages over these datasets. First, GDELT enables users to download the automatically derived news article metadata (see Global Knowledge Graph below) for conducting independent, research-driven analyses. In contrast, other platforms (e.g., MediaCloud) only provide the output of a few available analyses without granting access to the underlying news data. Second, GDELT's unprecedented Global Content Analysis Measurement (GCAM) system provides the output of over 40 automatic content analytic measures. To the best of our knowledge, no other available news dataset administers such an extensive content analytical data stream. Third, GDELT processes global online news sources daily at 15-minute intervals and thus offers an unparalleled spatiotemporal resolution. Fourth, GDELT provides access to both news and event metadata. While other platforms (e.g., ICEWS, Phoenix) rely on the same Conflict and Mediation Event Observations codebook (CAMEO; Gerner, Schrodt, Yilmaz, & Abu-Jabr, 2002) to extract events from news records, GDELT is unprecedented in providing both news and event metadata.

Despite these advantages that GDELT offers, without a research agenda that adopts a strong synergy between method and theory (Greenwald, 2012), granting mere accessibility to GDELT risks simply stacking up the "tool pile" that showcases data-mining capabilities while neglecting the advancement of knowledge for communication research (see "better science versus better engineering"; Lin, 2015).

In order to leverage the strengths of GDELT while taking measures to avoid these pitfalls, we have developed the GDELT interface for Communication Research (iCoRe). iCoRe is an easy-to-use web interface specifically tailored to diminish the aforementioned challenges of harnessing GDELT for communication research. In this manuscript, we describe three advantages of iCoRe that emphasize its utility for communication research. First, iCoRe makes GDELT's data *accessible* to scholars independent of their computational skills or technical knowledge. Second, iCoRe encourages the replication of analyses conducted using the system by providing transparent, standardized query and analysis protocols. Third, iCoRe is designed from a theory-driven perspective to allow the advancement of pressing communication research questions.

In the following sections, we accomplish three primary goals. First, we provide an overview of GDELT's datasets as they pertain to addressing communication research questions. Second, we introduce iCoRe and provide usage guidelines. Finally, we present three theory-driven case studies that draw on iCoRe. These case studies serve to illustrate the capability of iCoRe for addressing pertinent communication questions. We conclude with a discussion of iCoRe's future potential to advance communication research by bridging large-scale, macro-level analyses and controlled, micro-level studies.

GDELT: An Introduction for Communication Researchers

The Global Database of Events, Language, and Tone (GDELT; Leetaru & Schrodt, 2013) arose out of a collaborative project between Google Jigsaw and the Center for Social Complexity at George Mason University. GDELT monitors tens of thousands of news websites around the globe and automatically extracts entities, themes, quotes, sentiment, images, and events present

in articles posted on these sites. Article metadata is archived at 15-minute time intervals 24 hours a day. In addition to analyzing text content from monitored sites, GDELT transcribes video and audio broadcasts from monitored sites, and translates them from 65 different languages into English text. Further, the database currently contains over a quarter of a billion unique event references and adds over three quarters of a trillion sentiment assessments, 1.5 billion location references, and 70 billion images on a yearly basis, made possible by the overall global increase in the capacity to store, communicate, and compute information (Hilbert & López, 2011).

GDELT was designed to serve as a "passive crowdsourcing" platform (Leetaru, 2011), providing a dashboard that keeps a pulse on global events and news reporting by assessing linkages between communication processes and societal-scale physical behavior. GDELT's main database can be divided into three interrelated sub-datasets: the Global Knowledge Graph (GKG), Events, and Special Collections. We now provide an overview of each dataset and illustrate how their components may be harnessed to address communication relevant questions.

The Global Knowledge Graph

GDELT's Global Knowledge Graph (GKG) serves as the main database for storing metadata extracted from print, broadcast, and news web portals. Once a news report appears online, GDELT scrapes and computationally analyzes its content by relying on various image recognition and natural language processing techniques. On each obtained textual news report, GDELT employs more than 40 content-analytic dictionaries to extract over 2,230 identified emotions, topics, themes, and named entities. These dictionary-based content analysis pipelines draw on a preselected list of keywords (i.e., the dictionary) that are assumed to reflect a construct of interest. The respective goal of these dictionaries is to measure "the rate at which keywords appear in a text to classify a document into categories or to measure the extent to which

documents belong to particular categories" (Grimmer & Stewart, 2013, p. 274). In the GCAM string of the GKG, each dictionary and its subcategories is assigned a variable. In turn, these variables indicate how frequently certain keywords pertaining to certain categories appeared in a news document.

Among the currently implemented dictionaries are the popular Linguistic Inquiry and Wordcount (LIWC; Pennebaker, Francis, & Booth, 2001), the Moral Foundations Dictionary (MFD, Graham, Haidt, & Nosek, 2009), Hogenraad's (2003) Motive Dictionary, and numerous sentiment analysis libraries such as SentiWordNet 3.0 (Baccianella, Esuli, & Sebastiani, 2010) and VADER (Hutto & Gilbert, 2014). Yet, it must be emphasized that the validity of these GCAM measures is bound to the validity of the respective dictionaries. Thus, users drawing on a certain dictionary that is implemented in the GCAM string should critically evaluate the validity of this dictionary based on previous research. We recommend that users consider the following factors in their evaluation: First, were the words for the creation of the dictionary drawn from news texts or other, structurally different media formats? Second, was the dictionary validated against reliable human codings and across media formats? Third, does the dictionary extract enough signal to allow statistically powerful analyses? In a similar vein, communication researchers may be interested in drawing on GDELT's theme dictionaries (Leetaru, 2013) to identify news articles that discuss certain topics. Currently, GDELT monitors the presence of 284 topics by applying various pattern matching techniques. For example, news articles attributed the PROTEST theme contain mentions of protesting, demonstrating, rioting, striking, activists, agitators, and so forth. Furthermore, GDELT has started to assign topics certain meta-tags to allow a higher level of aggregation of similar topic areas. For instance, HEALTH_PANDEMIC, HEALTH_SEXTRANSDISEASE, and HEALTH_VACCINATION all

refer to discussion of human health, albeit from various perspectives. A researcher interested in analyzing general trends in the discussion of health topics can thus rely on these meta-tags as means to aggregate semantically similar topics.

After a news report has fully been processed and analyzed, its extracted content metadata is stored in a distinct GKG record, along with the article's associated news outlet, time of publication, and Uniform Resource Locator (URL). Notably, GDELT does not provide the full text of the article due to copyright restrictions. This limitation is common across many other large scale news databases (e.g., MediaCloud). Computational communication scholars can often overcome this limitation by creating web scrapers that follow the URL associated with each individual GKG record (see R. Weber et al., 2018, study 6).

The GKG offers numerous applications for conducting theory-driven and/or data-driven communication research. For communication scholars interested in framing, the dictionary-based GCAM metrics open several avenues to assess news framing. According to Entman's (1993) conceptualization, news frames can be identified via "the presence or absence of certain keywords" (p. 52). Hence, the dictionary-based analyses pipelines included in the GCAM allow to detect keywords that signal the presence of a certain frame. For example, researchers have examined news framing by relying on LIWC and the MFD; dictionaries that are implemented in the GCAM. Fulgoni and colleagues (2016), for instance, have relied on the MFD to measure partisan differences in moral news framing, whereas Sagi and Dehghani (2014) utilized the MFD to assess generic moral frames surrounding the 9/11 terror attacks and issue specific moral news frames surrounding abortion debates. Further, such fine-grained analyses on the moral framing of a discussed topic may reveal meaningful partisan differences in news coverage that drive present polarization debates (Boxell, Gentzkow, & Shapiro, 2017). Lastly, the framing of certain topics

(e.g., terror, climate change, democracy, science, etc.) can be assessed by relating the GCAM content metrics to topics identified via GDELT's theme dictionaries.

Beyond these wordcount based dictionary approaches, latent framing and agenda-setting aspects may also be uncovered through the use of network-analytical approaches on GCAM data. For example, constructing networks that link co-occurring entities and topics across news articles and outlets has proven useful for advancing intermedia agenda-setting research (Guo & Vargo, 2017; Vargo & Guo, 2017). Moreover, the integration of identified fake news websites within the GKG stream (Guo & Vargo, 2018; Vargo, Guo, & Amazeen, 2018) may prove useful to better understand and counter misinformation as it is constructed and spread.

GKG metadata may also have merit in manual content-analytic paradigms. As R. Weber and colleagues (2018) have demonstrated, the GKG can be utilized to preselect URLs of news articles based on certain criteria of interest, such as date of publication, source, topic, or average article length. After scraping the respective text of these URLs, R. Weber and colleagues (2018) employed a crowd-sourced content analysis procedure to extract the fine-grained, latent moral information represented in these articles, leading to the development of improved moral foundations dictionaries (Hopp, Cornell, Fisher, Huskey, & R. Weber, 2018). In a similar vein, Fisher, Cornell, Hopp, and R. Weber (2018) have relied on the GKG to preselect and scrape articles that were subsequently subjected to a semantic network analysis, uncovering latent partisan and entity framing in large news corpora.

Lastly, the GKG presents a rare opportunity to bridge macro (i.e., large-scale, behavioral data) and micro (i.e., experimental/self-report data) paradigms. Scholars interested in the perception of climate change frames (Nisbet, 2009), for example, may obtain articles discussing the issue from various sources and framing perspectives. Subsequently, participants in an

experiment could be exposed to these preselected articles. In turn, variance in cognitive processing or behavioral intentions may be explained via GKG's automatically derived content features, such as keywords pertaining to fear appeals. Likewise, the GKG presents a fruitful resource to further advance the science of news sharing (Milkman & Berger, 2014; Scholz et al., 2017). While GDELT does currently not provide the share counts of a given GKG record, freely-available application programming interfaces (APIs) such as *sharedcount.com* can be utilized to obtain social media sharing counts of a certain URL. Combined with the GKG's article metadata, sharing counts can be linked to topics and keywords of interest on a cross-national scale.

Events

The main purpose of GDELT's EVENT database is to store geopolitical events that are recorded in news articles. To computationally extract a mentioned event from a news report, GDELT relies on the CAMEO codebook (Gerner, et al., 2002). CAMEO consists of manually predefined and extensively validated verb phrases to detect up to twenty different types of event occurrences concerned with international and domestic conflict, ranging from diplomatic event categories (e.g., "APPEAL" , or "AGREE") to conflict categories (e.g., "REDUCE RELATIONS", or "USE CONVENTIONAL FORCE"). Each event type is then attributed one of the twenty *EventRootCodes* that specify its type, as well as a more detailed *EventCode* that specify the event type in a more fine-grained fashion. For example, *EventRootCode* 14 reflects any PROTEST event, whereas *EventRootCode* 1422 reflects a hunger strike for policy change. Finally, all event types are ultimately organized under four primary classifications: Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict. This *Quadclass* variable supports the analysis of event types at the highest level of aggregation and is less prone

to GDELT's automated coding errors in the underlying events data (Wang et al., 2016), due to the higher levels of aggregation in mapping event types to symbols.

Beyond assessing the event type, CAMEO also extracts the two main actors that characterize an event. Accordingly, CAMEO assesses actions where actor 1 performs a respective action (the event type) on actor 2. Hence, additional information about these actors is provided—for example, whether the actors belong to a certain country, ethnic/religious group, or are part of a governmental, military, or educational institution. Furthermore, GDELT provides the geographical location of the actors and the event by recording the location information closest to the point in the event description that contains the actual statement of action. Geographic information is provided in various resolutions, spanning country, administrative region, city, and latitude-longitude coordinates.

Moreover, each event type is assigned a value on the Goldstein index (Goldstein, 1992), which assesses the conflictive-cooperative nature of the event on a scale from -10 (very conflictive) to +10 (very cooperative). In addition, during the first 15 minutes after an event has been recorded for the first time, GDELT provides the total number of mentions of this event across all source documents (*NumMentions*), the total number of information sources (e.g., news outlets) containing one or more mentions of this event (*NumSources*), the total number of source documents containing one or more mentions of this event (*NumArticles*), and the average sentiment of all documents containing one or more mentions of this event (*AvgTone*) on a scale from -100 (extremely negative) to +100 (extremely positive). The average sentiment of an event reflects the average "tone" of all documents containing one or more mentions of this event during the 15 minute update in which it was first seen.

While GDELT's event metadata–to the best of our knowledge–has not yet been explored with a CCR lens, we argue that it provides various interesting avenues to examine communication-relevant questions. First, we see merit in advancing traditional news value theory (see Eilders, 2006 for an overview) from novel methodological perspectives utilizing GDELT's event data. For example, drawing on classical typologies of news factors (e.g., geographical distance, conflict, human interest, positivity and negativity, etc.), one could examine the "shareworthiness" (Trilling, Tolochko, & Burscher, 2017) of news records mentioning event types that vary in the degree to which they pertain to these news factors. Furthermore, GDELT's *NumSources*, *NumArticles*, and *AvgTone* variables may prove useful in linking an event's news factor to the density and significance of coverage it receives within its first minutes of occurrence.

Second, communication scholars interested in *issue-specific* news framing, i.e., frames surrounding the depiction of certain event types, may find a worthwhile research avenue in combining GDELT's automatically assessed event types with underlying automatically extracted content features of articles that discuss these events. For example, one could investigate how (densely) a terror attack in France is subsequently covered by European, Middle-Eastern, East-Asian, or North American news outlets. Third, combining the news frames provided by the GKG with event records made available through the EVENT table may yield novel insights into the temporal dynamics in which events drive certain types of news frames and vice versa, which types of news frames become motivationally relevant and contribute to the development of novel events, culminating in dynamic-transactional event-news careers (Früh & Schönbach, 1982).

Special Collections

GDELT applies its automated GKG content-analytical pipeline on a selection of "special collections" spanning American television, academic literature discussing the Middle East and Africa, human rights reports, and historical American books. While the primary focus of this article is on GDELT's GKG and EVENT data, we envision that these special collections provide additional, unexplored research outlets for computational communication scholarship. For instance, the dataset on human rights reports may provide a worthwhile archive for communication researchers interested in the framing and spatio-temporal development of human rights issues (e.g., Barel, Hopp, & Weber, 2018). Likewise, the American television repository may prove useful to uncover latent differences in news coverage between print and broadcasting. Furthermore, broadcasting coverage may be linked to geopolitical events and complemented with audience ratings to explore the impact of outstanding events on television audience ratings (R. Weber, 1993).

The GDELT Interface for Communication Research (iCoRe)

In light of GDELT's massive datasets, GDELT's data has been integrated into Google BigQuery, a web service designed to store and provide access to large-scale datasets through standard SQL queries. Accordingly, the majority of studies accessing GDELT have utilized Google BigQuery (e.g., Qiao et al., 2017) or developed independent scripts that download and parse the raw GDELT data in comma-separated value (CSV) format to address a specific research question (e.g., Guo & Vargo, 2017; R. Weber et al., 2018). While Google BigQuery provides unprecedented querying speed, it is a fee-based service that can quickly become expensive with increasingly data-heavy operations. Likewise, relying on independently

developed scripts necessitates knowledge and details about their execution and the specific purpose, preprocessing, and analysis steps undertaken to obtain the data from GDELT. In addition, GDELT has started to provide their own APIs for data analysis. Yet, these APIs are unwieldy, largely exploratory in nature, and restricted to a limited number of returned data points. In addition, instead of being tailored to specific research questions, they serve as a "proof of principle" demonstrating that the data can be accessed. Hence, none of the above approaches appear helpful to (a) spur collaboration among experienced and novice computational communication researchers (b) standardize querying, preprocessing, and analysis pipelines that are increasingly becoming common practice in other data-heavy fields and (c) fulfill guidelines for conducting reproducible, open-science communication scholarship (e.g., scalability, open source, adaptability, and easy-to-use interfaces; Trilling & Jonkman, 2018).

With these issues in mind, we set out to develop the GDELT interface for Communication Research (iCoRe; http://icore.medianeuroscience.org/), a freely-available, web-based API that mitigates the aforementioned shortcomings of GDELT by providing the following features: First, iCoRe provides fast, scalable, and open-source access to an unlimited number of GKG and EVENT records contained in GDELT. In contrast, GDELT's own API restricts the number of rows allowed to be indexed at any given time, limiting the scale of computational analyses. Second, iCoRe allows a user to filter data queries to only include content-analytic measures of interest, parsing GDELT's complex GCAM string and returning these values in a preprocessed, human-readable format. Third, iCoRe allows users to select news articles from specific, carefully pre-selected news sources. This enables fast comparisons across relevant news sources, and more efficient query protocols. In addition, this pre-selection substantially improves data quality by excluding non-news websites (such as blogs or classified

advertisements), which are also monitored by GDELT, but usually provide uninteresting information for communication scholars. Fourth, iCoRe is actively under ongoing development by communication researchers. The platform will very soon contain query protocols that allow a combined retrieval of GKG and EVENT records for direct insights into the framing of events as well as for examining news-event careers. Finally, iCoRe is unique in that its technical backend, its integrated analyses, and usage guidelines are well documented, facilitating collaboration and future extension of the platform.

Technical Backend Overview

Given the large amount of unstructured data contained within GDELT, traditional relational database management systems (RDBMSs; e.g., PostgreSQL) quickly reach their computational limits when it comes to data ingest and scalability across spatiotemporal datasets (Fox et al., 2013). Hence, researchers and corporate institutions are increasingly turning to distributed databases in which data is stored across a cluster consisting of multiple virtual machines (for an overview, see Moniruzzaman & Hossain, 2013). The data is usually distributed and partitioned according to shared properties (e.g., time, location, etc.), resulting in a faster execution of query protocols. These distributed databases necessitate larger storage capacities as data is being duplicated in the process, but they simultaneously ensure higher throughput and availability as the chances of having a single point of failure (SPF; e.g., downtime due to a single machine failure) are mitigated.

Accordingly, we built iCoRe on a distributed database (Apache Cassandra) across a virtual cluster currently consisting of ten nodes. iCoRe contains multiple tables to store the GDELT data according to different partition keys to speed up queries that span, for example, certain locations or time periods. Queries are driven by the DataStax Cassandra connector, with a

lightweight front end written in Java and hosted in Apache Tomcat. Cassandra queries are enhanced by Apache Spark support, which allows for in-memory distributed queries, such as filters and joins. Each query consists of an AJAX HTTP request, where the endpoint specifies the data source and the URL query string specifies the filters or constraints on the data. Users are able to acquire data in CSV or JavaScript Object Notation (JSON) formats depending on need. Moreover, iCoRe is delivered by a lightweight HTML5 user-interface. This interface offers documentation on how to query, interpret, and analyze data and provides an intuitive, easy-to-use query interface. Both the interface and the complete API specification are available at http://icore.medianeuroscience.org/. Interested readers can gain further insights into the technical backend of iCoRe via the dedicated, open-source Github repository[1].

Data Ingest and Preprocessing

Instead of querying all of the thousand sources monitored by GDELT, iCoRe is currently drawing on a specifically constructed and extendable whitelist of 111 international, major news outlets.[2] This whitelist serves as a first attempt to increase the quality of data obtained via iCoRe. By pre-selecting major news sources and excluding rather insignificant sources for large-scale communication research (i.e., sources with a small audience), iCoRe maintains a focus on news reporting, thereby excluding non-news websites (such as blogs or classified advertisements) that are also monitored by GDELT. In addition, for each source, iCoRe's whitelist provides its country of origin, whether it is state-run or not, and its political orientation (conservative, centrist, liberal) - information that GDELT does not provide. This information was gathered with the help of research assistants who collected information on a news outlet through the source's website and online encyclopedias. iCoRe in turn allows the selection of articles from any of the

1 https://github.com/medianeuroscience/icore
2 See source_whitelist in the Open Science Framework repository of this project: https://osf.io/24n6a/

111 whitelisted sources, along with the stored metadata of a given news outlet. Yet, we emphasize that this whitelist serves as a first data quality filter and will continuously be extended based on communication researchers' needs. For example, iCoRe's current whitelist only considers news outlets in the English language or foreign-language outlets that feature an English online version. Yet, in the near future this whitelist will be extended to include non-English news sources using GDELT's various translation capabilities.

Applying the aforementioned filters, iCoRe utilizes a Python script that automatically retrieves and parses GDELT's GKG and EVENT data at 30-minute intervals. Since the GKG GCAM string contains over 2,230 variables with varying length from document to document, we adopted various automated and manual pattern matching strategies to parse and preprocess this large and complex string. Currently, among other variables, iCoRe provides the GCAM output of all categories included in LIWC (Pennebaker et al., 2001), the MFD (Graham et al., 2009), Hogenraad's Motive Dictionary (2003), and GDELT's theme dictionaries. Importantly, while GDELT's EVENTS table dates back to 1979, the GKG was introduced in February 2015. Hence, iCoRe currently includes only events and GKG records that were mentioned/published after January 1st, 2015.

Computational Communication Research with iCoRe: Three Case Studies

We present three distinct, theory-driven use cases that highlight how iCoRe can be harnessed to address communication research questions from various theoretical and data-driven perspectives. In the first case study, we draw on a classical framing paradigm to demonstrate how iCoRe can be utilized to conduct a macro-level analysis of how U.S. based news sources differ in their framing of climate change and how the obtained article metadata can be harnessed

to explore the large-scale, behavioral outcomes of climate change framing. In the second study, we demonstrate iCoRe's potential utility within news value and agenda-setting theory, exploring how countries differ in their news coverage patterns following the death of journalist Jamal Khashoggi, an event that attracted global media attention in October and November of 2018. The third case study taps into the largely unexplored dynamic-transactional nature of news-event careers by highlighting how densities of news coverage following certain events ebb and flow in the United States.[3]

First Case Study: Framing Climate Change

In recent years, public and scientific interest in framing messages that relate to climate change has increased (Feinberg & Willer, 2013; 2015; Markowitz & Shariff, 2012; Nisbet, 2009). Scholars from communication and social psychology have started to examine how messages can be tailored to maximize the persuasive appeal of environmental messages among certain audiences. One branch of this scholarship investigates how individuals' moral sensibilities play a central role in shaping the effectiveness of environmental messages (Markowitz & Shariff, 2012). For instance, when focusing on the political orientation of audiences, liberals tend to place greater value on *individualizing* moral values that are concerned with care and fairness, whereas conservatives are more likely to emphasize *binding* moral values that emphasize loyalty, authority, and purity (Graham et al., 2009).

In light of these differences, Feinberg and Willer (2013; 2015) demonstrated that liberals are more receptive to environmental messages that conform more strongly to individualizing moral values (e.g., "an animal species will go distinct if global temperature levels keep rising"). In contrast, conservatives were more receptive to messages that emphasize the binding moral

3 All of the code, queries, data, and analyses for the following case studies can be viewed and executed interactively on the Open Science Framework repository for this project. https://osf.io/24n6a/

values over individualizing moral values (e.g., “protecting the soil of one’s nation from damage”). In addition, communication research has demonstrated that such morality subcultures not only exist among members of certain political camps, but also among producers (Bowman, Lewis, & Tamborini, 2014) and respective audiences (Mastro, Enriquez, & Bowman, 2012) of media content.

With these findings in mind, we use iCoRe to provide a brief example of how liberal versus conservative newspaper sources (a) differ in their climate change framing, specifically with regard to moral attitudes and (b) how these differences become motivationally relevant when predicting share counts of climate change articles among liberal and conservative audiences. To address these questions, we used iCoRe to pull articles and their URLs from two liberal-leaning news sources (*The Huffington Post* and *The New York Times*) and two conservative-leaning news sources (*Breitbart* and *Fox News*).[4] Further, we only considered articles that were published in 2017 and that contained the “ENV_CLIMATECHANGE” topic as proxy whether an article discusses climate change. This resulted in a total of 3,043 articles (see Figure 1, for a comparison of article counts across sources).

INSERT FIGURE 1 HERE

To explore differences in climate change framing between sources, we used iCoRe to pull the content-analytic output from the MFD and LIWC stored in the GCAM string of the GKG. The MFD contains preselected wordlists that aim to reflect the five moral foundations as identified by Moral Foundations Theory (MFT, Graham et al., 2009) such as notions of care and fairness. Each moral foundation in the MFD is further split into “virtue” and “vice” categories,

4 To replicate these queries, all query endpoints are available at https://osf.io/47b5m/

with respective word lists that aim to capture whether a foundation has been adhered to or violated. The content-analytic output of the MFD reflects word counts showing how often a given concept (e.g., “care” in the MFD) appeared in form of a single word in a document. Likewise, the LIWC dictionary contains wordlists pertaining to a larger, more general array of concepts (see http://liwc.wpengine.com/compare-dictionaries/), such as perceptual (e.g., seeing, hearing, and feeling), cognitive (e.g., insight, cause, certainty), and biological (e.g., body, health/ disease, sexuality) processes. We also included the total article length of each news document in our query and divided the MFD and LIWC wordcounts by the total number of words per article to control for the influence of variations in article length across sources. Figure 2 provides an interpretation of the returned CSV data.

INSERT FIGURE 2 HERE

As a first analysis, we contrasted LIWC’s *Biology* (e.g., words pertaining to health, illness, and the human body) and *Personal Concerns* (e.g., words pertaining to money, religion, or death) categories across all four sources (see Figure 3). While all sources appear to frame climate change more strongly in biological terms, liberal sources show slightly higher means than conservative sources in the *Biology* category. Likewise, conservative sources stress the *Personal Interest* slightly more strongly than liberal sources, suggesting that conservative news sources tend to frame climate change in regard to its relevance to the well-being of individuals and families whereas liberal outlets tend to focus more on the effects of climate change on humans and non-human organisms on a broader scale.

INSERT FIGURE 3 HERE

Next, to address differences in moral framing, we contrasted the ten moral foundation categories contained within the MFD across sources (see Figure 4). Interestingly, despite the hypothesized differences in moral message framing for liberal versus conservative audiences, all sources appear to adopt a similar framing strategy when discussing climate change: Words pertaining to loyalty, followed by authority, are most prevalent across all news sources independent of their political leaning. In addition, there seems to be no clear pattern between individualizing and binding moral foundations when comparing liberal and conservative sources as all included news sources score higher on the binding foundations than the individualizing foundations. This finding is unexpected, since results by Feinberg and Willer (2013; 2015) suggest that newspaper outlets would be more successful in communicating issues of climate change towards their audiences when considering the underlying moral cognitive structures of message receivers.

INSERT FIGURE 4 HERE

Finally, to test whether news documents of liberal (conservative) outlets that emphasize the individualizing (binding) moral foundations are of greater behavioral relevance to their respective audiences, we entered the URLs of the obtained GKG records into the *SharedCount* (https://www.sharedcount.com/) API to obtain the number of times a given URL has been shared on Facebook. After obtaining the share counts for each article, we grouped the data according to the liberal and conservative news outlets. We excluded articles that were shared less than one time or more than 1000 times and randomly selected 500 articles from each of the two (i.e., liberal and conservative) news source groups. Next, we estimated two negative binomial

regressions to predict the number of times each article was shared within audiences of liberal and conservative news outlets (see Table 1).

INSERT TABLE 1 HERE

When comparing the model fit between liberal and conservative sources, the conservative news source model indicates a slightly better fit, suggesting that moral values are of stronger motivational relevance among audiences of conservative outlets than liberal news outlets. This partially supports earlier research suggesting that conservatives tend to place greater emphasis on all moral foundations compared to liberals (Graham et al., 2009), which would explain the greater motivational relevance of moral values for news article sharing among conservative audiences. Furthermore, there are interesting commonalities and differences among the direction of coefficients when predicting share counts: Whereas violations of the purity and authority foundation seem to decrease share counts, adherences to the purity and authority foundation appear to increase share counts across political camps. Likewise, climate change articles that highlight unfairness and human harm appear to be positively related to share counts. Yet, considering the rather small sample size of 500 articles per group, only the *Subversion* category was a statistically significant predictor for share counts among liberal audiences, and the *Loyalty* and *Degradation* categories the only significant predictors among conservative outlets. Hence, we remain cautious with the interpretation of the aforementioned findings and invite communication researchers to utilize iCoRe for obtaining a broader range of articles, spanning multiple years and a greater, more heterogeneous selection of news outlets.

Drawing a short conclusion, the aim of the first case study was to demonstrate iCoRe's utility for communication researchers interested in exploring macro-level news framing

differences with regard to a certain topic of public interest. By relying on GCAM's content-analytic outputs such as LIWC and the MFD, we showed how differences in message framing across news sources can be crystallized. Furthermore, we demonstrated how data obtained via iCoRe can be complemented with additional behavioral data in the form of news article sharing to explore the macro-level effects of message framing among audiences. Finally, we encourage fellow computational communication researchers to scrape the article text of URLs obtained through iCoRe. The obtained articles, along with their computationally derived content metrics, may then be used in an experimental paradigm to better understand what micro-level cognitive processes give rise to large-scale behavioral outcomes in the form of news article sharing or commenting behaviors (see Scholz et al., 2017). We envision that a combination of such micro-macro analyses through iCoRe opens numerous fruitful future research directions.

Second Case Study: The Death of Jamal Khashoggi

In our second case study, we demonstrate the utility of iCoRe for communication scientists interested in trans- and international media coverage following certain outstanding events (see Wessler & Brüggeman, 2012). For example, according to news value theory (Galtung & Ruge, 1965), certain characteristics of an event (i.e., news factors) such as proximity, conflict, human interest, or positivity and negativity determine the value of an event to become *news* in a given country. In determining a recent event that received international attention, we chose to focus on the death of Jamal Khashoggi, a Saudi-Arabian journalist and U.S. green card holder who died on October 2nd, 2018 in Turkey.

To approach this case study, we first illustrate how the average tone (i.e., sentiment) of news coverage of six countries (Canada, China, Germany, India, United Kingdom, and the United States) mentioning Saudi Arabia shifted over time leading up to and following the death

of Khashoggi. We were curious whether the physical proximity of Saudi Arabia to various countries shapes the tone of country-level news coverage leading up to and following the death of Khashoggi. Using iCoRe, we pulled all articles, along with their publication date and average tone that were published in 2018 and contained ‘SA’ (i.e., FIPS country code for Saudi Arabia) in their named entities. We decided to compare the news coverage of the following countries: Canada, Germany, United States, India, United Kingdom, and China. This resulted in a total of 36,113 articles being selected. As can be seen in Figure 5, the average tone of news coverage mentioning Saudi Arabia across all countries started to decline in July 2018 and witnessed a significant drop towards the end of October 2018, the month in which Khashoggi died. However, our analysis does not suggest a direct association between the proximity of a reporting country and the tone of news coverage. While German news sources appear to adopt the most negative tone through the obtained time period, similarly distant countries like India or China adopt a more positive tone, whereas Canada appears to produce more negative coverage.

INSERT FIGURE 5 HERE

To further distill how countries covered the killing of Khashoggi, we adopt a mediated associations paradigm. Following Arendt and Karadas (2017), we conceive of mediated associations as the repeated pairing of an object (e.g., Khashoggi) with specific attributes (e.g., topics and themes) and other mentioned objects (e.g., entities, organizations, etc.). We operationalize these associations as the co-occurrence of Khashoggi within news articles along with certain themes and entities as identified by GDELT. For example, if the same article mentions Khashoggi, Donald Trump, and a topic concerned with assassination, then a connection (i.e., an edge) is incremented between these objects (i.e., nodes). Furthermore, while previous

research utilizing such mediated associations to study media stereotyping of certain issues and entities has remained restricted to media outlets within *single* countries, iCoRe enables the comparison of mediated associations of media outlets *across* countries.

To explore these mediated associations across countries, we utilized iCoRe and selected articles that were published between September 1st 2018 and November 20th 2018 and mentioned the entity “jamal khashoggi”. Furthermore, to allow for a subsequent country-level comparison, we focused on articles that were published by news sources situated in Canada, the United States, and the United Kingdom. Lastly, we selected themes and named entities as provided by GDELT. This query resulted in a total of 2,260 articles. After querying these articles, we first grouped articles together based on their country of publication. Next, for each country, we constructed a co-occurrence network that highlights how often Khashoggi was mentioned along with certain themes and other named entities (see Figure 6).

INSERT FIGURE 6 HERE

When comparing the resulting networks, it is salient that all three countries most often refer to Khashoggi along with the topic of *armed conflict*, followed by the topic of *alliance*. When examining associated entities, the *Washington Post* shares the same edge weight across all three countries, highlighting that news coverage in each country emphasized Khashoggi’s position as a columnist for the *Washington Post*. Each co-occurrence network also contains King Salman, Tayyip Erdogan, Donald Trump, and Mike Pompeo, highlighting the roles of these individuals in the developing narrative. To further study these differences, future studies could test the statistical significance or equivalence of differences in mediated associations between countries. For instance, the networks could be reduced to nodes that occur in each country to

allow for a comparison of their edge weight across countries. A simple analysis of variance may be computed in which nodes and associated edge weights are entered as factors and countries as group levels to detect statistically significant differences in coverage about Khashoggi.

Third Case Study: Event-News Careers in the United States

In the third case study, we set out to demonstrate iCoRe's utility for examining the dynamic nature of the relationship between real-world events and the reporting of these events in news media. Early research by Früh and Schönbach (1982) has recognized that there exist dynamic transactions between real-world events and news coverage that accompany these events, culminating in reciprocal *event-news careers*. Certain events (e.g., a nation-wide protest) will likely lead to intensified news coverage, which in turn may evoke heightened attention among news audiences, who subsequently seek additional information about the event, driving sustained coverage of the issue until either audiences' interest in the issue attenuates or other, more newsworthy events occur (R. Weber, 1993). Yet, except for a few recent attempts (e.g., Hopp, Fisher, & Weber, 2019), it has remained largely untested to what extent these dynamic associations indeed exist, and if they do exist, whether they are more prevalent for certain topics and events that may differ in motivational relevance. The empirical assessment of these dynamics from a communication perspective has remained restricted, largely due to the limited access to large-scale archival datasets of real-world event prevalence and the diffusion of events into news coverage.

iCoRe circumvents this data access problem as it provides insights into the spatio-temporal trajectories of real-words events and news coverage in regards to specific topics (and other content dimensions covered by iCoRe not shown in this case study). To provide a basic example of such an analysis, we apply classical time series analysis approaches (ARIMA and

Temporal Causal Modeling; Box, Jenkins, & Reinsel, 1994) in combination with time series transfer functions (Montgomery & Weatherby, 1980) to examine how the occurrence of events concerned with immigration policies in the United States are related to news coverage about immigration and vice versa.

Using iCoRe, we pulled articles that were (a) published by a U.S. news source between January 1st, 2017 and November 20th, 2018, and (b) contained the theme "IMMIGRATION." This query resulted in a total of 87,372 articles. Next, we pulled events from iCoRe that (a) occurred between January 1st 2017 and November 20th, 2018, (b) happened on U.S. soil, and (c) contained the *EventRootCode* 14, which reflects any mention of a protest event (see the CAMEO codebook for an overview of event types and associated codes)[5]. To ensure that we are capturing news coverage about protests that are concerned with immigration, we only included social movements that were identified within the above news articles discussing immigration. This resulted in a total of 902 retrieved protests. Figure 7 illustrates the z-standardized counts of news articles about protest events and news articles discussing immigration for each week in 2017 and 2018.

INSERT FIGURE 7 HERE

As a first visual validation for GDELT's events and news data, it becomes salient that spikes in protest events targeting immigration policies correspond to major social movements that were happening at that respective time point (see Table 2). To capture a few other mentionable events as "interventions" in transfer functions for demonstration purposes we dummy coded three of President Trump's top retweeted tweets with an immigration background

5 data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

(*Change Immigration Law*, 6/18/17; *Abolish DACA*, 9/7/17; *Migrant Caravan*, 10/18/18), and three extraordinary government events in which immigration played an important role (*Government Shutdown*, 1/21/18; *Senate Immigration Debate*, 2/12/18; *Midterm Elections*, 11/6/18). For simplicity in this example case study, all interventions were considered as a one-time pulse (see Montgomery & Weatherby, 1980). As can easily be spotted in Figure 7, both time series follow a slightly negative trend, but using first temporal differences resulted in two stationary time series.

INSERT TABLE 2 HERE

As a first, simple analysis of the dynamic relationship between immigration protest events and news articles discussing immigration, we looked at the cross-correlation function of the stationary time series at 1 to 7 lags (i.e. immigration discussions lagging protest events) and 1 to 7 leads (i.e. protest events lagging immigration discussion). Specifically, we inspected whether the occurrence of protest events is more likely to precede the discussion of immigration, or vice versa. This analysis, for this specific topic area, revealed no particular (direct) dynamic relationship between protest events and immigration discussions. While both variables were clearly cross-correlated at a lag/lead of zero ($r = .37$, $p < .05$, $n = 105$), all cross-correlations at higher order lags and leads for properly stationarized time series were insignificant and close to zero. However, simple cross-correlations at different lags and leads cannot fully represent the potentially more complex dynamics between two time series, and they do not take into account the dynamic influence of non-probabilistic (i.e. without measurement error) interventions in form of pulse transfer functions.

Hence, as a next step, we identified appropriate ARIMA models to accurately account for the data generating process in the protest event and immigration discussion time series, and then used both the pre-defined interventions and one of the time series as predictor of the other at various lags and leads. Due to the sharp increase in variables and parameters for such a temporal causal modeling approach it is advisable to consider more than just 105 weeks or data points. Thus, we conducted a simplified version of this analysis using daily (instead of weekly) data pulled from iCoRe (n = 688 days; otherwise the same query commands as above).

We found that protest events can be more accurately predicted by previous protest events (F (5, 642) = 4.81, p < .001) and previous immigration discussions (F(5, 642) = 5.63, p < 001) than immigration discussions by previous immigration discussions (F(5, 642)=20.68, p < .001) and protest events (F(5, 642)=1.57, p = .17). More specifically, given the temporal dependencies in this dataset and using a Granger causality perspective, the results suggest that our selected non-probabilistic event interventions (e.g. President Trump's top retweeted tweets and the midterm elections) drove the immigration discussions to some extent which in turn were the cause of subsequent protest events.

In summary, our example analysis demonstrated that iCoRe can be used to study the complex dynamics between events, news coverage, and other relevant interventions. More detailed analyses should cover a wider range of topics, a longer time frame, specific hypotheses about the nature of interventions and their influence (not just one-time pulses), and more sophisticated modeling of the *interdependence* between real-world events and their diffusion into news coverage. In addition, future analyses should also consider audiences' general media use behaviors (including audience behavior on social media).

Conclusion and Future of iCoRe

The goal of this article was to introduce the GDELT interface for communication research (iCoRe) and to showcase its value for communication scientists to answer theory-driven questions. Part of this effort is to provide a workable solution for problems facing communication researchers attempting to use large-scale open-source datasets like GDELT. iCoRe is a well-documented, thoughtfully pre-processed, and easy-to-use interface that allows communication researchers to easily access one of the most promising extant datasets for communication research. iCoRe is built on a robust, scalable architecture and uses non-relational database technology in conjunction with pre-programmed partition keys. This ensures that the system is reliable, interpretable, and quick to use for researchers of a variety of skill levels. In addition, iCoRe is open-source and under continuous development,[6] ensuring that new features, user interfaces, and analysis pipelines will become available within the API.

To demonstrate the utility of iCoRe for addressing questions of interest to communication scholars, this manuscript presented three example case studies. In the first case study, the iCoRe interface was used to access the GDELT Global Knowledge Graph and pulled 3,403 GDELT records. Subsequently, we demonstrated how the obtained records can provide insights into the framing of climate change using moral language in liberal and conservative news sources. In the second case study presented here, the iCoRe interface was used to pull 36,113 articles that mentioned Jamal Khashoggi, a *Washington Post* journalist who died on October 2nd, 2018 in Turkey and whose death attracted global media attention. This analysis investigated the tone of international news articles mentioning Saudi Arabia before and after Khashoggi's death as well as the mediated associations between Khashoggi and various entities and themes. The final case

6 https://github.com/medianeuroscience/icore

study presented here highlighted iCoRe's capability to combine GDELT's Global Knowledge Graph and event database in order to probe the dynamic transactional relationship between U.S. immigration protests and news articles that preceded and followed these events. We demonstrated that protest events can be more accurately predicted by previous protest events and previous immigration discussions than immigration discussions by previous immigration discussions and protest events.

Limitations

While iCoRe circumvents many pitfalls concerned with large-scale datasets (e.g., accessibility, transparency, and scalability), various limitations remain that need to be addressed. First, iCoRe is inevitably constrained by the validity and reliability of GDELT's automated content-analytic measurement systems. Therefore, erroneous data points that exist within GDELT are transferred via iCoRe. Based on our experience, discrepancies may occur between GDELT's reported article length and publication date and the actual length and date of the story when manually scraping and examining the news document. Although initial concerns by the International Studies Association (ISA) have limited the publication of papers drawing on historical GDELT data, the ISA has recently revoked their decision to "unsubmit" manuscripts that include GDELT data (see https://www.isanet.org/Publications/ISQ/Posts/ID/321/GDELT) As of now, the ISA "will treat articles using GDELT like any other submissions. Such papers will be subject to any and all rules related to replication data as enacted by the ISA."

Furthermore, Wang and colleagues (2016) have demonstrated that GDELT's event detection algorithms may at times over-report the densities of certain events at given time points. Communication scholars concerned with these discrepancies may engage in a two-step data acquisition process: First, iCoRe can be utilized to quickly obtain a large selection of news

articles that fulfill certain inclusion criteria (e.g., discussion of a certain topic). In a second step, manual coders can assess the accuracy of these content-analytic metrics by comparing them to the real-world online news article. For example, similar procedures have been applied for constructing the Global Terrorism Database (GTD; LaFree, & Dugan, 2007). Yet, we argue that these discrepancies become largely negligible when mining large quantities of GDELT's data. As our case-studies and other research drawing on GDELT has demonstrated (e.g., Qiao et al., 2017; Vargo & Guo, 2017; Vargo, et al., 2018), GDELT does provide meaningful, valid insights into larger societal trends and news coverage patterns. Second, communication researchers using iCoRe to obtain computationally derived content-analytic measures must be conscious of the opportunities and shortcomings that underlie these methods (Grimmer & Stewart, 2013). Lastly, due to copyright restrictions, iCoRe does not provide the actual text of a news document. Scholars wishing to obtain the full text of an article hence need to rely on a scraper that utilizes the URL of an article to download its text.

Future Directions

We envision that the herein provided case studies can serve as a helpful "jump start" for scholars in communication that are interested in conducting computational analyses, but are unsure where to begin. Likewise, we look forward to seeing how advanced computational communication researchers utilize GDELT's data provided through iCoRe and extend the herein discussed case studies to novel contexts. In addition, we argue that iCoRe will serve as an important outlet to complement large-scale, content-analytical studies with lab-controlled experiments to better understand how groups and individuals respond to certain content characteristics that, for instance, drive polarization, message virality (Brady et al., 2017; Scholz

et al., 2017), or protest behaviors (Leetaru, 2011; Mooijman, Hoover, Lin, Ji, & Dehghani, 2018).

Furthermore, we emphasize that iCoRe's capabilities can be vastly extended by collaborative efforts among communication scholars using and further developing iCoRe. For instance, we imagine that our current source whitelist of 111 sources can be extended with additional sources to monitor a broader selection of the world's news media. In addition, this whitelist may be extended with news sources that have been identified as "fake news" and are monitored by GDELT (Guo & Vargo, 2018). We hence invite our fellow communication scholars to submit data to iCoRe to enable and complement certain analyses. We are currently in touch with various institutions and research groups whose datasets may serve as valuable, additional asset to iCoRe, such as country levels of press freedom (see www.freedomhouse.org), densities and metadata of geo-located terror events (see https://www.start.umd.edu/gtd/), and regional levels of historical pathogen stress (Murray & Schaller, 2010) and moral sensibilities (Graham et al., 2011).

In conclusion, we hope that the development and application of iCoRe will "push the envelope" of computational communication research by making GDELT, a promising, vast repository of online news and events, more accessible and relevant to many of our fellow communication scientists.

Acknowledgements

We would like to thank Andreas Boschke and Jeff Oakes from Aristotle Cloud Federation for their assistance with virtual machine image development and cloud training, made

possible by National Science Foundation grant ACI-1541215. We extend our thanks to J. Michael Mangus, who contributed to an earlier version of the database backend.

References

Arendt, F., & Karadas, N. (2017). Content analysis of mediated associations: An automated text-analytic approach. *Communication Methods and Measures, 11*(2), 105–120. doi:10.1080/19312458.2016.1276894

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Lrec, 10*, 2200–2204. Retrieved from: https://esuli.it/publications/LREC2010.pdf

Bowman, N., Lewis, R. J., & Tamborini, R. (2014). The morality of May 2, 2011: A content analysis of US headlines regarding the death of Osama bin Laden. *Mass Communication and Society, 17*(5), 639–664. doi:10.1080/15205436.2013.822518

Barel, A., Hopp, F. R., & Weber, R. (2018). *The moral framing of human rights reports: An exploratory data analysis of the human rights global knowledge graph*. Poster presented at the 2018 Summer Undergraduate Research Experience Project Showcase, University of California, Santa Barbara, USA.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences, 114*(40), 10612–10617. doi:10.1073/pnas.1706588114

Box, G. E. P., Jenkins, G. M. , & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control* (3rd edition). Englewood Cliffs, N.J.: Prentice Hall.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. doi:10.1073/pnas.1618923114

Eilders, C. (2006). News factors and news decisions. Theoretical and methodological advances in Germany. *Communications*, *31*(1), 5–24. doi:10.1515/COMMUN.2006.002

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x

Feinberg, M., & Willer, R. (2013). The moral roots of environmental attitudes. *Psychological Science*, *24*(1), 56–62. doi:10.1177/0956797612449177

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, *41*(12), 1665–1681. doi:10.1177/0146167215607842

Fisher, J., Cornell, D., Hopp, F. R., Weber, R. (2018, May). *But how are they talked about?": A novel measure of entity framing in online news*. Paper presented at the annual meeting of the International Communication Association (ICA), Prague, Czech Republic, Prague, CZ.

Fox, A., Eichelberger, C., Hughes, J., & Lyon, S. (2013). Spatio-temporal indexing in non-relational distributed databases. *IEEE International Conference on Big Data* (pp. 291–299). doi:10.1109/BigData.2013.6691586

Früh, W., & Schönbach, K. (1982). The dynamic-transactional approach. A new paradigm of media effects. *Publizistik*, *27*, 74–88.

Fulgoni, D., Carpenter, J., Ungar, L. H., & Preotiuc-Pietro, D. (2016). *An empirical exploration of moral foundations theory in partisan news sources*. *LREC*. Retrieved from: www.lrec-conf.org/proceedings/lrec2016/pdf/1076_Paper.pdf

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, *2*(1), 64–90. doi:10.1177/002234336500200104

Gerner, D. J., Schrodt, P. A., Yilmaz, O., & Abu-Jabr, R. (2002). Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association*, New Orleans.

Goldstein, J. S. (1992). A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, *36*(2), 369–385.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267-297. doi:10.1093/pan/mps028

Guo, L., & Vargo, C. J. (2017). Global intermedia agenda setting: A big data analysis of international news flow. *Journal of Communication*, *67*(4), 499–520. doi:10.1111/jcom.12311

Guo, L., & Vargo, C. J. (2018). “Fake news” and emerging online media ecosystem: An integrated intermedia agenda-setting analysis of the 2016 US presidential election. *Communication Research*, 1–23. doi:10.1177/0093650218777177

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. doi:10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. doi:10.1037/a0021847

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, *7*(2), 99–108. doi:10.1177/1745691611434210

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi:10.1093/pan/mps028

Hester, J. B., & Dougall, E. (2007). The efficiency of constructed week sampling for content analysis of online news. *Journalism & Mass Communication Quarterly*, *84*(4), 811–824. doi:10.1177/107769900708400410

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65. doi:10.1126/science.1200970

Hogenraad, R. (2003). The words that predict the outbreak of wars. *Empirical studies of the Arts*, *21*(1), 5–20. doi:10.2190/HJWQ-QRBX-0C2E-VJYA

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247. doi:10.1111/j.1540-5907.2009.00428.x

Hopp, F. R., Fisher, J., & Weber, R. (2019, May). *The dynamic relationship between news frames and real-world events: A hidden markov model approach*. Paper presented at the annual meeting of the International Communication Association (ICA), Washington D.C., USA.

Hopp, F. R., Cornell, D., Fisher, J., Huskey, R., & Weber, R. (2018, November). *The moral foundations dictionary for news (MFD-N): A crowd-sourced moral foundations dictionary for the automated analysis of news corpora*. Paper presented at the annual Convention of the National Communication Association, Salt Lake City, UT, USA.

Huberman, B. A. (2012). Sociology of science: Big data deserve a bigger audience. *Nature*, *482*(7385), 308. doi:10.1038/ 482308d

Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.

LaFree, G., & Dugan, L. (2007). Introducing the global terrorism database. *Terrorism and Political Violence*, *19*(2), 181–204. doi:10.1080/09546550701246817

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., ... & Jebara, T. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, *16*(9). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/3663/3040

Leetaru, K. (2013). The GDELT global knowledge graph (GKG). Available at http://gdeltproject.org/

Leetaru, K., & Schrodt, P. A, (2013). GDELT: Global data on events, location and tone, 1979–2012. Paper presented at the International Studies Association Meeting, San Francisco, CA, USA. Retrieved from http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

Lin, J. (2015). On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 33–47. doi:10.1177/0002716215569174

Markowitz, E. M., & Shariff, A. F. (2012). Climate change and moral judgement. *Nature Climate Change*, *2*(4), 243–247. doi:10.1038/nclimate1378

Mastro, D., Enriquez, M., & Bowman, N. D. (2012). Morality subcultures and media production: How Hollywood minds the morals of its audience. In Tamborini, R. (Ed.) *Media and the moral mind* (pp. 99–116). Routledge.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176–187. doi:10.1086/267990

McCombs, M. (2005). A look at agenda-setting: Past, present and future. *Journalism Studies*, *6*(4), 543–557. doi:10.1080/14616700500250438

Milkman, K. L., & Berger, J. (2014). The science of sharing and the sharing of science. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13642–13649. doi:10.1073/pnas.1317511111

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*, 389–396. doi:10.1038/s41562-018-0353-0

Moniruzzaman, A. B. M., & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*. Retrieved from: https://arxiv.org/abs/1307.0191

Montgomery, D.C., & Weatherby., G. (1980). Modeling and forecasting time series using transfer function and intervention methods, *AIIE Transactions*, 12(4), 289–307. doi:10.1080/05695558008974521

Murray, D. R., & Schaller, M. (2010). Historical prevalence of infectious diseases within 230 geopolitical regions: A tool for investigating origins of culture. *Journal of Cross-Cultural Psychology, 41*(1), 99–108. doi:10.1177/0022022109349510

Nisbet, M. C. (2009). Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development, 51*(2), 12–23. doi:10.3200/ENVT.51.2.12-23

Noelle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, *24*(2), 43–51. doi:10.1111/j.1460-2466.1974.tb00367.x

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334(*6060), 1226–1227. doi:10.1126/science.1213847

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic inquiry and word count: Liwc 2001*. Mahway: Lawrence Erlbaum Associates.

Qiao, F., Li, P., Zhang, X., Ding, Z., Cheng, J., & Wang, H. (2017). Predicting social unrest events with hidden Markov models using GDELT. *Discrete Dynamics in Nature and Society*, *2017*. doi:10.1155/2017/8180272

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, *32*(2), 132–144. doi:10.1177/0894439313506837

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, *49*(1), 103–122. doi:10.1111/j.1460-2466.1999.tb02784.x

Scholz, C., Baek, E. C., O'Donnell, M. B., Kim, H. S., Cappella, J. N., & Falk, E. B. (2017). A neural model of valuation and information virality. *Proceedings of the National Academy of Sciences*, 201615259. doi:10.1073/pnas.1615259114

Smith, E. M., Smith, J., Legg, P., & Francis, S. (2017). Predicting the occurrence of world news events using recurrent neural networks and auto-regressive moving average models. In Chao, F., Schockaert, S., & Zhang, Q. (Eds.) *Advances in Computational Intelligence Systems* (pp. 191–202). Wiesbaden: Springer.

Trilling, D., & Jonkman, J. G. (2018). Scaling up content analysis. *Communication Methods and Measures*, *12*(2–3), 158–174. doi:10.1080/19312458.2018.1447655

Trilling, D., Tolochko, P., & Burscher, B. (2017). From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & Mass Communication Quarterly*, *94*(1), 38–60. doi:10.1177/1077699016654682

Van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92. doi:10.1080/19312458.2018.1458084

Vargo, C. J., & Guo, L. (2017). Networks, big data, and intermedia agenda setting: An analysis of traditional, partisan, and emerging online us news. *Journalism & Mass Communication Quarterly*, *94*(4), 1031–1055. doi:10.1177/1077699016679976

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, *20*(5), 2028–2049. doi:10.1177/1461444817712086

Wallach, H. (2016). Computational social science: Towards a collaborative future. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (p. 307). Cambridge, UK: Cambridge University Press.

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science, 353*(6307), 1502–1503. doi:10.1126/science.aaf6758

Weber, M. S. (2018). Methods and approaches to using web archives in computational communication research, *Communication Methods and Measures, 12*(2–3), 200–215, doi:10.1080/19312458.2018.1447657

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., ... & Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures, 12*(2–3), 119–139. doi:10.1080/19312458.2018.1447656

Weber, R. (1993). The impact of outstanding events on television audience ratings of the *Tagesschau* against the background of the dynamic transactional model. Master thesis, University of the Arts, Berlin, Germany.

Wessler, H., & Brüggemann, M. (2012). *Transnational communication*. *An introduction.* Wiesbaden: Springer-Verlag.

Tables

Table 1

Negative Binomial Regressions Predicting the Number of News Article Shares on Facebook for Liberal and Conservative News Outlets

	Liberal News Outlets			Conservative News Outlets		
MFD Categories	Coefficient	0.025	0.975	Coefficient	0.025	0.975
Care	– 6.53	– 41.93	28.86	7.71	– 22.09	37.53
Harm	24.06	– 10.72	58.85	6.93	– 22.65	36.52
Fairness	– 4.45	– 52.53	43.62	20.76	– 28.52	70.04
Cheating	79.80	– 111.09	270.69	20.79	– 222.38	263.96
Loyalty	7.04	– 8.06	22.14	– 16.75*	– 32.81	– 0.69
Betrayal	– 25.66	– 82.21	30.88	8.18	– 30.48	46.85
Authority	2.87	– 28.24	33.99	4.56	– 18.52	27.65
Subversion	– 186.23**	– 318.33	– 54.14	– 0.76	– 48.71	47.18
Purity	42.41	– 6.47	91.29	26.91	– 24.57	78.40
Degradation	– 33.03	– 128.67	62.60	–127.54*	– 231.32	– 23.75
AIC		6327.56			6076.19	
BIC		– 2021.15			– 1923.45	
Log Likelihood		– 3152.8			– 3027.1	

Note. Coefficients with confidence intervals. N = 500 articles per source. * $p > .05$. ** $p > .01$.

Table 2

Identified Protest Events Linked to Immigration in the United States From January 1st 2017 – November 20th 2018

Protest Event	Date
Nation Against Trump	1/22/17
Travel Ban	1/5/17
Day without Immigrants	2/12/17
Million Hoodie March	2/26/17
March for Trump	3/5/17
May Day	5/7/2017
Protest after Trump rally in Phoenix	8/20/17
Protests against phasing out of the Deferred Action for Childhood Arrivals (DACA) program	9/10/17
Families Belong Together	7/1/18, 1/29/18

Figures

Figure 1

Frequency of Articles Discussing Climate Change Per Source in 2017

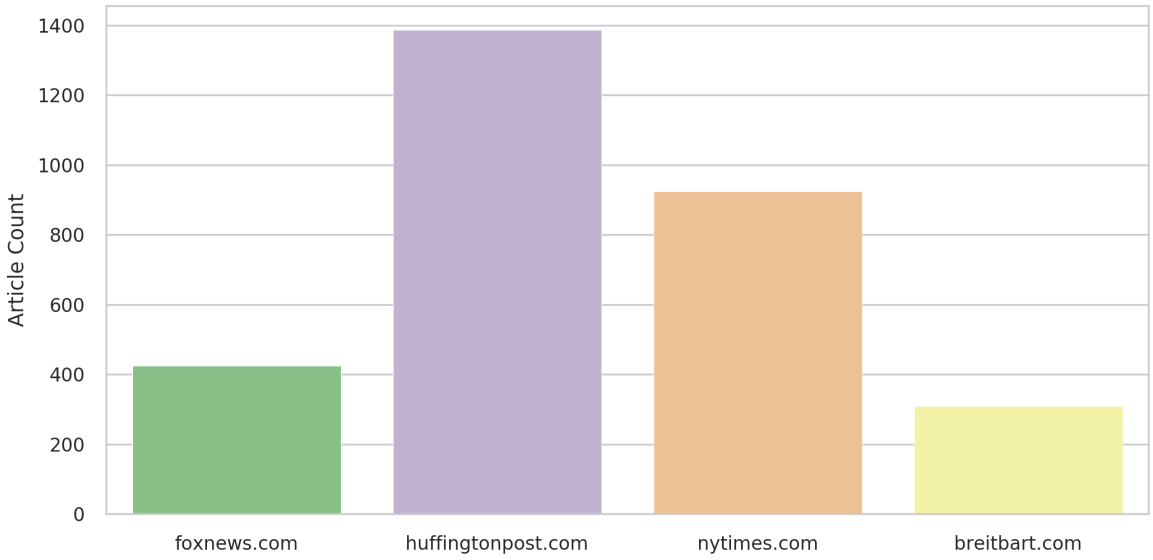


Figure 2

Excerpt and Description of Query Result for the First Case Study

id	date	url	tone	care	harm	themes	named_entities	source	wordcou	source_location
20170228003	Tue Feb 28 00:(	http://abcnews.g	0.8057296	1		[AFFECT|BAN	[AS|IR|MX|US|USCA|	go.com	1026	US
20170228133	Tue Feb 28 00:(	http://abcnews.g	0.4524887		1	[CRISISLEX_CR	[UK|US|donaldtrump|	go.com	200	US
20170228163	Tue Feb 28 00:(	http://abcnews.g	0.75757575			[CRISISLEX_C0	[CA02|citycouncil|dor	go.com	120	US
20170228140	Tue Feb 28 00:(	http://abcnews.g	-0.4234026	3	5	[AFFECT|CRIS	[RS|SP|US|USCA|USD	go.com	2193	US
20170228040	Tue Feb 28 00:(	http://abcnews.g	0.2141327	5	3	[AFFECT|ARRE	[IZ|SP|US|USKY|USM	go.com	424	US
20170228090	Tue Feb 28 00:(	http://abcnews.g	-0.7337526	3	1	[CRISISLEX_C0	[RS|US|USDC|USNH|(	go.com	899	US

Note. For each news article, from left to right: unique record identifier; publication date; URL; average tone/sentiment (–100 most negative through +100 most positive); wordcounts for identified *care* and *harm* words (other dictionary categories omitted for visualization purpose); identified themes/topics; identified entities; news source; article length; FIPS country code of news source location.

Figure 3

Comparison of Selected LIWC Categories Across Sources

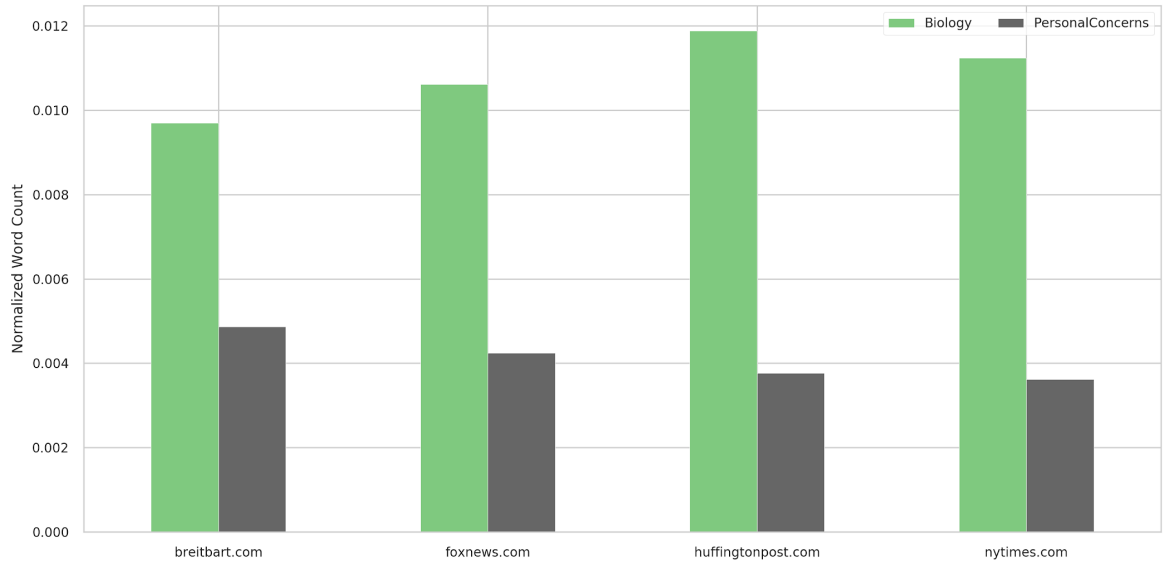


Figure 4

Comparison of MFD Categories Across Sources

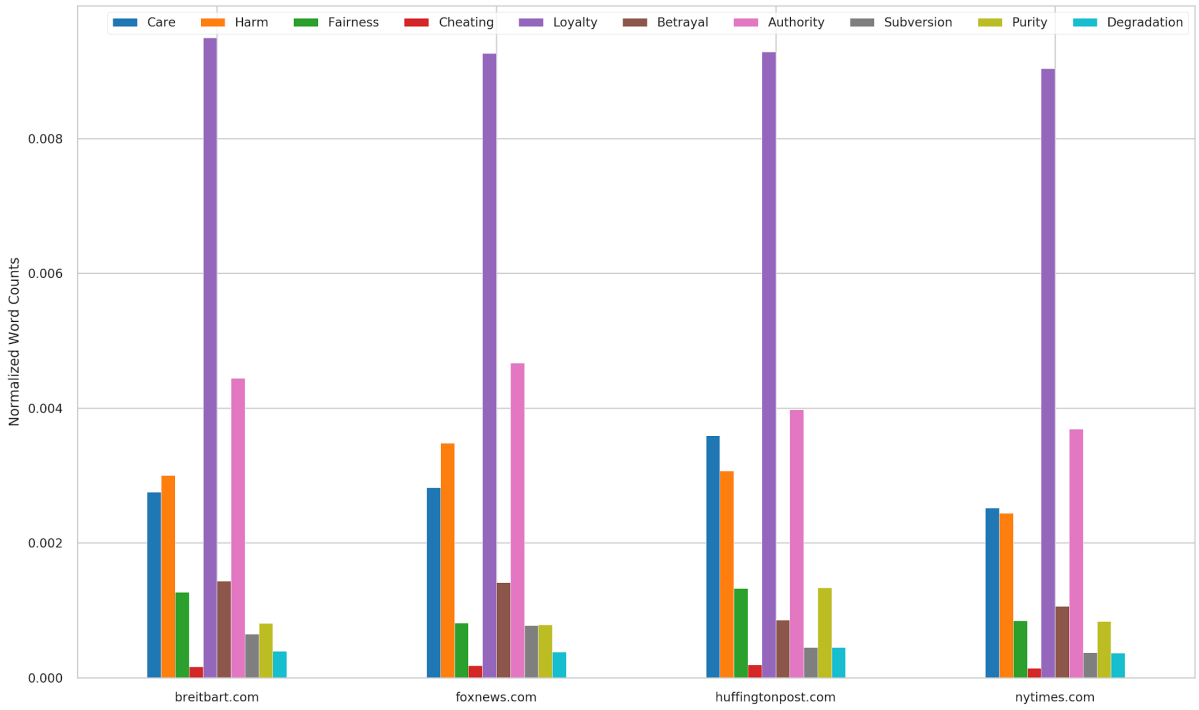
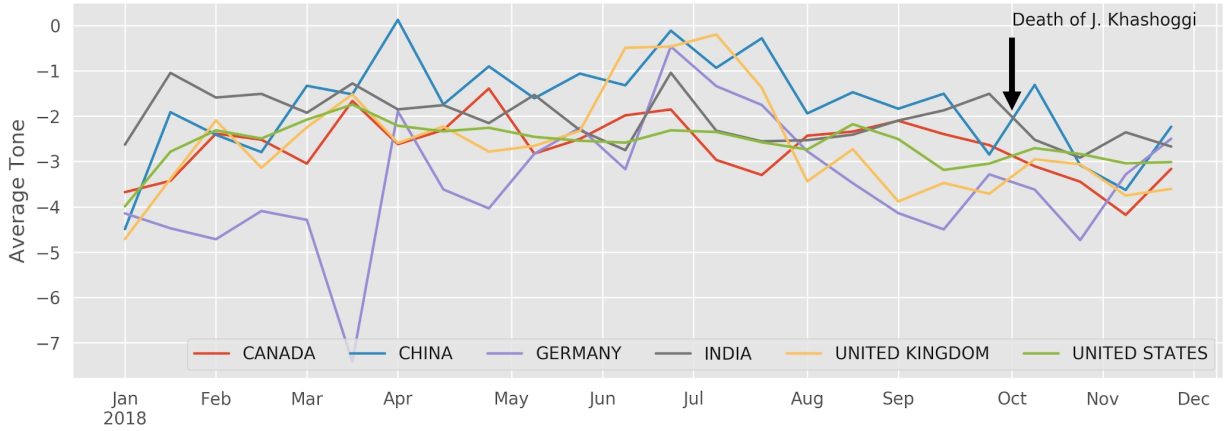


Figure 5

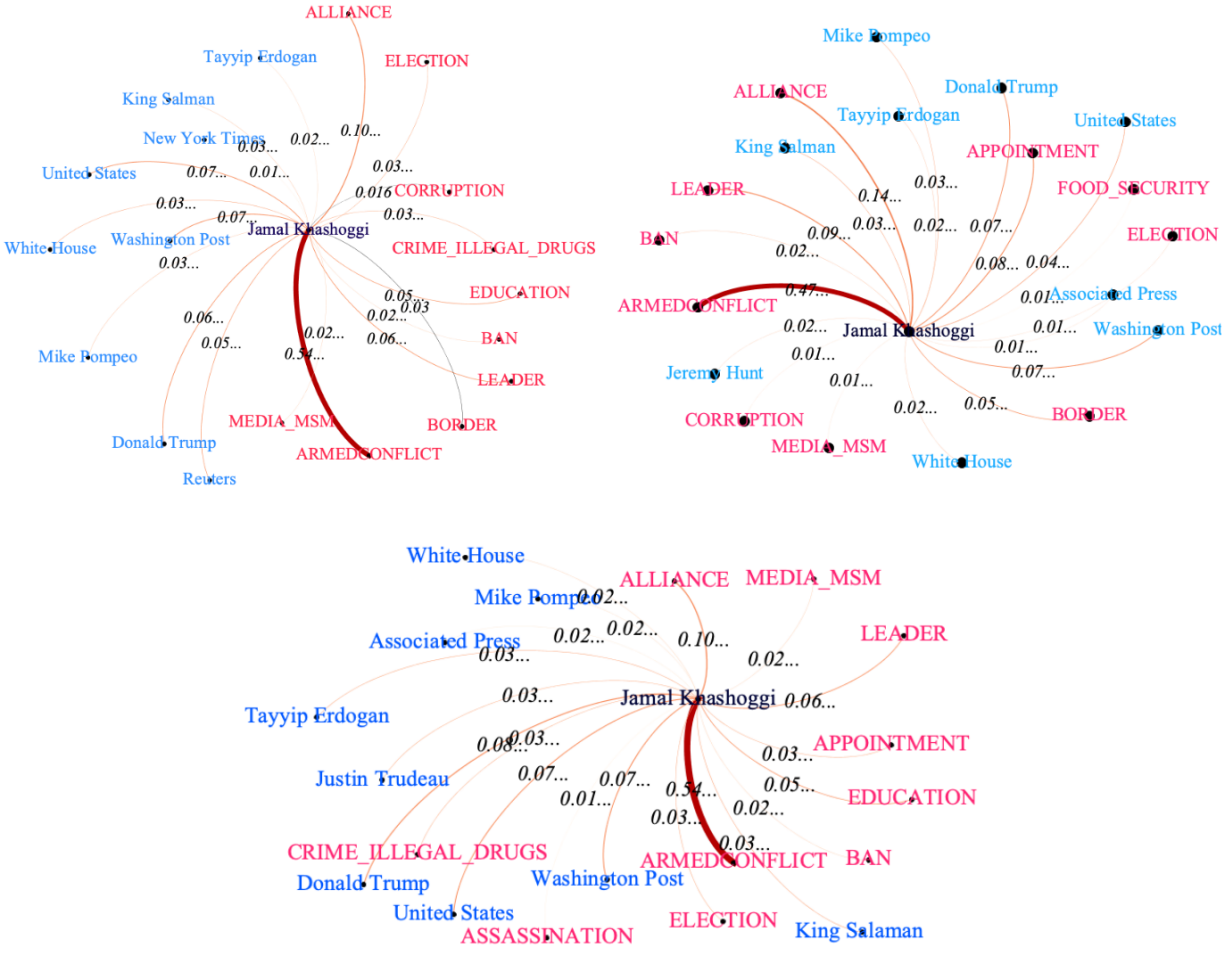
Average Tone of News Sources by Country Mentioning Saudi Arabia



Note. Average tone was computed using a two-week moving average.

Figure 6

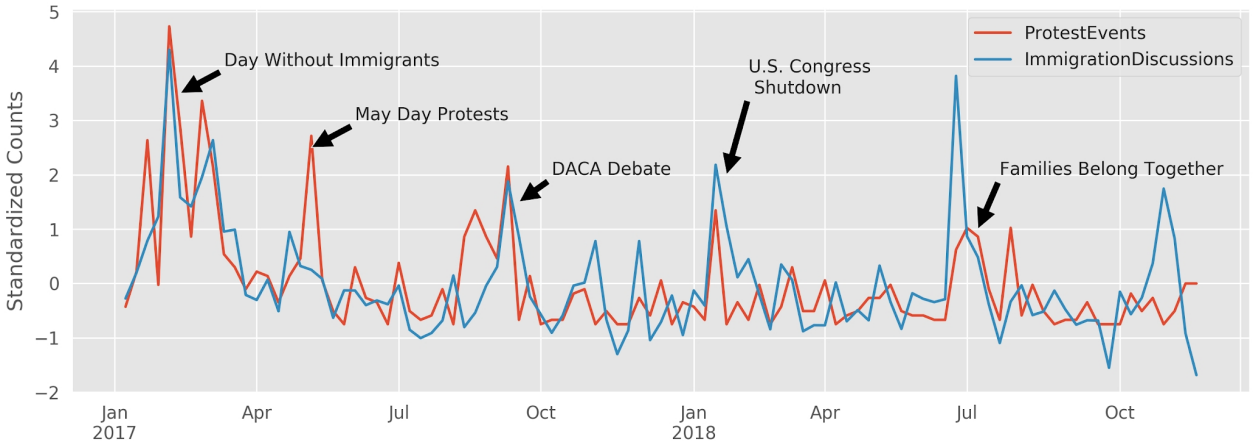
Comparison of Mediated Association Networks Between the United States, the United Kingdom, and Canada



Note. Top-left: United States, Top-right: United Kingdom, Bottom: Canada. Red nodes reflect GDELT themes, blue nodes reflect named entities. Edge weights reflect normalized co-occurrence frequencies between J. Khashoggi and the top ten themes and named entities that were mentioned with J. Khashoggi in news sources within each country. Networks were visualized with Gephi.

Figure 7

Time Series of Immigration Protest Events and News Articles Discussing Immigration in the United States



Note. Counts were computed with a one-week moving average and then z-transformed