Evaluation of Open-Source Data for Gray-zone Operations Decision-Systems

Robert Ducharme, Colby McAlexander, Brian Mills
CAE USA
Arlington, TX
robert.ducharme@caemilusa.com, colby.mcalexander@caemilusa.com, brian.mill@caemilusa.com

Jay Freeman
CAE USA
Tampa, FL
jay.freeman@caemilusa.com

ABSTRACT

Gray-zone activities—behaviors and/or actions potentially leading to, but below the threshold of armed conflict—executed across actors' instruments of national power present significant national security and global stability challenges. Successful gray-zone maneuver depends on an actor's ability to model, then implement effective strategies whilst managing the associated risks and chaos of possibly destabilizing activities. One approach to modeling the evolving nature of global competition and conflict is examining the history of international relations encoded in multiple Conflict and Mediation Event Observations (CAMEO) open-source databases. An exemplar – the Global Database of Events, Language and Tone (GDELT) project is 55TB of events and related data from public news sources accumulated for four decades. This paper examines the feasibility and suitability of this data as a means for decision-makers to explore complex, dynamic gray-zone phenomena, anticipate competing incentives, and assess consequences of choices. There are four facets to this proposed approach. First, it will be shown gray-zone news events fall on a Pareto distribution in terms of the number of mentions each gets in the media. Second, Reflective Thematic Analysis (RTA) is used to extract relevant data from GDELT to train statistical topic models for actor behaviors. Thirdly, results—including newsfeeds and thematic signatures—are generated for two actors over the first four months of 2023. Regarding data quality it will be shown that filtering events with low mention counts can be used for data conditioning, but unfiltered and filtered topics appear statistically similar so that strong filtering is not usually worth the information loss. Finally, we discuss utilizing open-source intelligence (OSINT) for potential model generation for wargaming capabilities. In this, emphasis will be placed on the usefulness of mention counts for cost-benefit-risk analysis to aid decision-making as well as the power of RTA to adapt OSINT to alternate analyst frameworks.

ABOUT THE AUTHORS

Robert Ducharme is a Senior Scientist and Data Science Technical Authority at CAE USA. He has worked in the training and simulation industry for 23 years and currently specializes in applications of AI / ML in battlespace simulations. He gained his BSc and PhD degrees in theoretical physics from the University of Essex in England.

Colby McAlexander is an Advanced Concepts Engineer at CAE USA. Retired Naval Aviator, Information Warfare Officer (25-years). Experience in joint, multi-national, and all-domain operations and planning; lead IW expert in the highest multi-classification, pedigreed multi-threat war games at U.S. Naval War College. B.S. in Computer Science, Texas A&M University, College Station, TX; MOAS, Air University, Montgomery, AL.

Brian Mills is a Modeling & Simulation Engineer at CAE USA. He leads the Cross-Domain Simulation, Training & Analysis Range (XDSTAR) engineering team. Previous employment includes Toyota, FlightSafety International, and Fives. Brian is a USAF veteran and received his B.S. in Electrical Engineering Technology, University of Cincinnati.

Jay Freeman is a Synthetic Environment Technical Fellow at CAE USA. He oversees various project, to include R&D investments in operational decision support, USSOCOM Mission Command System and Common Operating Picture, and Joint Staff J7 Environmental Development Division's development of a Joint Training Tool. Mr. Freeman previously served as the System and Software Architect for SE Core DVED, TERREX, Lockheed Martin STS (ATARS - SOFPREP) and Intergraph Services Company. Hobart College (Geneva, NY) for undergraduate and the University of Alabama in Huntsville (Huntsville, AL) for graduate studies.

Evaluation of Open-Source Data for Gray-zone Operations Decision-Systems

Robert Ducharme, Colby McAlexander, Brian Mills
CAE USA
Arlington, TX
robert.ducharme@caemilusa.com, colby.mcalexander@caemilusa.com, brian.mill@caemilusa.com

Jay Freeman
CAE USA
Tampa, FL
jay.freeman@caemilusa.com

INTRODUCTION

To achieve national objectives, state actors—hereafter, actors—may execute gray-zone campaigns such as territorial grabs and ideological subversion—under the guise of 'territorial integrity'—foreign investments, and foreign partnerships. A Center for Strategic & International Studies (CSIS) report (2018) identifies the following non-military actions actors may use in their campaigns: election meddling; economic coercion; ambiguous use of forces. In addition to these actions, actors may employ information warfare and/or military brinkmanship to achieve incremental strategic victories. A common factor is an actor "'can test the waters' with gray-zone activities to determine the relative strength of domestic and international commitment to an endeavor without resorting to the more lethal violence of war [or armed conflict]" (Kapusta, 2015). Furthermore, Mazarr (2015) makes the case regarding individual gray-zone actions; They are often an integrated part of gradual and methodic efforts "to change important aspects of the global distribution of power and influence in their favor." For instance, regimes with a 'long-game' grand strategy often exhibit subtle implementation of a 'chaos theory,' or 'deterministic chaos,' methodology to create a 'butterfly effect.' In other words, small alterations can give rise to strikingly great consequences;" "Whereby a minute localized change in a complex system can have large effects elsewhere" (Oxford Dictionary of English). This chaos coupled with an actor imposing increased cost and risk on what an opponent holds to be of value, presents calculated decision-making challenges. Adding to these challenges, gray-zone confrontation is often defined as occupying the space between peace and war; It is distinct from the concept of hybrid warfare, which includes a component of conventional war. Both gray-zone and hybrid warfare terms have been subject to criticism (Stoker and Whiteside, 2020). Specifically, these terms introduce confusion into foreign policy discussion and blur the line between peace and war. To note, it is not this paper's intent to fully define or argue the finer details of the much-discussed gray-zone, hybrid warfare, associated activities, or 'actions are in the eyes of the beholder' statements. For this paper: an actor has knowledge on how its opponent may perceive its actions, albeit lacking in full or 'perfect' understanding; Gray-zone activities are behaviors and/or actions potentially leading to, but below the threshold of armed conflict; The actor purposely acts contrary to widely accepted or ratified—on a global stage—responsible norms and behaviors.

Successful gray-zone maneuver depends on an ability to model effective strategies whilst managing the risks and chaos of possibly destabilizing, subtle activities; These maneuvers, when synchronized across an actor's instruments of national power, inject confusion and present significant national security and global stability challenges. To address the confusion issue, we acknowledge gray-zone activities injects disorder (entropy) into international relations and propose this must be tolerated to some degree to support the longer-term goal of learning to adapt. Our project's goal is to investigate and/or create frameworks and models for a future solution which feeds decision support systems to aid decision-makers' examination of gray-zone tactics in terms of strategic campaigns and effects. This paper is an interim step to advance a theory for our larger gray-zone project; Our goal is to develop an algorithm for extracting gray-zone newsfeeds from open-source intelligence (OSINT). In turn, we will evaluate this algorithm's ability to discover data facilitating the generation of gray-zone behaviors themes to construct topical models. Our view is such a system could prove useful in the early identification and characterization of gray-zone campaigns as they unfold, and thus better position decision makers to counter such campaigns before a competitor's objectives are achieved.

BACKGROUND

Gray-zone activities, disparate with normal international relations, occur often and tend to generate news topics with multiple events which have many mentions (i.e., a reference to something or someone). Thus, one means of gathering

OSINT is through publicly available news reports. Related research projects examples include: (Murphy, Boyd, Mandrick & Dannewitz, 2017); (Nadolski, & Fairbanks, 2019); (Sullivan, 2022). A significant common factor is they all use the Global Database of Events, Language and Tone (GDELT) as the news source. GDELT (Leetaru & Schrodt, 2013) is an open platform (https://www.gdeltproject.org/); It uses sophisticated natural language and data mining algorithms to draw data from print, broadcast, and web news media in over 100 languages across every country in the world. The data is comprised of:

❑ a Conflict and Mediation Event Observations (CAMEO) encoded (Schrodt, 2012) events table;
❑ a mentions table relating events to all sources which reference each event;
❑ a global knowledge graph (GKG) which characterizes each of the sources using themes.

GDELT archives date back to January 1, 1979; GDELT 2.0 format was introduced in February 2015. The events database is roughly 55-TB of events and related data from public new sources currently updated every 15 minutes. GDELT event data contains a SOURCEURL field which usually serves as a headline for the news articles that describe the events. Joining the events table to the mentions table enables the number of mentions to be calculated for events. Similarly, the events table can be joined to the GKG in support of determining a list of themes for each event. The GDELT themes are generated from a mix of other sources such as the crisis lexicon (Olteanu, Castillo, Diaz & Vieweg, 2014). This paper examines the feasibility (i.e., accomplish tasks within established limitations) and suitability (i.e., appropriate for task or intent) of GDELT data as a potential means for decision-makers who must anticipate competing incentives and assess consequences of choices as they maneuver through complex, dynamic gray-zone phenomena.

To generate gray-zone analysis results from GDELT data, we constructed the data pipeline illustrated in Figure 1. Next, we applied Structured Query Language (SQL) to extract a gray-zone data dump from GDELT using Google's 'BigQuery' tool. This raw data is conditioned to generate a more refined gray-zone dataset for further use in topic modelling and data analytics. Our mixed-method implementation of topic modelling is different from fully automated approaches performed on textual data using techniques like Latent Dirichlet Allocation (LDA), since the gray-zone topic models are constrained to use the themes GDELT has extracted from the GKG. Regardless, in common with LDA (Adams & Janowicz, 2015), the goal is to generate a statistically trained topic in support of data analytics. For our specific use-case the topics will represent gray-zone actor behaviors.

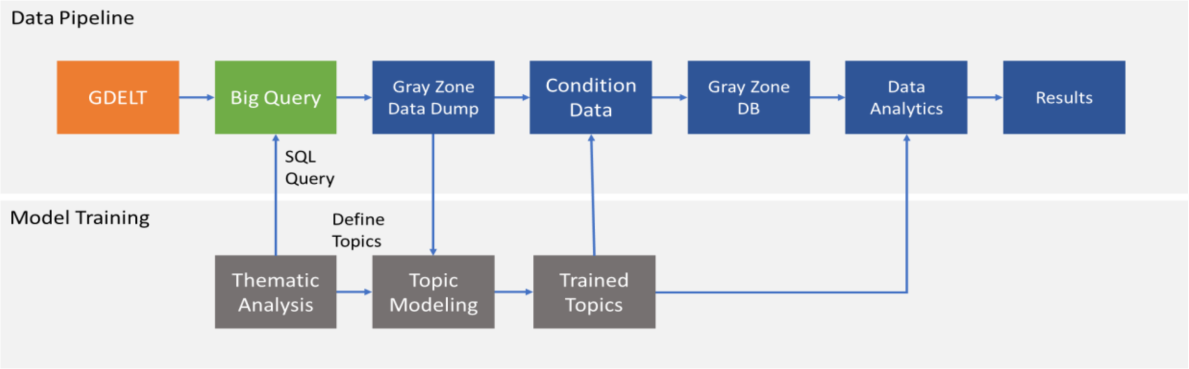


Figure 1. High-level Diagram of Gray Zone Data Pipeline and Offline Training System

In fully automated approaches to topic modeling, human analysts do not choose the themes. Thereby, it is of interest to consider Reflective Thematic Analysis (RTA) used in the qualitative analysis field as it: formalizes human analysis of textual data; emphasizes the analysis results will be a function of the themes the analyst chooses, as well as the actual data being reviewed (Braun & Clarke 2006). Often, multiple analysts conduct the same study with the objective of consensus building. The key point is the field of qualitative analysis is broadly shared between human-centric disciplines which all purposefully aim to assemble different human perspectives on human problems—significant since international relations are between humans. RTA is labor intensive and therefore not practical for application to big data. Regardless, past investigators recognized they both have a similar purpose and developed hybrid big data solutions for topic modeling which do integrate human analysts' perspectives (Gillies, Murthy, Brenton & Olaniyan, 2022). Building on this argument, we found although GDELT themes do have limitations, there are still multiple opportunities for analysts to add value, through: selection of the most relevant themes; merging of overlapping themes; compilation of related themes into a hierarchy of topic labels. Furthermore, owing to the absence of tailoring in GDELT, themes for gray-zone topic mining, analyst intervention is in fact essential to the process.

From a mathematical standpoint, a topic can be represented as a probabilistic vector (thematic signature) in a statistical themes space. The number of relevant themes gives the dimensions of the space, and the components of the vector scale to the mention counts for each theme. GDELT algorithms have performed the hard work in finding the theme references in each source. As such, all we need to complete the training of the topic is to count the mentions. In plain language, journalists vote for a theme each time they reference it in a news topic. To emphasize, this research is based on journalists' points of view and how one can leverage their newsfeeds for decision-support systems. Thus, GDELT provides the record of all the actor behaviors the journalists have voted on and the thematic signatures we shall compute for each actor by simply tallying the votes. Accordingly, this paper will present findings in these sections:

- ❑ **Data Discovery—**Explains the process of querying the GDELT events and mentions tables for a sampling of past gray-zone activities by applying keyword search to the SOURCEURL field. In post processing steps, it will reveal event mention counts lie on a Pareto distribution showing the possibility of trimming the data, to include only those events with high mention counts but this leaves the problem of setting the cutoff value.
- ❑ **Thematic Analysis—**The event database is joined to the GKG to generate themes and mention counts for past gray-zone news stories. It will be shown the results: (1) are suggestive of a human analyst workflow to correlate the themes with gray-zone actor behaviors; or other news topics (2) themes can be used to form rules that detect actor behaviors that are useful for both constructing SQL queries for GDELT, as well as topic modelling; (3) themes can serve as a more general replacement for keywords in the SOURCEURL.
- ❑ **Topic Modeling—**Describes thematic signatures and explains how they are derived from statistical topic modeling and their usage to represent gray-zone actor behaviors. Results including newsfeeds and thematic signatures are then generated for two actors over the first four months of calendar year 2023. In consideration of data quality, it will be shown it is useful to measure the degree of similarity between topics generated using different mentions cutoff values. This indicates the impact (usually small) that inclusion of the low mention rate data has on the statistical results from the model.
- ❑ **Discussion—**Presents the suitability of GDELT, alongside other OSINT sources, for decision-support. Results of both thematic analysis and topic modeling will be revisited, highlighting the approach's limitations and improvement opportunities. It is noted potential applications of OSINT include tracking gray-zone activity as it emerges and harvesting data to inform wargaming models. One avenue we will analyze, with respect to wargaming, is the feasibility of gauging the impact of gray-zone activities and events across the Diplomatic, Information, Military, Economic (DIME) instruments of national power. For instance, if employing a different RTA generates the possibility to model gray-zone actors in terms of DIME as an alternative framework.

DATA DISCOVERY

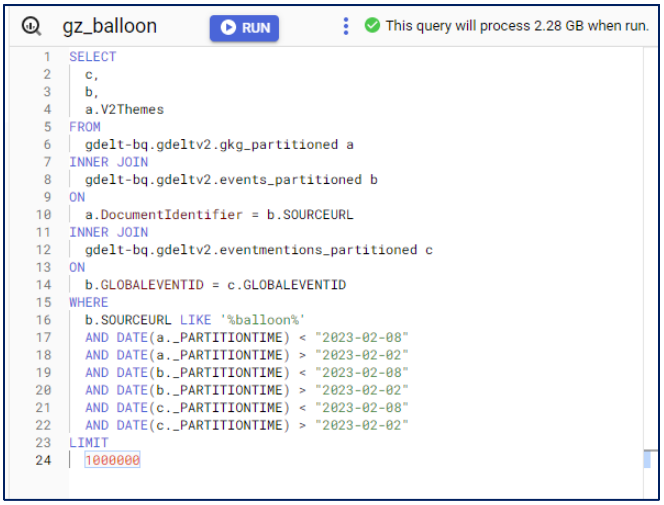
This investigation requires OSINT for two purposes: (1) provide historical data for training of statistical topic models; (2) to be the source of a streaming news feed to support data analytics. For brevity, the analysis in this section focuses on the suitability of the GDELT project to generate training data for statistical models of topics. Thereby, consideration of alternative OSINT sources to GDELT and the possibilities for data fusion will be deferred to the discussion section.

Data Interfaces

GDELT is free to use and downloadable through a web interface, which is useful for some purposes; however, the GDELT database qualifies as big data. This means if queries need to be run against significant portions of GDELT, there are challenges of scale to store the massive dataset and process it in a reasonable amount of time. Google BigQuery cloud data warehouse solves these problems since it hosts the entire GDELT project and executes Structured Query Language (SQL) requests in seconds, vice hours. All SQL processing for this paper was performed in BigQuery. To begin the data discovery phase of this work, we will analyze the Chinese spy balloon news story which started in early February 2023. This choice has no special significance compared to other gray-zone news stories; It is just helpful to have a recent, concrete example of a news story to exemplify a process we will later apply to many.

Extraction of Training Data

After creating a BigQuery account and accessing the search box on the web, GDELT projects display as a long list of tables. To generate data, the user scripts and runs a SQL query. Figure 2 shows a sample query which selects all data from the joining of the GDELT, GKG, events, and mentions tables designed to generate all records containing the word “balloon” in the SOURCEURL field—from the 2nd to 8th of February 2023. This query generated 67962 records, each containing fields from all the joined tables. The basic structure is each event has a unique GLOBALEVENTID field in the events table which may have multiple mentions in the mentions table. Note, the number of mentions for an event is indicative of the significance the world’s journalists attached to it compared to other news events. The GKG table is joined to the events table through the SOURCEURL field to access themes from the database which will be utilized in RTA. In the overall workflow, the limited function of the SQL query is to extract training data for topic modeling. One significant task is determining better content for the query’s WHERE section, since the current content is specific to one news story, whereas a general solution needs to be inclusive of broad range of activities.



This query will process 2.28 GB when run.

```sql
SELECT
  c,
  b,
  a.V2Themes
FROM
  gdelt-bq.gdeltv2.gkg_partitioned a
INNER JOIN
  gdelt-bq.gdeltv2.events_partitioned b
ON
  a.DocumentIdentifier = b.SOURCEURL
INNER JOIN
  gdelt-bq.gdeltv2.eventmentions_partitioned c
ON
  b.GLOBALEVENTID = c.GLOBALEVENTID
WHERE
  b.SOURCEURL LIKE '%balloon%'
  AND DATE(a._PARTITIONTIME) < "2023-02-08"
  AND DATE(a._PARTITIONTIME) > "2023-02-02"
  AND DATE(b._PARTITIONTIME) < "2023-02-08"
  AND DATE(b._PARTITIONTIME) > "2023-02-02"
  AND DATE(c._PARTITIONTIME) < "2023-02-08"
  AND DATE(c._PARTITIONTIME) > "2023-02-02"
LIMIT
  1000000
```

Figure 2. SQL Query—Chinese Spy Balloon News Story

Pareto Distribution

The Chinese spy balloon news story contains thousands of CAMEO events. Therefore, it is desirable to have an adjustable means of filtering out less important ones. Thus, a Python script has been developed to read the mentions data outputted from GDELT; It determines the mention count for each event and extracts only those events above a threshold number of mentions. Figure 3 shows a vast majority of events in the Chinese spy balloon news story have low mention counts, but a few have very high mention counts. This is called a Pareto type distribution. It implies much can be learned from restricting the analysis to a few high-quality events per news topic. Nevertheless, it does not give an optimum value for this cutoff such that we need to quantify the actual impact of the cutoff. Table 1 shows the top 7 CAMEO events in the Chinese spy balloon news topic by mention count. The result confirms the potential of a GDELT newsfeed to detect and convey the essentials of a breaking gray-zone news story. The following events are clearly reported on the days they happened in just 7 of 4000+ CAMEO events GDELT recorded for the entire news story.

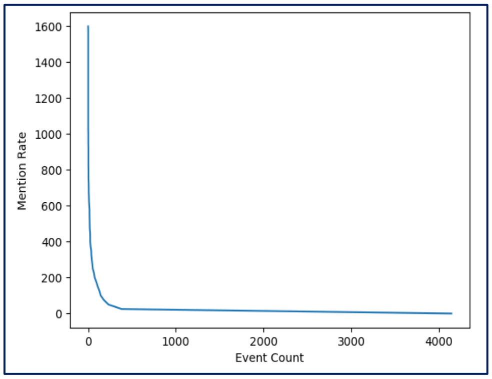


Figure 3. Mention Count vs. Event Count

Table 1. Chinese Spy Balloon News Story as Reported in Top-7 GDELT Events

Date	Time (EST)	CAMEO Code	Headline
2023-02-02	18:00:00	Host a visit	Chinese Spy Balloon Spotted Over Western US Pentagon Says
2023-02-03	03:00:00	Express intent to meet or negotiate	China Looking Report Spy Balloon United States
2023-02-03	10:15:00	Consult	Blinken Postpones His China Trip After a Chinese Balloon is Spotted Over Montana
2023-02-03	11:15:00	Consult	Blinken Postpones Beijing Trip After Chinese Balloon Enters US Airspace
2023-02-03	14:45:00	Consult	Chinese Balloon Soars Across US Blinken Scraps Beijing Trip
2023-02-03	23:45:00	Consult	China Balloon Questions Suspected Spy Sky
2023-02-04	16:15:00	Occupy territory	US Downs Chinese Balloon a Flashpoint in US China Tensions

Limitations of OSINT

A more careful comparison of Table 1 to the reported facts, which have since emerged, shows GDELT misses the first portion of story when the spy balloon was first spotted over Alaska and then floated into Canada. To note, these events were not made public at the time they occurred. Also, it is clear the timestamp on the events typically lags a few hours behind the actual event occurrence's, owing to the time delay in the recording of these events and subsequent assimilation into the GDELT project. It's noted some alternative news sources such as twitter are often faster. Another limitation of the analysis, so far, is the word 'balloon' had to be entered into the SQL query. It would not have been intuitive to do until the spy balloon story started breaking. Thus, a system which relies on keywords in the SOURCEURL field is not fully automated. For this reason, we will turn next to examining the themes GDELT stores for each news source in the GKG as a possible replacement for keywords in the detection of gray-zone activity.

THEMATIC ANALYSIS

Thematic analysis provides a means for human analysts to shape GDELT trained topic models. There are two puzzles to solve in the use of themes: (1) acquire better training data for actor behavior topics; (2) optimize thematic signatures for analysts. In this section we will focus on constructing rule conditions for identifying gray-zone actor behaviors in GDELT then discover a default set of behavior themes for the actor behavior topics has emerged.

Global Knowledge Graph

The SQL query depicted in Figure 1 extracts a set of themes for each event. The same theme can occur in multiple places in each source article. Thus, the number of mentions for a theme can be calculated to be the product of the number of mentions in the SOURCEURL for an event and the mention count for the same event. Table 2 shows the most mentioned themes for the Chinese spy balloon news story (See Table 1) from a total list of 400 themes GDELT detected in the news story. The main gray-zone themes being China and Surveillance. Thereby, based on CSIS categorization, the Chinese balloon news story is being primary reported as an information warfare story.

Table 2. Top Themes for Chinese Spy Balloon News Story

GDELT Theme	Mentions	Note
TAX_WORLDLANGUAGES_CHINESE	291459	Additional information on and the complete listing of GDELT GKG Themes can be found at: ❑ https://blog.gdeltproject.org/new-november-2021-gkg-2-0-themes-lookup/ ❑ https://www.gdeltproject.org/data.html#rawdatafiles (Both links last accessed on June 16, 2023)
TAX_ETHNICITY_CHINESE	291459	
TAX_FNCACT_OFFICIALS	170640	
TAX_FNCACT_OFFICIAL	161429	
SURVEILLANCE	137041	
TAX_FNCACT_SECRETARY	134481	
TAX_FNCACT_SPY	120987	

Better Training Data

Themes can be used to write more general SQL queries than Figure 2 illustrates. Evidentially, the Chinese spy balloon news story maps to SURVEILLANCE and TAX_FNCACT_SPY themes. Thereby, we then reverse this mapping to show these themes can retrieve the spy balloon news story. Next, we modified the SQL query in Figure 2 to replace the balloon statement with the following line of code: (a.V2Themes like 'SURVEILLANCE' OR a.V2Themes like 'TAX_FNCACT_SPY'). Both the query and Python post-processor were rerun exactly as before, with no coding changes to the post-processor. The result is the thematic query generated approximately twice the number of events and mentions as the original query and included >99% of the original content. Table 1 was reproduced perfectly; Yet it included one additional spy balloon event from a foreign language source. Thus, the surveillance themes are an improvement over the balloon keyword in two respects. First, the query is better for use in automated systems since it does not contain details specific to individual news topics—generally not known a priori. Second, the themes capture more relevant news events owing to the fact they are more general. Surveillance of a country by another is just one form of gray-zone activity. Therefore, it is desirable to define a broad range of gray-zone topics and seek out strong themes which are indicative of the activity and statistically likely to be present in news stories relating to the topic. One means of accomplishing this is to repeat the process generated the surveillance theme. Specifically, work though a list of past gray-zone news topics, generate a set of themes for each, and identify the best gray-zone candidates.

In further exploration of ‘better training data’ for models, we generated Tables 3 and 4 to provide examples of a simple taxonomy to discover news events tied to specific gray-zone behaviors using GDELT themes. We took an orthogonal approach, in the statistical meaning, to provide a wide sampling. To best highlight gray-zone news stories—and thereby news events—we selected themes and topics which are well-known current and historical events, widely accepted paradigms, and nominally symbolic of totalitarian and/or authoritarian regimes. Table 3 ‘composite themes’ and Table 4 ‘campaigns’ (i.e., actions an actor may pursue to successfully realize their strategy/priorities), with associated ‘news topics,’ were chosen to present concrete examples of news stories where its data can be used for trending analysis and model training. We recognize more than one theme and news topic can align with one or more campaigns; we are only offering a sample to present our research method vice a fully authoritative guide. Table 3 is a crosswalk of composite themes to GDELT themes for both a Table 4 cross-reference and brevity. Table 4 offers context for GDELT searches and linkages to news topic classifiers of regimes through: (1) representative campaigns and associated priorities; (2) concepts to provide insight into examples of advantages sought in an actor’s gray-zone campaign. To note, a single composite theme alone is not necessarily negative or gray-zone (e.g., foreign investment); however, the combination of additional specificity (see Table 4, Debt-Trap diplomacy) brings one closer to the target.

Table 3. Crosswalk of Composite Themes to GDELT Themes

Composite Themes	GDELT Themes
tension	CRISISLEX_C07_SAFETY or EPU_CATS_NATIONAL_SECURITY or WB_2432_FRAGILITY_CONFLICT_AND_VIOLENCE
diplomats	TAX_FNCACT_FOREIGN_MINISTER or TAX_FNCACT_AMBASSADOR or TAX_FNCACT_DIPLOMAT
criminal justice	TAX_FNCACT_POLICE OR ARREST OR TRIAL OR WB_840_JUSTICE
corruption	CORRUPTION or WB_832_ANTI_CORRUPTION or WB_2024_ANTI_CORRUPTION_AUTHORITIES
digital support	WB_678_DIGITAL_GOVERNMENT or WB_133_INFORMATION_AND_COMMUNICATION_TECHNOLOGIES
foreign invest	ECON_DEVELOPMENTORGS or ECON_FOREIGNINVEST
indebtedness	WB_1104_MACROECONOMIC_VULNERABILITY_AND_DEBT or WB_450_DEBT or ECON_DEBT
trade support	WB_2601_TRADE_LINKAGES_SPILLOVERS_AND_CONNECTIVITY or WB_772_TRADE_FACILITATION_AND_LOGISTICS or WB_866_CONNECTIVITY_AND_LAGGING_REGIONS
tech transfer	WB_1084_TECHNOLOGY_TRANSFER_AND_DIFFUSION or WB_1274_TECHNOLOGY_TRANSFER_OFFICES

Table 4. Tying GDELT Themes to News Topics and Gray-Zone Campaigns

Campaign	News Topic	Composite and/or GDELT Themes (See Table 3)
Territorial Integrity or Territorial Grab	Debt-Trap Diplomacy	Foreign invest, trade support, indebtedness and corruption
	Cyber Attacks	CYBER_ATTACK or WB_2457_CYBER_CRIME
	Disinformation	Use “SOURCEURL” field, which are the articles’ headline.
	Political Interference	CYBER_ATTACK AND ELECTION
	Maritime Incidents (includes artificial islands)	Tension, diplomats, and MARITIME_INCIDENT. Use SOURCEURL for ‘artificial island building’.
	Aviation Incidents	Tension, diplomats, and AVIATION_INCIDENT
	Concepts: 9-Dash Line (PRC); Control Lines of Communication; Territorial Waters and/or Economic Exclusion Zone Expansions; Access to Resources; Sovereign Rights and Control	
	Priorities: Regime Survival (Legitimacy; Authoritarian Governance); Territorial Integrity; National Unity; Comprehensive National Power Acquisition.	

Campaign	News Topic	Composite and/or GDELT Themes (See Table 3)
Ideological Subversion	Covert Police Abroad	Use SOURCEURL
	Cyber Monitoring	Tension, diplomats, and SURVEILLANCE
	Disinformation	Use SOURCEURL
	Human Rights Violations	Tension, diplomats, and WB_2203_HUMAN_RIGHTS
	Concepts: Sovereign Control; Rise to Power; Power Projection; International Influence; Social Development; Population Control; Narrative Control; Foreign Academic Institutes	
	Priorities: Regime Survival (No Challengers) Internal Security; National Unity	
Foreign Investments and Partnerships	Debt-Trap Diplomacy	Foreign invest and digital support and indebtedness and corruption
	Cyber Attack	CYBER_ATTACK or WB_2457_CYBER_CRIME
	Espionage	TAX_FNCACT_SPY and (ARREST or TRIAL)
	Criminal Corruption	Tension, diplomats, corruption, and crime
	Foreign Business Pressure	Tech transfer and corruption
	Concepts: International Influence; Minimize Western Influence; Control– Narrative, Market Corporate, Global System Advantageous to Regime; Power Projection; Resource Access	
	Priorities: Regime Survival (International Legitimacy); Territorial Integrity; Comprehensive National Power (CNP) Acquisition	

TOPIC MODELING

The goal is to construct statistical topic models of gray-zone actor behaviors. Thematic analysis has provided us with a good starting point since Tables 4 defines a set of twelve behaviors in the 'News Topic' column and gives a corresponding rule for extracting events, mentions and themes from the GDELT database for each of them.

Topic Creation

The concept of a topic is evident in Table 2; It takes the form of an N–dimensional vector Pi (i = 1, 2, …, N), where the themes are unit vectors, and mention counts size the components in an abstract statistical space. This simple geometric construction conjures up a clear image of each news topic in the GDELT database as existing as a unique vector in the space of all GDELT themes. One reason to do this would be to classify actor behavior topics into similar groups. Specifically, if Pi and Qi denote actor behavior topics, the statistical distance (degree of similarity) between them can be calculated using the Kullback-Liebler Divergence (KLD) formula (Figure 4). Thus, we are set up to use a trained statistical topic Q_i to recognize a second topic P_i as being either similar or dissimilar depending on the value of D_{KL} from the formula. One delightful aspect of working with GDELT data is themes have been extracted; Therefore, counting mentions for actor behaviors is all that is needed to complete the training of topics to the point of creating the probability vectors for topics. Our topic modelling algorithm referenced in Figure 1 is therefore a simple python script that applies the rule set from Table 4 to the raw data dump from GDELT to count the mentions. The script also includes some basic data post-processing logic such as the application of a mentions threshold. The additional data conditioning step in the diagram alludes to the fact that once created, the topics facilitate a more advanced data conditioning solution as further explained below.

$$D_{KL} = \sum_{i=0}^{N} P_i \ln\left(\frac{P_i}{Q_i}\right)$$

Figure 4. Kullback-Liebler Divergence.

Results

GDELT events are timestamped. Therefore, it is straightforward to classify each as having an association to one or more gray-zone behaviors and output them as part of a time ordered sequence as indicated in Table 1. This is a gray-zone newsfeed. GDELT themes contain language and ethnicity references that enable each event to be reliably further particularized to an actor. Figure 5 shows the result of counting each mention of these events per behavior per actor (China and Russia) over the first four months of calendar year 2023. It can be seen, for example, the February peak in the 'Surveillance' news topic for China is attributable to the Chinese spy balloon news story which generated massive worldwide interest over a few days. For presentation purposes, the mention counts have been grouped by month since the strong fluctuations evident in the data, evident in the graphs, are even more pronounced in a weekly reporting. Figures 6 and 7 show the normalized gray-zone thematic signatures for China and Russia respectively.

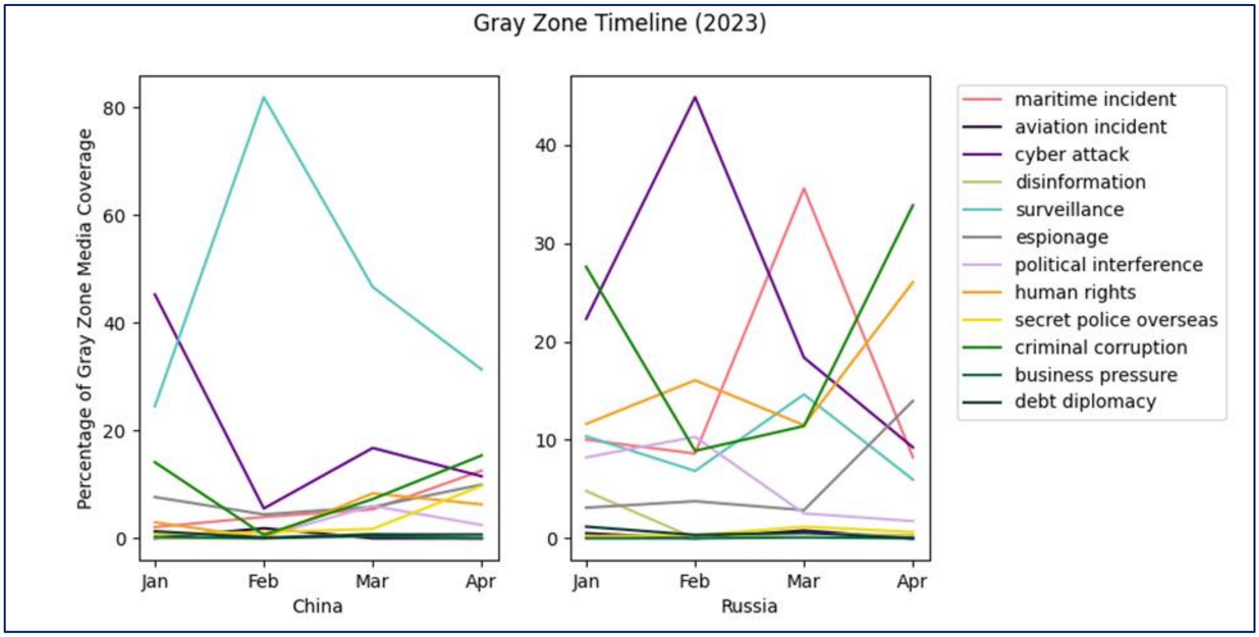


Figure 5. Gray-Zone Timeline (2023)

In comparison to Figure 5, it can be seen the mention counts per behavioral theme are summed over the entire four-month period the data was collected instead of for each individual month. It can also be seen that the thematic signatures have each been generated for two different mention count cutoff values of N = 10 and N=100 whereas Figure 5 was generated using just N = 10. Observe now, the Chinese data in Figures 5 and 6 is surveillance dominated making it very different from the Russian data in Figures 5 and 7 which shows a much more even distribution of mentions between multiple different themes. By contrast, the Chinese thematic signatures in Figure 6 look very similar for the N = 10 and N = 100 cases, also the Russian signatures in Figure 7 look quite similar. Ahead of explaining why this is important, it will be helpful to first put the result on a more formal footing using the KLD formula in Figure 3.

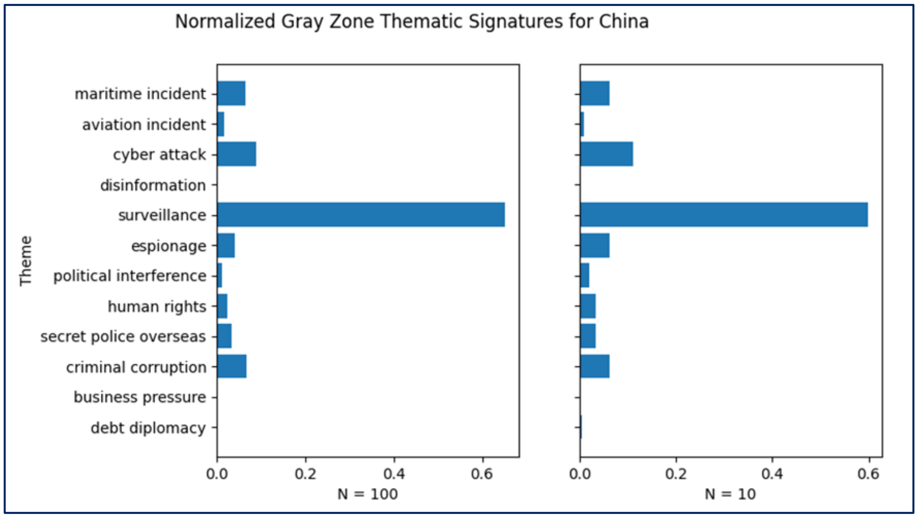


Figure 6. Normalized Gray Zone Thematic Signatures for China

Let P and Q denote the respective Chinese and Russian thematic signatures given a mention cutoff of N = 100. Table 5 shows that putting P and Q into the KLD formula gives $D_{KL}(P, Q) = 1.18788$ as the statistical distance between the Chinese and Russian topic models. Although this result is not meaningful by itself, Table 5 shows the statistical distance between N = 10 and N = 100 cases of the Chinese signature is only 0.01597 and another similar result for the Russian signatures. The point is the KLD formula is indicating the N = 10 and N = 100 cases for each actor are statistically close to each other compared to the statistical distance between the Chinese and Russian signatures like what we observed by visual inspection. Two reasons this is important. First, it shows using a low mention count cutoff like 10 is not significantly corrupting our statistical topic model. Second, the fact that we can show this using the KLD formula means we now have an automated check to include in our data pipeline to maintain N as low as possible to avoid information loss (e.g., debt diplomacy in Figure 5) whilst being alerted to anomalies of which the difference in the cyber-attack mention count between the N = 10 and N = 100 cases of Figure 6 is a minor example. Table 5 also shows gray-zone activity is time-dependent but that smaller statistical distances separate different time periods per country than different countries at least for this dataset.

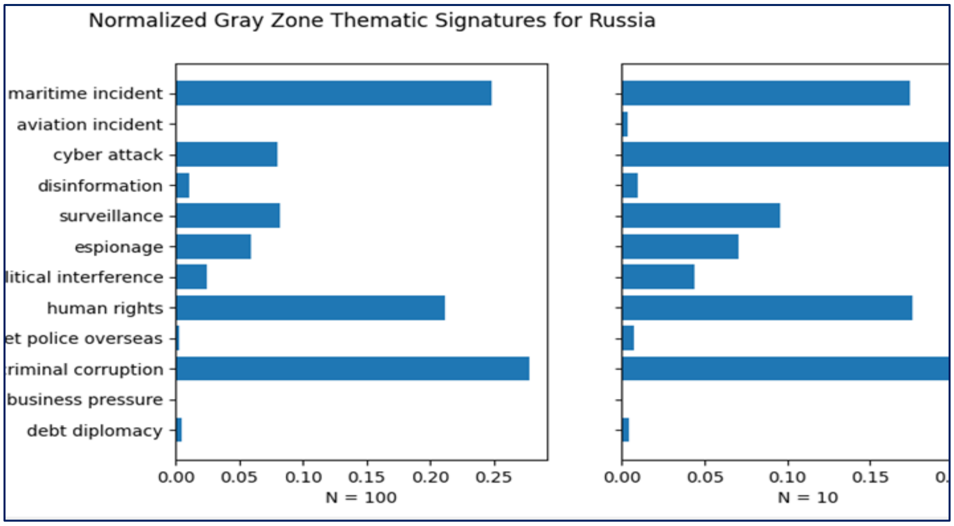


Figure 7. Normalized Gray Zone Thematic Signatures for Russia

Table 5. Statistical distances calculated using the KLD formula.

P	Q	$D_{KL}(P,Q)$
State = China, N = 100, Jan-Apr	State = Russia, N = 100, Jan-Feb	1.18788
State = China, N= 100, Jan-Apr	State = China, N= 10, Jan-Apr	0.01597
State = Russia, N = 100, Jan-Apr	State = Russia, N= 10, Jan-Apr	0.08663
State = China, N = 100, Jan-Feb	State = China, N = 100, Mar-Apr	0.427219
State = Russia, N = 100, Jan-Feb	State = Russia, N = 100, Mar-Apr	0.630314

DISCUSSION

It is taken for granted in the economic space, decisions can be costed and risks assessed. Yet, economics is just one of the four levers of power available to governments in the DIME model. One appealing feature of OSINT is it provides a metric (mentions) which spans all the elements of DIME enabling the relative importance of events happening in the world to be gauged alongside the decision-making that led to them. Therefore, mentions provide a mathematical means of assigning value to what economists call intangibles; Thus, enabling them to be quantitively integrated into cost-benefit-risk calculations as a plausible basis for decision support systems. We recognize many newsfeeds are tied to business requirements (e.g., sales) or who 'owns' the media (e.g., state-media; suppression; interference); Thereby, driving what journalists find news-worthy, or compelling, to report. These newsfeeds may only be interesting, or mildly useful, for analysts and decision-makers. What may be fascinating to decision-makers are unreported events or actions. Such events could be benign, noise, or could be purposeful hiding below the 'noise floor.' Thereby, we pose the question, 'would several low mention newsfeeds collated together reveal a hidden gray-zone campaign:' In turn, other arguments can be made: competitor conditioning of the 'battle-space;' journalistic personal interests; etc. As such, these discussions are beyond the scope of this paper and worthy of further exploration in our follow-on research. More generally, we acknowledge the further need for a much more extensive analysis of historical data than it has been to provide here for building in predictive capabilities into the system and estimating there worth.

Summary of Results

Regarding our hypothesis—mentions have potential use to inform decision-makers—we created a data pipeline to mine mentions from GDELT through Google's BigQuery engine. We discerned Google and GDELT have made the task straightforward. The key to extracting useful gray-zone data from what is a big data source (i.e, roughly 1-GB/ month) is a set of twelve rules derived from thematic analysis. We encoded both in SQL for data extraction from

GDELT, then in python for classifying the downloaded data into an approximate orthogonal set of behavior themes; In turn, we constructed statistical topic models for two gray-zone actors—China and Russia. Results which include a summary report for a gray-zone newsfeed for the first four months of calendar year 2023, alongside Chinese and Russian thematic signatures, were presented in the main-body of this paper. Our preliminary results showing statistical similarity between gray-zone activities for individual countries for different time periods is indicative of how history can inform the future but much more work of this kind is needed to put our hypothesis on a firmer footing.

Overall, our data pipeline's principal strength is its simplicity; It only applies rules to identify and classify events, then counts the mentions per behavior theme per actor. Our two main challenges were in the application of thematic analysis to develop the rules and data quality. The rules are still a work in progress, as the work is iterative. For data quality, we were concerned there is visibly both much 'noise,' as well as beneficial data in GDELT events with low mention rates. Therefore, we developed a test using the Kullback-Liebler divergence formula to flag if using lower mention rate cutoffs has a significant impact on the thematic content of topics. We found it usually does not impact.

Suitability of GDELT for Gray-zone Operations Decision Systems

We have demonstrated GDELT can be used to generate a newsfeed which reports twelve different kinds of activity for multiple gray-zone actors. This presents multiple possibilities, including data fusion coupling with other data sources (e.g., Twitter and Integrated Crisis Early Warning System (ICEWS, 2023)). Another idea, for a more holistic decision-making approach, is mentions can be used as a kind of currency for cost-benefit-risk calculations that factor in economic intangibles. A full analysis of this concept lies beyond this paper's scope. Our follow-on intent will apply TA to map GDELT themes to DIME actions vice gray-zone behaviors; This would enable a cost assessment of an actors' different courses of action and the impact it has on the rest of the world. The evolution of the twelve gray-zone rules the paper presented will still be part of this solution; It is currently the best mechanism we have identified to-date to extract a reasonably comprehensive gray-zone newsfeed from GDELT into the statistical DIME space.

CONCLUSION

We developed a sample algorithm set which successfully extracted gray-zone newsfeeds from OSINT sources (e.g., GDELT); Evaluation of the utility of the results reveals a potential feasibility to incorporate these newsfeeds data in decision-support systems. Specifically, the tools, techniques, and methods overviewed in this paper facilitated our ability explore large newsfeed datasets of complex, dynamic gray-zone phenomena (the symptoms)—i.e., this preliminary framework can act as an alternative sensor to collect, process, and exploit data to facilitate a better understanding of competing incentives and explore maneuver through the gray-zone. Our findings have revealed:

- ❑ Gray-zone newsfeeds fall on a Pareto distribution in terms of the number of mentions each receives.
- ❑ RTA can be used to extract relevant data from GDELT to train statistical topic models for actor behaviors.
- ❑ Regarding data quality, we determined filtering events with low mention counts can be used for data conditioning, but unfiltered and filtered topics appear statistically similar.
- ❑ OSINT, public newsfeed tools, have potential for generating models to aid alternate analyst frameworks.
- ❑ The suitability of the mechanisms discussed are worthy of follow-on development, and testing of a novel rapid prototype solution, with strategic and operational analysts and decision-makers is warranted.

Thereby, extracting gray-zone newsfeeds could be used to create models could support synthetic environments which provide decision-makers and their staff a space to examine a competitor's gray-zone campaigns for identification and characterization as they unfold. (Note: We purposely did not identify or discuss acceptability of this proposed initial solution, as this should be determined by those who must balance managing the 'chaos' (or 'butterfly effects') and cost-benefit-risk tradeoffs). For example, incorporating means such as GDELT, and other conditioned data sources, for model generation in simulations used in wargaming or exercises to facilitate the assessment of consequences of choices in a safe environment. Our further research in this area would include: historical comparisons; aforenoted limited-to-unreported events or actions which could be benign, noise, or purposefully hiding below the 'noise floor" (i.e. 'butterfly effects'). Moreover, to examine the question would several low mention newsfeeds collated together reveal a hidden gray-zone campaign? Finally, our follow-on rapid prototype for the concepts presented in this paper (e.g. synthetic environment and/or decision-support systems) will examine feeding the results of these newsfeeds into 4-dimensional statistical DIME space as a basis for cost-benefit-risk calculations in the gray-zone arena.

REFERENCES

Adams, B. & Janowicz, K. (2015). Thematic signatures for cleansing and enriching place-related linked data. Int. J. Geogr. Inf. Sci. 29, 556–579

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. Qual. Res. Psychol., 3, 77–101

Bunker, R. (2019). *China's Securing, Shaping, and Exploitation of Strategic Spaces: Gray-zone Response and Counter-Shi Strategies*. Small Wars Journal.

Center for Strategic International Studies. (2018, December 7). *Understanding Gray-zones Video and Survey*. COMPETING IN THE GRAY-ZONE: Countering Competition in the Space between War and Peace. https://www.csis.org/analysis/competing-gray-zone-countering-competition-space-between-war-and-peace

Gillies, M., Murthy, D., Brenton, H., & Olaniyan, R. (2022). Theme and topic: *How qualitative research and topic modeling can be brought together*. arXiv preprint arXiv:2210.00707. https://doi.org/10.48550/arXiv.2210.00707

ICEWS (2023). Integrated Crisis Early Warning System. Retrieved May 6, 2023 from https://www.lockheedmartin.com/en-us/capabilities/research-labs/advanced-technology-labs/icews.html

Kapusta, P. (2015). United States Special Operations Command Whitepaper: The Gray-zone. 09 September 2015.

Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global Data on Events, Location, and Tone, 1979–2012. In ISA Annual Convention (Vol. 2, No. 4).

Lin, Bonnie et al. (2022) *Competition in the Gray-zone: Countering China's Coercion Against U.S. Allies and Partners in the Indo-Pacific*. RAND. https://www.rand.org/pubs/research_reports/RRA594-1.html

Mazarr, M.J. (2015). Mastering the Gray-zone: Understanding A Changing Era of Conflict. US Army War College Press. Retrieved May 4, 2023 from https://press.armywarcollege.edu/monographs/428/.

Murphy, D., Boyd, D., Mandrick, B. & Dannewitz, D. (2017). Improving the Utility of Open-Source Event Data for the Design of Training Exercises. Proceedings of the 2017 MODSIM World Conference.

Nadolski, M. & Fairbanks, J. (2019). Complex systems analysis of hybrid warfare. Procedia Computer Science 153, 210-21.

Olteanu, A., Castillo, C., Diaz, F., Vieweg, S. (2014). CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In: Proc. of ICWSM, pp. 376–385

Schrodt, P. A. (2012). CAMEO Conflict and Mediation Event Observations Event and Actor Codebook Version 1.1b3. Retrieved April 16, 2023 from http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf.

Stevenson, Angus. (2015). *Oxford Dictionary of English* (*3 ed*.). Oxford Reference. https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/reference_list_electronic_sources.html Last accessed June 21, 2023.

Stoker, D. and Whiteside, C. (2020). Blurred Lines: Gray-Zone Conflict and Hybrid War—Two Failures of American Strategic Thinking," Naval War College Review: Vol. 73. No. 1. Article 4. Retrieved April 16, 2023 from: https://digital-commons.usnwc.edu/nwc-review/vol73/iss1/4.

Sullivan, J. R. (2022). Gray-zone Tactics as Catalysts for Balancing Coalitions: A Level of Analysis Approach. Master's thesis, Harvard University Division of Continuing Education. Retrieved May 8, 2023 from: https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37371532.

Troeder, E G. (2019). A Whole-of-Government Approach to Gray-zone Warfare. US Army War College Press. Retrieved May 6, 2023 from https://press.armywarcollege.edu/monographs/937/

Wasser, B. et al. (2019). *Gaming Gray-zone Tactics: Design Considerations for a Structured Strategic Game*. RAND. https://www.rand.org/pubs/research_reports/RR2915.html