Re-imaging data collection: the promise and pitfalls of using non-state actor media for large-scale event datasets

Emily K. Gade*
Shahryar Minhas†
Bree Bang Jensen‡
Kai Ping Leung§
Michael Gabbay¶
Mohammed Hafez‖
Emily Thomas**
Kayla Salehian††
Ava Sharifi‡‡
Tara Farkouh
Ramin Farrokhi
Arica Schuett
Willa Jeffers
Bernadette Bresee
Ben Lefkowitz

April 3, 2024

Abstract

Non-state actors produce their own media with a scope and frequency unmatched by conventional news. Specifically, can rebel media—which should be understood as propaganda—be used in creating *reliable* event data sets of ongoing or recent conflicts? We suggest that while rebels do lie, there are certain types of data—where rebels are on the ground, and with whom they fight and collaborate—that will be more accurately represented in rebel-produced media than in news data. Using a novel dataset coded from some 28,000 rebel social media posts during the Syrian civil war (2012-2018), we devise a novel approach to reconstructing networks from event data using latent variables approaches, and test our dataset against existing news-based data sources. Our findings reveal that rebel media can predict many of the relationships found in news media nearly as well as news media can predict itself, suggesting that rebel-generated information is a valuable source of information for understanding conflict and collaboration dynamics in conflict settings.

[*]Senior Researcher, Fors Marsh
[†]Associate Professor of Political Science, Michigan State University
[‡]Assistant Professor, University of Georgia
[§]Ph.D. student, University of Washington
[¶]Principle Physicist, Applied Physics Laboratory, University of Washington
[‖]Professor of Political Science, Naval Postgraduate School
[**]BA, Emory University
[††]BA, Emory University
[‡‡]BA, University of Washington
BA, Emory University
BA, University of Washington
Ph.D. candidate, Emory University
BA, University of Washington
BA, University of Washington
BA, Emory University

We thank Daniel Arnon, David Davis, Dan Reiter, Adam Glynn and the Emory Political Science Department Speaker Series for comments on drafts of this manuscript, as well as Emory's IR Journal Club (2021, 2022), Anna Zelenz and members of the O/R Lab @ Emory for their critiques and ideas about this broader project. This work was generously supported with a United States Air Force Academy Young Investigator Award #FA7000-18-1-0017 (PI: Emily K. Gade); research funds from the Moore/Sloan & WRF Innovation in Data Science Fellowship at the University of Washington eScience Institute; research funds from Emory University; and Office Naval Research grant N00014-16-1-2919. Shahryar Minhas acknowledges support from the NSF, Award 2017180. Authors are listed in order of contribution to this article (rather than to the broader project). Authors are listed in order of contribution to this manuscript (rather than to the broader project). Any mistakes are our own.

1 Introduction

Buoyed by technological advances and the proliferation of information platforms, non-state actors—from criminal organizations to protest movements to rebels—are increasingly producing their own "news." Their in-house media reports on their activities, political stances, and ideologies, as well as information about other actors in the same political space. This media is not held to the same standard as professional journalism, yet may offer important opportunities for researchers to evaluate questions that require information that conventional media covers less completely.

Here, we consider one such case: rebel-produced social media from ongoing conflicts. We ask: Can rebel-produced media—which should be understood as propaganda—be used to build reliable observational datasets about ongoing conflict? We first develop a theory of how rebels have incentive to lie in their own media, and detail the likely strengths and weaknesses of rebel-produced media as a source for large-scale observational data efforts. Next, using a novel dataset of rebel-produced media in the Syrian civil war (§3.2) and a new methodological approach (§4), we demonstrate that the rebel-produced media data outperforms the existing gold standard in predicting rebel relationships with other rebels (both conflictual and collaborative). Based on our findings, we suggest that this approach to data collection may generalize to other case contexts—especially protest actors.

Indeed, our findings highlight notable, substantive faults in our existing understanding of rebel behavior in multiparty civil wars. First, the vast majority of studies of civil wars exclude minor rebel groups for the sake of data collection and justify this with the contention that small actors do not shape the landscape of civil conflict. While future studies are needed to verify our findings, our data seem to suggest that the behavior of small rebel actors substantively shapes the conflict behavior of larger actors and indeed the conflict network as a whole. Second, while existing literature highlights times with groups move from fighting to collaborating or visa versa, groups in our data appear to move back and forth between fighting and collaborating and then fighting again—suggesting "Frenemies" relationships are more common that existing

literature suggests. Thus, rebel media and micro event data of conflicts more generally have the potential to open up new research questions, windows of understanding and provide modest correctives to existing assumptions of conflict studies.

The Syrian civil war is characterized by a complicated web of contentious rebel group dynamics. As Ostovar and McCants (N.d., 6) note, "There are hundreds if not thousands of small groups that have been active in the Syrian conflict, nearly countless of which have split or changed alliances, political outlook, or leadership over time. Tracking these grassroots interactions would be an endless, if not impossible task." Such micro-dynamics frequently go unreported by news outlets. Given the importance of the Syrian conflict, we take on this 'endless, impossible' task. To do so, we turn to a novel data collection approach and source—Syrian rebel social media data—in conjunction with more traditional approaches, to build the first networked, micro-event dataset of its kind: the SAGAA—Syrian Armed Groups, Alliances, and Actions—dataset (§3.2; Appendix §A–C). SAGAA is primarily drawn from 28,000 rebel social media posts depicting on-the-ground conflict engagements between 2012-2018.[1]

Theoretically, we build on existing work to develop a theory about the incentives for honesty in rebel self-reporting. Gabbay and Thirkill-Mackelprang (2010) note that the primary audience for rebel media posts in Iraq was on-the-ground followers, who have extensive information about how events unfold. As a result, rebels have significant incentives to be truthful in their media. Greene and Lucas (2020) and Whiteside (2020) likewise find rebels seek to be transparent in their reporting, especially as concerns alliance and fighting partners. We expect rebels to be forthright in their reporting about where they were, who they were with and when they were there, but we also anticipate less veracity regarding contentious topics such as civilian targeting.

Methodologically, we devise a technique for comparing our SAGAA dataset with the existing gold standard event dataset of the Syrian Civil War, the UCDP. We utilize a network-based latent variable approach to tackle incomplete information in event data. We infer non-

[1] See Appendix C for more details. We augment this with data from the Syrian Observatory for Human Rights, the Institute for the Study of War, the Carter Center, and Airwars as appropriate.

zero events and probabilistic ties between groups based on the underlying network structure (e.g., clustering and transitivity) and assess how well these predictions correspond with the true values in each dataset. This method provides a unique perspective on conflict dynamics and highlights the potential to recover missing information by solely leveraging the inherent interdependencies between actors and their connectivity patterns.

Empirically, we find that the SAGAA—while containing fewer overall events—has a much denser network of ties than UCDP,[2] and as a result, does a *better* job recovering the UCDP collaborative network that UCDP does at predicting itself. This indicates a key theoretical finding as well: that small-scale engagements and less prominent groups that may be absent from major news coverage are important in structuring the overall conflict landscape of fighting and collaboration. Indeed, higher-order dependencies in the network and a denser structure of ties are important for understanding network relationships as a whole. We also conduct a qualitative evaluation of our understanding of when rebels lie, comparing rebel media from major Syrian group Hay'at Tahrir al-Sham (HTS) to news coverage from that same group. We find that the group's coverage largely matches our expectations.

We offer four primary contributions in this article. First, we build on existing research (Aday et al. 2010, Jones and Mattiacci 2019, Loyle and Bestvater 2019, Walter 2017, Zeitzoff 2017) and highlight rebel media as a trove for conflict researchers, and evaluate the value and challenges of its use. Many militant groups leverage a range of text-based social media platforms to communicate with their public and highlight their accomplishments or actions. Second, our approach to comparing these networks builds on the latent variable literature developed in political science (Desmarais, Harden and Boehmke 2015, Fariss 2018, Nieman 2016). Specifically, we develop a novel approach for evaluating network similarity and recovering missing network data (§4) using latent factor modeling that we believe has a host of downstream implications. Third, we build hierarchical network data with event data and group-level covariate data alongside state and international actor involvement to create a clearer picture of a

[2]Indeed, more than 75% of UCDP ties involve a single dyad: Syrian state v. Syrian Insurgents.

civil conflict ecosystem. Civil conflicts—and indeed security-related affairs more generally—contain a complex web of interrelated actors whose behavior is dependent on what other actors in the network are doing, in ways that matter for how wars unfold (Dorff 2017, Dorff, Gallop and Minhas 2020, Dorff and Minhas 2017, Gabbay 2008, Gade et al. 2019, Gade, Hafez and Gabbay 2019, Larson 2016, Maoz 2012, Minhas, Hoff and Ward 2019, Parkinson 2013, Staniland 2014, Zech and Gabbay 2016).

Finally, the Syrian conflict has changed the social and political landscape of the Middle East, cost hundreds of thousands of lives, altered politics and policy across Europe, emboldened international actors that threaten the security and stability of the Middle East, stoked regional proxy wars, and provided a fertile breeding ground for violent extremism. Ongoing—and spiraling—conflicts in Yemen, Afghanistan, Pakistan, Iraq, Mali, Nigeria, and Somalia bear similarities to the Syrian conflict. As a result, understanding the micro-dynamics that facilitated such destruction has immediate real-world benefits. Our dataset provides a significantly more comprehensive treatment of its intra-war rebel dynamics than prior efforts (Gade et al. 2019, Gade, Hafez and Gabbay 2019). More granular data allows for a greater level of complexity in understanding the effects of intervention and action, allowing more effective decision-making in an inter-rebel conflict dynamic. We hope this dataset aids policymakers, scholars, and activists alike in comprehending the factors that fueled this conflict with the hopes of preventing recurrence in other similar contexts.

2 Observational Datasets and civil conflict

2.1 News data in civil wars

Observational datasets of conflict are critical to understanding how wars unfold, why actors do what they do, and the durability of peace (Kalyvas 2012). These datasets are used and or funded by conflict scholars and major international actors like the United Nations, the US Department of Defense, and the World Bank (Shaver and Lab 2022). Though major event datasets—e.g.,

ACLED and UCDP—have broken critical ground in this area, data collection in conflict zones faces significant, well-recognized challenges. Research demonstrates that data collected from news media encounters numerous issues in terms of reliability, over and under-reporting, and bias (Davenport 2009, Earl et al. 2004, Eck 2012, Weidmann 2015*b*, 2016). When considering data about civil conflict, terror attacks, or rebel behavior in particular, these biases can be exacerbated (Shaver and Lab 2022). For example, news media tends to emphasize larger, more lethal attacks (Behlendorf, Belur and Kumar 2016), often fails to distinguish between subgroups (Warner and Lizzo 2021), and is heavily influenced by local political actors and context (Baum and Zhukov 2015). These shortcomings are generally accepted as a consequence of the chaos of war, but they can also increase the emphasis on certain dimensions of conflict while blinding researchers to others (Bakke, Cunningham and Seymour 2012).

This data can provide researchers with more dynamic, thorough information on the relational shifts during conflict to better understand rebel group dynamics (Kalyvas 2003). As the access to and use of social media in contested contexts rises, scholars have turned to non-state actor-produced media as an alternative source of information about conflict (Aday et al. 2010, Jones and Mattiacci 2019, Loyle and Bestvater 2019, Zeitzoff 2017), as well as theorizing the role of this media in the behavior of armed actors (Seib 2017).

We suggest that rebel media has significant and increasing potential to complement news data in our study of violent unrest and civil war, and that these benefits may hold sway when considering media constructed by other non-state actors. For instance, rebel-rebel relationships in conflict are less likely to be reported in the news as compared to state-rebel relationships or one-sided violence, yet rebels may report on their relationships with other non-state actors (Greene and Lucas 2020). As a result, rebel-produced media may have significant advantages over traditional news media in yielding micro-event data about rebel actions on the ground, particularly regarding inter-rebel group relationships. Networked data of rebel group relationships is crucial to understanding how civil wars unfold (Dorff, Gallop and Minhas 2020, Minhas, Hoff and Ward 2019, Zech and Gabbay 2016), making rebel media a critical and com-

paratively untapped resource for understanding micro-dynamics within civil wars. Building on recent work (Gabbay 2008, Gade et al. 2019, Gade, Hafez and Gabbay 2019, Greene and Lucas 2020), which leverages rebel statements about their relationships with other rebels, state, and pro-government actors to trace rebel relationships (rather than relying on retweets or follower networks), we use rebel-produced social media data to build a networked event dataset of the Syrian conflict.

2.2 Rebel-produced media as a data source

Identifying reliable observational data—and the ways in which data sources fall short—is a challenge that pervades the study of political phenomena, from political bias (Arnon, Haschke and Park 2022, Harmon, Arnon and Park 2022) and data challenges (Fariss 2014) in human rights data, to a western-centric focus in major IR datasets (Colgan 2019), to the impacts of under-reporting on the broader study of terrorism (Drakos and Gofas 2006) or even mainstream language models designed to surmount data challenges (Gururangan et al. 2022). These challenges are exacerbated when considering violent behavior by armed actors (Baum and Zhukov 2015, Behlendorf, Belur and Kumar 2016, Shaver and Lab 2022, Warner and Lizzo 2021). Yet, while we frequently overlook the biases present in government, NGO, scholarly, or news-produced data (Baum and Zhukov 2015, Möller 2011, Öberg and Margareta 2011), the idea of leveraging rebel propaganda as a mainstay of conflict data collection may still raise concerns. Also, the proliferation of social media globally gives rise to additional challenges (and opportunities) for data veracity and collection (Olteanu et al. 2019).

Rebels and other non-state actors have been making leaflets, pamphlets, and manifestos for hundreds of years, but social media has fundamentally changed the reach and access of rebel-produced media, in ways that are important for how conflicts unfold (Aday et al. 2010, Jones and Mattiacci 2019, Loyle and Bestvater 2019, Zeitzoff 2017). Today, rebel groups everywhere use Twitter, Facebook, YouTube, and other media platforms to circulate propaganda (Azman 2014), recruit new members (Bestvater and Loyle 2023, Klausen 2015), garner pop-

ular (Blaker 2015, Cohen 2016) and financial (Berman 2012) support and, indeed, to govern (Loyle et al. 2021, Weidmann 2015*a*). Most rebel-produced media available today appears at face value to provide information unlikely to be covered in mainstream news—e.g., day-to-day operations, small practices or policy changes, the behavior of smaller groups, alliance shifts within parts of the conflict, and prisoner exchanges.

While anti-regime periodicals have a long and storied history—e.g. Thomas Paine's 'Common Sense' (Jones and Mattiacci 2019, 740)—the study of rebel and terrorist social media usage is nascent considering the rise of social media use is roughly fifteen years old. These new media and technologies can have a significant and surprising influence on the dynamics of conflict (Aday et al. 2010, Loyle and Bestvater 2019, Shapiro and Weidmann 2015, Zeitzoff 2017), including on recruitment (Blaker 2015) and international persuasion (Jones and Mattiacci 2019). Social media and the ever-evolving Web make it easier to form, finance, and sustain a rebel group (Walter 2017, 481).

Terrorist groups sometimes choose not to claim responsibility for attacks they have undertaken due to a number of strategic considerations. This decision may be driven by a desire to avoid alienating potential supporters who might be sympathetic to the group's broader goals but opposed to specific violent tactics (Hoffman 2010). Additionally, terrorist groups may choose not to claim attacks that failed to achieve their intended impact or resulted in unintended civilian casualties, which could damage their reputation and bolster the image of their competitors (Kearns, Conlon and Young 2014). While violence can demonstrate a group's capability and resolve through costly signaling, it can also undermine its credibility if the violence is seen as illegitimate or excessive by the group's constituencies. As Abrahms and Conrad (2017, 281) note: "When operatives strike a target, their leaders claim credit only if the expected political return is positive." Furthermore, by not claiming responsibility, groups may be seeking to sow confusion and uncertainty about culpability, manipulating media and public perceptions by injecting conspiratorial conjecture surrounding who is really behind the attack. If the deniability is plausible, it may help groups avoid immediate retaliation by complicating the response

efforts of authorities and potentially leading to mis-attribution of the attack.

However, there are reasons to believe that rebel and insurgent groups in active civil wars might use social media differently than do their terrorist counterparts. Terrorist incidents tracked in the Global Terrorism Database (GTD)—the base data source for many studies about attack attribution—occur predominately outside of the context of active civil war. In our data (described in §3.2), we find rebel groups—regardless of ideology or size—claim attacks daily or weekly, to such an extent that it seems highly improbable they could be conducting vastly more attacks than they claim. We posit that in the context of civil war, rebel groups may face different incentives for 'claiming' and for producing accurate information about their behavior on the ground than do international terrorist groups, who may spend months or years planning and executing each individual attack.

Rebel objectives are typically territorially and politically defined, which can make public claims of responsibility a useful tool for advancing their cause and negotiating with governments (Fortna 2015). Gabbay and Thirkill-Mackelprang (2010) provide particular insights into why this might be. While international terrorist actors often seek to influence government decision-making or increase mainstream news coverage with their attacks, rebels may be speaking most explicitly to their local followers and on-the-ground supporters. In Iraq, Gabbay and Thirkill-Mackelprang (2010) find that posts were largely directed at rank-and-file soldiers, who had a great deal of specific local knowledge that enabled them to discern about veracity of claims, leading to greater incentives for rebel actors to be truthful regarding alliance partners, attack timing and other features on which this audience would likely already have information. This dovetails from Mearsheimer (2011)'s argument that when leaders lie domestically, it is often to prevent panic, protect national security, achieve strategic advantages, or protect domestic backlash, but these lies are costly if there is a risk that the leader is caught lying. Thus, rebel group leaders may have greater incentive to lie during civil war, given security issues and the importance of strategic advantage, but these lies may be even more risky given that the population the leader is lying to may be armed, have specific knowledge which enables them

to detect lies, and does not need to wait for an election to force leadership change.

So, when do rebels lie in their own media? Considering Syria and the Islamic State specifically, Milton (2020) notes that rebel groups use deception in three primary ways in their own media: substantive deception (telling a one-sided version of a particular event to support their cause); source deception (misleading about the source of a given piece of information); and spread deception (in which a group lies/misleads about the total amount of support for a given idea or cause). Also studying the Islamic State, Whiteside (2020, 133) notes rebels do lie about information that could harm the group and provide misinformation related to enemies of the group, and Welch (2018) demonstrates strategic changes in the group's public statements across time.

Despite these cases of rebel disinformation, rebel groups tend not to stray far from the truth. "When the Islamic State claims that truth is an important aspect of its media operations, the amount of detail it puts in its reporting and its frequent corrections and clarifications demonstrate that outside of deliberate deception operations, the group walks the talk," (Whiteside 2020, 141). Likewise, Greene and Lucas (2020) find that rebel media can accurately articulate relationships between Hezbollah and other armed actors. Thus, there are also incentives for groups to tell the truth, and they often do. We suggest that the type of data one seeks from rebel-produced media will be important in determining its likely veracity as is the observable context of networked data. Researchers using rebel-produced data must be cognizant of these realities; all inferences drawn from rebel-generated data must be triangulated with other primary and secondary sources to form comprehensive, reliable, and valid conclusions.

It should also be said that the assumption that non-rebel media is inherently trustworthy is equally fraught with risks. Civil conflicts most often emerge in anocracies where freedom of the press is heavily restricted and access by independent journalists is either physically risky or curtailed by authorities. Additionally, readily available global news sources include reports from state media, which can also be viewed as propaganda for the incumbent regime.

2.3 What kind of information can be reliably gathered from rebel-produced media?

We expect there to be varied incentives and capacities for producing and reporting on different types of information. For the purposes of this dataset, we confine our universe of rebel media to "claims of attack"—tactical operations on the ground—to uncover text-based references (Gabbay 2008, Gade et al. 2019, Gade, Hafez and Gabbay 2019) in which rebels publish information about their tactical behavior. While rebels produce a significant amount of media (social and otherwise) regarding their governance behavior, beliefs and ideas, slurs, and exalting statements about the actions of others, our core dataset is comprised solely of rebel group statements about their own engagements on the ground (see §3.2).[3]

There are several benefits to the use of these textual references (Gabbay 2008, Greene and Lucas 2020) over "likes" and "retweets." First, militant groups have control over their textual references, but not over how their social media posts are interacted with. Second, engagement with content on social media may illustrate how surprising or newsworthy it is, not the closeness of the relationship between the author and the audience. Additionally, people engage in following and retweeting for a number of reasons aside from endorsement, making these variables unreliable measures of connection between actors. Finally, a social media post may "go viral" and receive follows and likes outside of the arena of the conflict.

When we expect rebels to lie: First, rebels may have few incentives (or may lack the knowledge) to produce reliable information about battle deaths and civilian casualties. In our data, the majority of statements about casualties contain vague statements: "killing many," "killing dozens," or "significant damage." Likewise, if a group loses an engagement (which they are unlikely to advertise), they are unlikely to accurately count the dead. Many rebel claims of attacks omit causality information altogether. We thus suggest rebel-produced media is not—in the context of Syria—a reliable source for precise estimates of fatalities, battle deaths, or

[3] We excluded some 7,000 rebel statements that we had originally collected because they were not "claims of attack."

civilian casualties; think tank reports, news media, and relief organizations/NGOs are likely to be more reliable than rebel media.

Next, groups have significant incentives to cast aspersions regarding who started a given engagement, as well as who 'won.' Groups likewise face incentives to avoid appearing weak by stating they were attacked or that they lost. This is reflected in how groups tend to describe engagements, sometimes portraying themselves as sympathetic victims who were "viciously attacked" and other times as brave heroes that "stormed the enemy." In our data, groups frequently do not state the outcome of a given fight at all; though they claim engagements without a discussion of a clear victor. For example, "Today our brave fighters stormed a *Nusaryi* checkpoint" does not clearly identify a 'victor' while "Today our brave fighters captured a regime checkpoint, killing many soldiers" appears to do so. Because of the nebulous nature of these statements, rebel media is not a reliable source of information about the outcome or onset of a given fight. However, we note that most news-based event data is likewise an undirected network (including ACLED and UCDP).

Finally, rebel groups may lie about targeting civilians. We have observed that rebels frequently claim to "shell" or "assault" a given town, market, or civilian center, and certain groups—like the Islamic State—are quite transparent in broadcasting their most heinous behavior; there are clearly cases in which groups are willing to claim overt attacks on civilians. However, there are incentives—especially for ideologically moderate groups—to hide the atrocities they commit.

Where we think rebel media is accurate: Groups appear to be quite transparent concerning where they were (location) and who they were with (tactical fighting and collaborative relationships). For example, on August 17, 2017, the official Twitter account of Fath al-Sham Front tweeted:

> "Yesterday, we killed dozens of regime soldiers with our rockets and guns alongside Jaysh al-Fath in Aleppo. The local battalions of the area helped us in our quest."

This quote provides significant detail regarding timing, location, allies, and adversaries: Fath al-Sham claims to have engaged the regime in Aleppo while fighting in coordination with Jaysh al-Fath on August 16, 2017. They also state that they collaborated with "local battalions." Although groups do sometimes state that they fought "with other local groups" without explicitly describing the name of the local group (as above), these incidents make up a comparatively small portion of the total claims of attacks. Far more frequently, rebel groups provide extremely detailed information about their relationships, often including the tactical battalions or units of larger groups that were present on the ground, regional operations rooms, and the hierarchical natures of their own groups, their allies, and their competitors. For example,

> On October 18, the Media Office of the Ahrar al-Sham Brigades posted a claim in which the Abu-Ubaydah Bin-al-Jarrah Battalion of the Al-Iman Brigade, of the Ahrar al-Sham Brigade, claimed to have "destroyed the pick-up vehicles of the 'traitorous security gangs' with a planted explosive device on the Old Industrial Road, killing three *Shabbihas* and wounding two others on board."

Claims most commonly contain detailed information about group relationships with other groups and the location and timing of the incident. This data is particularly valuable to conflict researchers because of the challenges in getting accurate, event-data-style information about tactical cooperation and fighting from the news or think tank data. As a result, we collect *undirected* data: data which states that two groups fought but not who initiated the fight. Rebel-produced media is valuable for an overall count and location of battlefield engagements for a given group, as well as undirected relational data among and within rebel groups (hierarchical networked data), but not casualties, the initiator of a given engagement, or who achieved victory.

2.4 The importance of militant networks in civil wars

When multiple rebel groups are present in a civil conflict, their behaviors affect one another and how civil wars unfold (Bakke, Cunningham and Seymour 2012, Christia 2012, Cunning-

ham 2011, Cunningham, Bakke and Seymour 2012, Krause 2017, Larson 2016). Given this behavioral interdependence and the proliferation of militant groups in modern insurgencies and civil wars, Zech and Gabbay (2016) proposed an inter-organizational network analysis framework for the quantitative and empirical investigation of political phenomena within fragmented conflicts, such as alliance formation, infighting, violent outbidding, splintering, and spoiling. Militant groups within a single movement or conflict are taken as the network nodes; network ties are constructed from discrete, well-defined indicators of inter-group cooperation or conflict; and the network data is integrated with salient group-level attributes such as power, ideology, targeting practices, and territorial presence.

This approach first sought to understand tactical and leadership-level cooperation between Iraqi insurgent groups (Gabbay 2008). Recent work used the social media discourse of Syrian rebel groups (Gade et al. 2019, Gade, Hafez and Gabbay 2019) to consider the drivers of cooperation and conflict among rebels, finding groups with similar ideologies were more likely to cooperate and rebel dyads with greater ideological disparity were more prone to clash.[4] Networks of group interactions have also been constructed from event data based on news reports rather than militant discourse and used to investigate how network structure affects the intensity of militant anti-regime activity (Metternich et al. 2013) and how the emergence of new militant groups affects the level of conflict within the network (Dorff, Gallop and Minhas 2020).

Many outstanding questions persist involving both the causes and effects of militant cooperation and conflict that require networked data. Violent outbidding, for example, has been studied qualitatively and quantitatively (Bloom 2005, Chenoweth 2010, Conrad and Greene 2015, Crenshaw 1985, 1987, Findley and Young 2012, Kydd and Walter 2006, Nemeth 2014). Yet, considerations of network structure may provide important insights into how militant group relationships may restrain or exacerbate shifts toward more extreme violence. For example, a core-periphery structure indicative of good cooperation among larger groups may better resist outbidding pressures from less central groups on the periphery who seek to increase their

[4]This finding of militant ideological homophily in the Syria case was extended in a cross-national (non-network) analysis (Blair et al. 2021).

popularity via more radical targeting practices (Zech and Gabbay 2016). Networked, hierarchical, micro-event data may be able to shed critical insights into a wide range of conflict processes that matter for civil war onset, duration, and outcome.

3 Data and methodology

3.1 Case selection: why Syria?

Over the past decade, the Syrian civil war has devastated a country, displacing as many as twelve million Syrians, with a death toll of roughly half a million, the overwhelming majority of which are civilians. The conflict merits in-depth study to better understand the factors that contributed to its intensity, duration, and outcomes. Three specific elements of the civil war make it particularly invaluable for the broader study of day-to-day rebel interactions—the fragmentation of rebels into distinct factions; the asymmetries in factional ideologies, military power and organizational structures; and the abundant use of social media.

Over the course of the war, rebel groups emerged and collapsed, gained and lost territory, and reframed their goals as the strategic landscape changed. New geographic, political, and administrative divisions emerged as rebel groups formed and disbanded, shifting power dynamics and altering networks of cooperation and infighting (Abboud 2016). This breadth of data creates the opportunity to understand both the behavior of those diverse groups and how their interactions changed over time.

Furthermore, state and parastate entities (such as the Islamic State and Hay'at Tahrir al-Sham), as well as transnational foreign fighters and pro-regime militias from Lebanon and Iraq, contributed to the complexity of alliances, requiring network analytic tools to both visualize and analyze their relationships. Regional states such as Turkey, Iran, and Gulf countries expanded an intrastate conflict into the realm of regional rivalries and intervention by Russia and the United States introduced strategic competition through proxy warfare into the conflict (Al-Sayed Hussein 2013, Darwish 2020). Thus, Syria offers fertile soil for considering the

impact on conflict trajectory of external states and nonstate actors that played significant roles in resourcing rebel groups or the Syrian government in battlefield operations.

Lastly, the social media landscape has enabled rebel groups in Syria to advance their cause by providing a platform for communication, recruitment, and propaganda. Social media platforms have been instrumental in highlighting political grievances and regime atrocities, and documenting daily activities, alliances, and inter-rebel conflicts. This treasure trove of information offers an opportunity to lift the fog of war on inaccessible conflict zones to better understand dynamics of alliances, fragmentation, and infighting.

The Syrian civil war is not unique in this regard. It is one of many fragmented, multiparty civil conflicts featuring hundreds of armed groups with varying interests, ideologies, and state sponsors. As such, Syria fits the prevailing pattern of civil conflicts since 1945 (Bakke, Cunningham and Seymour 2012). According to (Jones 2017, 168), of 181 insurgencies since 1946, more than half involved multiple insurgent groups; since the 1980s, 64% involved multiple rebel factions. Indeed, a growing body of scholarship substantiates the multi-factional nature of civil conflicts and proceed to analyze how this fragmentation shapes the choice between violent and nonviolent strategies (Pearlman 2011); increases the odds of protracted civil wars (Cunningham 2011); results in unstable rebel coalitions (Christia 2012); and affects the ability of nationalist movements to win their independence (Krause 2017).

The Syria civil war features multiple rebel groups that represent a diverse range of Islamist and non-Islamist factions, as well as variation in power capabilities and degrees of ideological moderation and extremism. This is not dissimilar to the Iraqi insurgency in the 2000s or Afghanistan in the 1980s and 2000s. The Libyan civil war since 2011 has also featured hundreds of armed groups and external state intervention, as did the mobilization of Iraqi Shia militias against the Islamic State in Iraq with the aid of external intervention beginning in 2014. Beyond the Arab or Muslim worlds, we can see the prevalence of fragmentation similar to Syria in the first decade of the Sri Lankan civil conflict during the 1980s, and during the Liberian first and second civil wars (1989-2003), which also featured several armed factions that competed

with each other as they fought the state. Presently, armed conflicts in Mali and Myanmar also exhibit a dizzying array of armed actors that form coalitions, splinter into competing camps, and fight with rivals. Because the Syrian civil war is not *sui generis*, it offers generalizable and testable inferences about conflict dynamics and the utility of network analytic methods in similar cases.

3.2 Dataset Design

The Syrian Armed Groups, Alliances and Actors (SAGAA) dataset is comprised of four complementary data components: (1) an event dataset based on militant social media posts; (2) collaborative and conflictual networks among rebel and state actors drawn from the event dataset; (3) a systematic tracking of the aliases of major militant groups in Syria (different names, transliterations, front memberships, etc.); and (4) group-level co-variate data for sixty-seven distinct groups, drawn from think tank reports, existing datasets, and web searches.

Dataset	Source	Details
Event Dataset	Translations of rebel social media posts	28,000 hand-coded social media posts, 7,000 of which were excluded because they were not claims of attack. Each of the remaining 21,000 claims of attacks were coded for the variables described in §C.
Collaborative and Conflictual Network Dataset	In-text references about relationships among groups drawn from claims of attacks	Direct textual references about collaborative and conflictual relationships among groups.
Hierarchical Dataset	In-text references about relationships among groups drawn from claims of attacks	This focused on hierarchical relationships including parent-subgroup ties, alliances, and joint operations rooms.
Group-level Covariate Dataset	Think tank reports, NGO reports, rebel's own media, news articles, google searches	We selected the 67 most prominent groups in our collaborative and conflictual networks (this cut-off was based on group degree in each network) and modestly updated that list as groups splintered and merged. Following detailed coding ontology presented in §C, we collected group-year level data for the most prominent groups in the network.

Table 1: Descriptions of each component of the SAGAA dataset

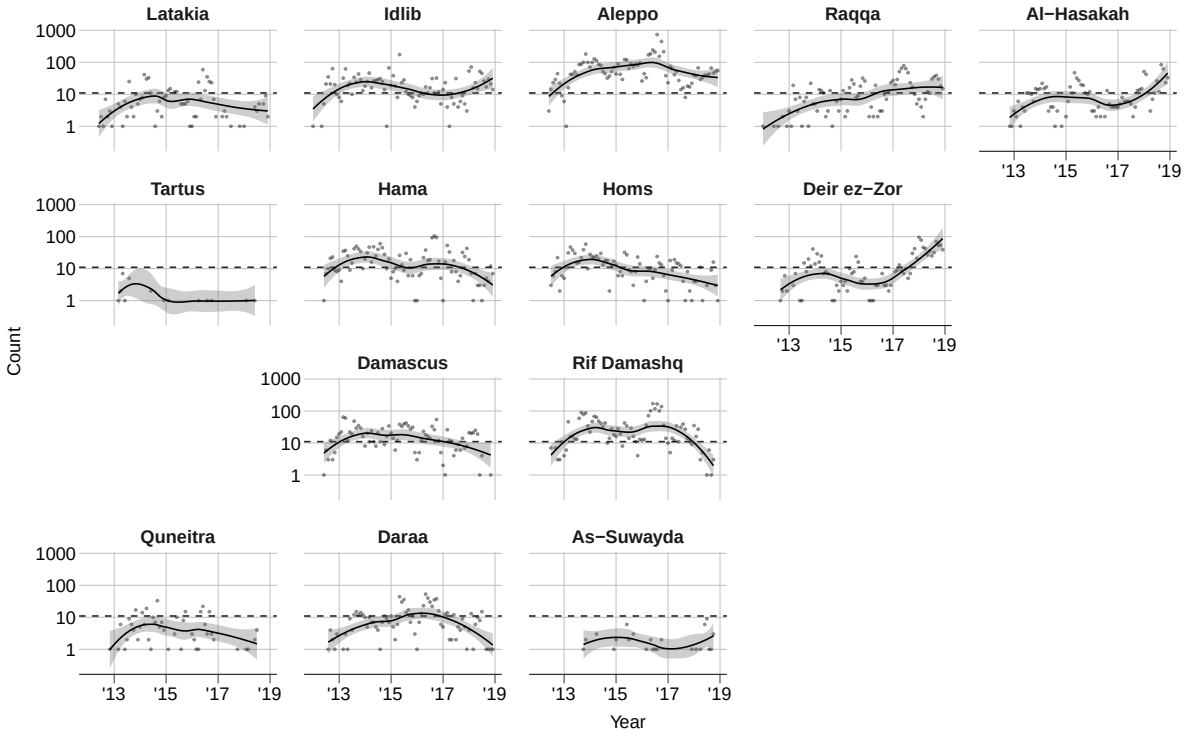

Figure 1: *Number of militant claims of attacks by Syrian governorate (2012-2018).* Each data point represents the number of attack claims by militant groups in a given governorate-month. The governorate panels are arranged in a grid that roughly resembles the geography of Syria. Smooth curves are estimated by the LOESS method. Dashed guide lines show the median number of claims (11). Y-axes are log(10)-scaled.

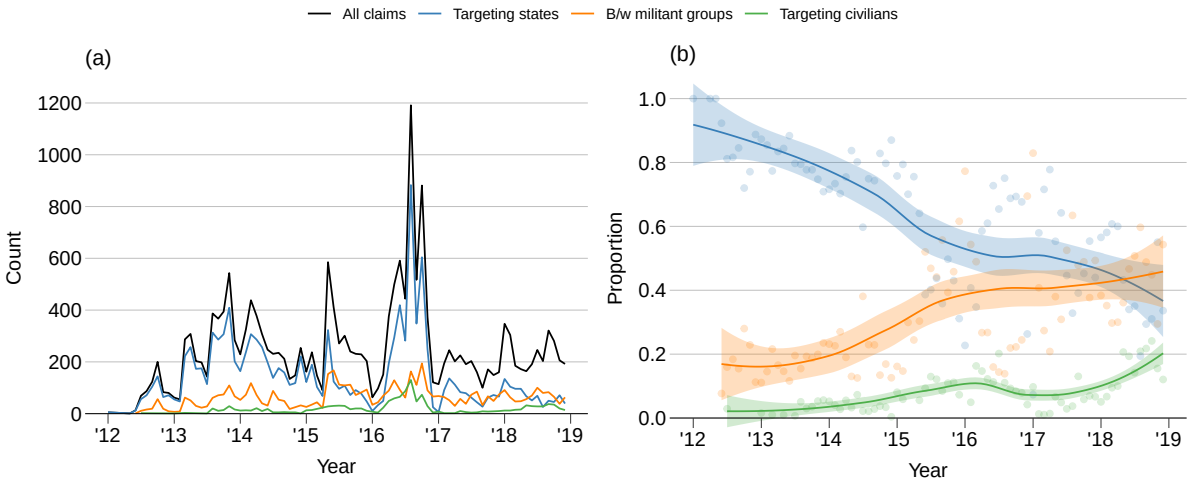

Figure 2: *Targets of militant claims of attacks (2012-2018).* Panel (a) shows the raw number of claims in each month across three types of targets – against states, militant groups, civilians – and their total sum. Panel (b) shows the target-type specific proportion (out of all attack claims) in each month. The smooth curves are estimated by the LOESS method.

Our core data components—Table 1—involve claims of attack reported by militant groups in Syria between January 2012 and December 2018. These 28,000+ claims (collected and translated by the U.S. Government) originated from Twitter, Facebook, YouTube, and local news within the conflict zone. Figure 7 displays the time series of attack claims in the SAGAA dataset by the geography of governorates, while Figure 2 visualizes the time trends across target types. Appendix §C contains details about data collection, coder training, and the complete ontology used (including what we count as civilian targeting, for example).

3.3 Constructing Rebel Conflictual and Cooperative Networks from Event Data

We construct a conflictual network among rebel groups with the following procedure: If one or more militant group(s) claimed to have attacked another group (or more than one), a conflictual tie is attributed to each unique initiator-target pair.[5] For example, if Groups A and B claim to have engaged in a conflict with Groups C and D, the following conflictual (and undirected) ties will be established: $A \leftrightarrow C$, $A \leftrightarrow D$, $B \leftrightarrow C$, and $B \leftrightarrow D$. Our construction of conflictual ties is intentionally undirected: given the incentive for militant groups to misrepresent the initiator of an attack, we do not have perfect information about who initiated a given fight. Therefore, as long as either observation i or j claimed that they engaged in a conflict with one another, both element x_{ij} and x_{ji} will receive a value of 1 in the adjacency matrix $\boldsymbol{X}^{Conflict}$.

Iterating this procedure over all claims in our dataset, we construct a conflictual network represented as an adjacency matrix $\boldsymbol{X}^{Conflict}$, which is of size $n \times n$ (where n represents the total number of unique rebel groups), and in which element x_{ij} represents the total number of instances where a conflict involving actor i and j was claimed to have occurred throughout our dataset. In the language of social network analysis, $\boldsymbol{X}^{Conflict}$ is an undirected (i.e. symmetric) and weighted network, summarizing the structure of conflictual relationships among rebel

[5] We apply a criterion that the targets must involve militant groups or Syrian or foreign governments, and exclude attacks solely targeting civilians or other non-state, non-militant actors.

groups.

Similarly, we construct a collaborative network based on the definition of joint operation in a combat: If Group A, for example, claimed to have engaged in a conflict with Group C and D in collaboration with Group B, a collaborative (and undirected) tie will be established for each possible pair on each side, which is $A \leftrightarrow B$ and $C \leftrightarrow D$ in this case. The resultant adjacency matrix $\boldsymbol{X}^{Collab}$ has each element x_{ij} representing the total number of instances where a collaboration involving actor i and j was claimed to have occurred throughout our dataset.

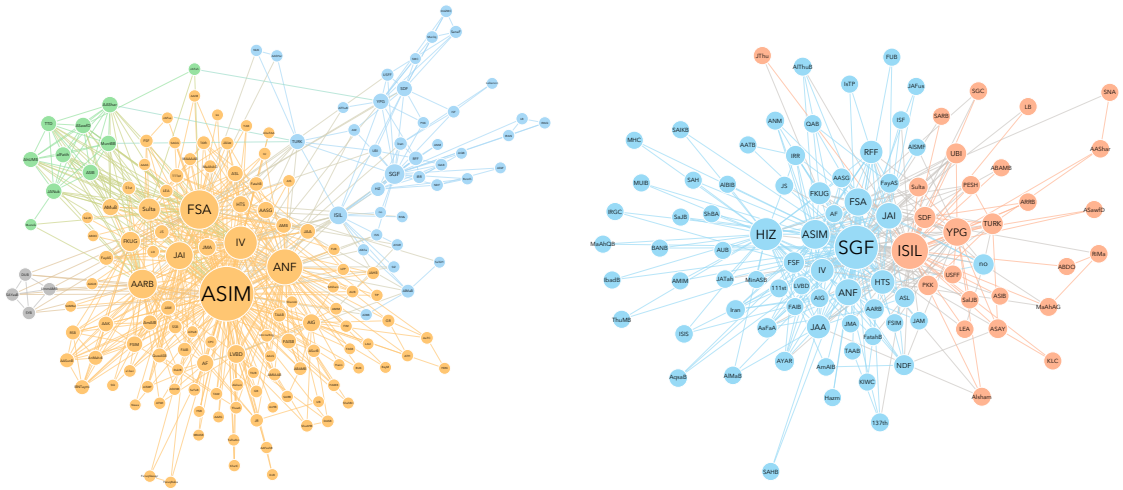

Figure 3: *Structures of collaboration and conflict in the Syrian Civil War.* The **left** panel shows the collaborative network in which a node represents a group. A line connecting two nodes indicates a collaborative relationship; the size of a node reflects the total number of relationships it has. The underlying community structure is detected by the Louvain algorithm and each node is colored accordingly. The **right** panel shows the conflictual relationships and communities using the same methodology.

Figure 3 visualizes these networks. In each graph, each node represents a rebel group, and each line connecting two nodes represents a collaborative (or conflictual) relationship between them. We scale the size of each node in proportion to the sum of collaborative (or conflictual) ties a group possesses. We further implement the Louvain algorithm (Blondel et al. 2008) to offer suggestive evidence of the community structure underlying the actors, and color and cluster the nodes according to the community they are assigned to.[6] The left

[6]The Louvain method is based on optimizing modularity (i.e., maximizing ingroup links' density) and produces non-overlapping communities (i.e., each node can only belong to one community).

panel of Figure 3 visualizes the collaboration network and reveals a cluster of rebel groups that collaborated frequently with one another, most prominently led by Ahrar Al Sham Islamic Movement (ASIM), the Free Syrian Army (FSA), Al Nusra Front (ANF) and others. The right panel of Figure 3 visualizes the conflictual network and maps the two main clusters of conflictual relationships, most prominently led by Syrian Government Forces (SGF) and Islamic State (ISIS) respectively. This second plot does not imply that many groups collaborate with the government—rather, that the conflict is divided into groups who primarily fight the state, and groups who primarily fight ISIS.

4 Evaluating Rebel Media

4.1 Comparison to Other Data Sources

We compare the SAGAA dataset to the current gold standard, UCDP Syria dataset, to evaluate the veracity of SAGAA as well as consider what rebel media might be able to tell us that conventional news cannot. The intuition here is that to be useful for conflict researchers, rebel media should be able to reproduce some or most of the networked information drawn from news data, which we know to be factually accurate, but likely incomplete. The Uppsala Conflict Data Program (UCDP) is the world's oldest collection of civil-war related data. UCDP's data is collected via string searches through the Dow Jones news aggregator to identify all articles which involve death or injury for each particular conflict. For full details and differences in coding protocol see Appendix §A.2.

To compare the data sources, we first mapped the group names from UCDP Syria to SAGAA.[7] The UCDP data contains 76 unique groups for our time period (2012-2018) while SAGAA contains more than 400 for the same period. The SAGAA dataset contains 92% of

[7]We carefully combed over our master list of group names/name variations and transliterations. When an exact UCPD unique group name was not found in the SAGAA dataset, we collected external information from think tanks, news sources, and search engines, in particular looking for potential variations in or alternative group names, spellings, etc. All coder decisions and outside sources are documented in the 'coder notes' and 'outside sources consulted' columns respectively.

the groups listed in the UCPD dataset, while the UCDP dataset contains less than 16% of the groups listed in the SAGAA dataset. As well, between 2012-2018 UCDP contains triple the number of incidents as the SAGAA dataset (74,000 as opposed to 21,000). However, 78.9% of the UCDP incidents are from a single dyad—described as "Syrian Government v. Syrian Insurgents." While the Gov. v. Syrian Insurgents dyad may be extremely useful for scholars seeking to understand government use of force or overall violence levels in the area, for scholars interested in conceptualizing the Syrian insurgency, "Syrian Insurgents" is not a helpful label. In this case context, only a tiny fraction of the data collected from mainstream sources is actually directly attributable to specific rebel groups on the ground. In contrast, rebel-produced media is extremely specific regarding who did what and to whom.

Moreover, a significant proportion of identified data (e.g. not "Syrian Insurgents") in the UCDP dataset is focused on government attacks and attacks by the Islamic State. This mirrors our expectations; the Syrian state and ISIS seem likely to be a center-point of conventional news reports while prominent but less 'splashy' (e.g. non-ISIS) rebel-rebel interactions might not make mainstream news coverage.[8]

To compare UCDP and the SAGAA dataset, we subset the SAGAA data to include only the universe of groups identified by UCDP—this provides a "tough test" as the SAGAA dataset will be penalized for any missing groups. Next, we construct separate conflict and collaboration adjacency matrices for both datasets. For SAGAA, we construct these matrices following the protocol detailed in §3.3. The UCDP conflict adjacency matrix is constructed by summing up the number of conflictual interactions a pair of actors had between 2012 and 2018, and the collaboration matrix sums up the number of times a pair of actors are recorded as having been involved in the same side of an event. However, the average density of the two datasets is quite different—UCDP weighted conflict network is ≈ 8.1 compared to ≈ 1.5 in SAGAA. However, much of this difference is because UCDP tracks the number of events involving the Islamic

[8]The primary dyads in the UCDP dataset are: Government of Syria v. Syrian Insurgents (78.9% of UCDP incidents); Gov. Syria v. Islamic State (6.7% of incidents), Islamic State v. Kurdish Syrian Defense Forces (3.6% of incidents) and Gov. Syria v. Civilians (1.7%).

State. Simply removing the Islamic State as a node drops the density of the weighted conflict network in UCDP to ≈ 0.6. We thus focus on comparing unweighted versions of the SAGAA and UCDP, the cross-sections in these adjacency matrices are 0 if no event was recorded between a pair of actors and 1 if there is an event recorded. Comparing the unweighted versions also allows us to better understand how these two datasets can complement one another.

4.1.1 Descriptive Comparison

We calculate common network level statistics for both the SAGAA and UCDP versions of the conflict and collaboration networks and present the results in Figure 4 below. The first column of plots describes the conflict network and the second collaboration. The network-level statistics that we calculate are (1) the density of the network, (2) heterogeneity in actor degree, and (3) transitivity. Density in the case of this binary network simply provides a representation of how many ties actually exist (out of all possible ties) in each dataset.[9] In both cases of conflict and collaboration, we can see that there are more ties captured in SAGAA than in UCDP. This difference is particularly notable given that we are restricting the SAGAA sample to only actors that could also be identified in the UCDP dataset. We investigate the heterogeneity in actor degree by first computing every actor's degree centrality (i.e., the number of ties a given actor has in total) and then taking the standard deviation of all degrees. SAGAA having higher values here indicates that the number of other actors a group is reported as interacting with is greater than what is in UCDP, but it is important to note that the difference between the two datasets here is marginal.

Furthermore, SAGAA reports higher transitivity in both networks compared to UCDP. Transitivity captures the degree to which "a friend of a friend is a friend,"[10] an important

[9] Note as described in the previous section to facilitate comparisons, the number of actors in both networks are the same.

[10] Transitivity is also known as a measure of "triadic closure." A triad in a network consists of three nodes, which can be connected by two ties (an open triad) or three ties (a closed triad). We first compute the local transitivity for each node, which is the ratio between the number of closed triads and the number of all triads (both open and closed) it involves in, then averaging the measures over all nodes. Intuitively, it represents how likely that a tie $i \leftrightarrow k$ exists, given we observe the ties $i \leftrightarrow j$ and $j \leftrightarrow k$.

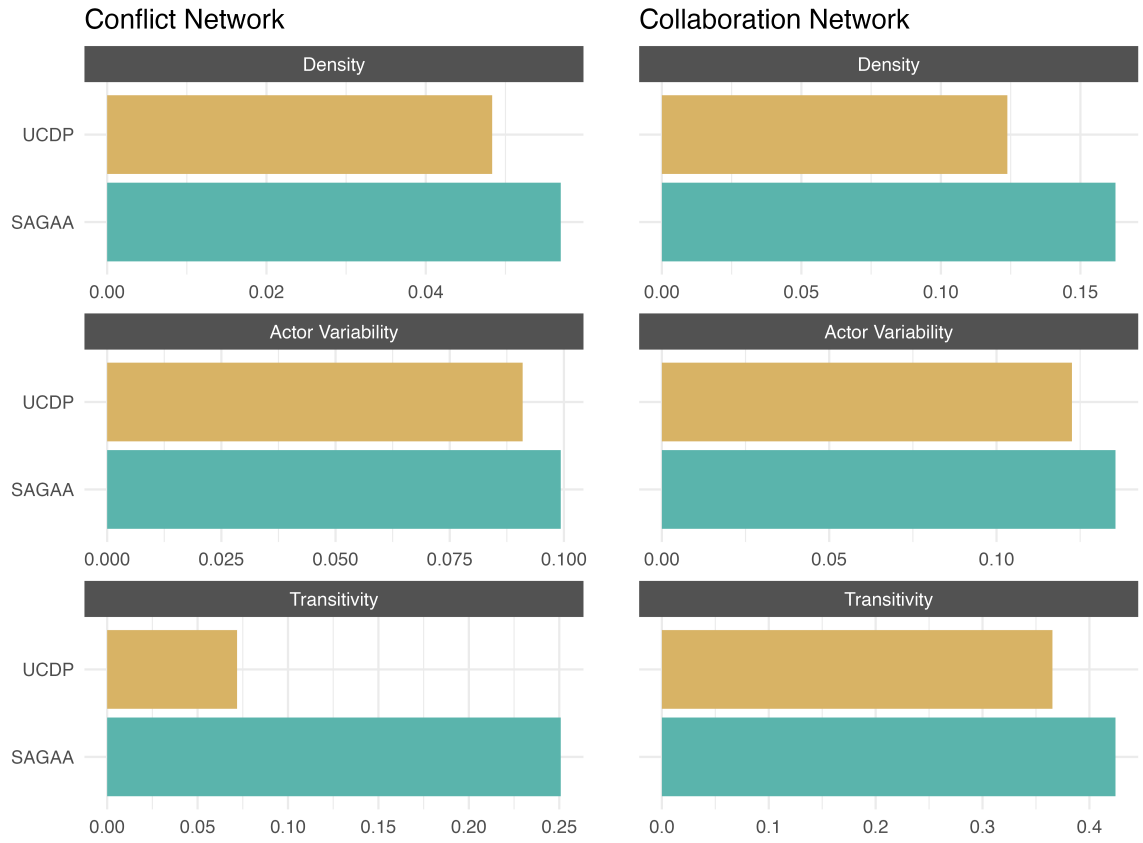

Figure 4: *Graph-level summary statistic comparisons of SAGAA and UCDP for Collaboration (on the left) and Conflict (on the right) networks.*

network concept in understanding civil conflicts (Dorff, Gallop and Minhas 2020) and international cooperation (Kinne 2013, Manger, Pickup and Snijders 2012). Most interestingly, higher transitivity in the conflict network from SAGAA indicates that we observe more rebel groups who shared a common enemy still fought one another, compared to the extent we see in UCDP. This type of "imbalanced" relationship (that the enemy of my enemy is not my friend but my enemy) has been investigated in the context of interstate wars (Maoz et al. 2007) and raises new questions in the civil conflict literature.

More generally, higher levels of transitivity in SAGAA networks hint that SAGAA is capturing small-scale engagements and less prominent groups more completely than UCDP, thus providing a more complex picture of the overall system of collaboration and conflict.

Figure 5 depicts the actual conflict and collaboration networks from both sources. The

left column of networks showcases conflict and the right collaboration. The nodes in each row of the figures are the same and actors are positioned in the same place; the difference between the data sources are the edges (or ties) between nodes. In the top panel of both figures, a tie is shown between a pair of nodes if they had conflict/collaboration with one another according to the SAGAA dataset but not in UCDP; in the middle panel if actors had a tie in UCDP data but not SAGAA; and the bottom panel highlights instances of conflict and collaboration captured by both datasets.

In the conflict network, 92 of the same dyadic ties exist in both datasets, while 82 only exist in UCDP and 113 only in SAGAA. The difference in collaboration is even more stark, where 98 exist in both, 164 only in UCDP, and 246 only in SAGAA. In both cases, more unique ties are captured by SAGAA than UCDP—38% more in the case of conflict and 50% more in the case of collaboration. Additionally, in the conflict networks, the central role of ISIS is recognized in both data sources, but SAGAA finds more conflictual interactions involving Al-Yarmuk Brigade (AYAR) while UCDP finds more involving Partiya Yekitiya Demokrat (PYD).[11] On the collaboration side, the differences between these two sources is quite stark. While the notable role of Fath al-Sham (FSF) is acknowledged in both data sources, the SAGAA dataset captures complexity unmatched in UCDP. In the next section, we show that the denser structure and greater complexity of ties captured by SAGAA provides a meaningfully different understanding of the overall conflict and collaboration system. We believe these findings demonstrate notable, substantive updates to our understanding of rebel behavior in multiparty civil wars.

[11]Because SAGAA is drawn from U.S. government translations of rebel media, we expect that there is an *overemphasis* on non-Kurdish groups. Kurdish groups have been U.S. allies in this conflict, and it would appear that the U.S. was less focused on collecting and translating Kurdish-produced media than it was on the media of Islamist/Jihadist armed actors.

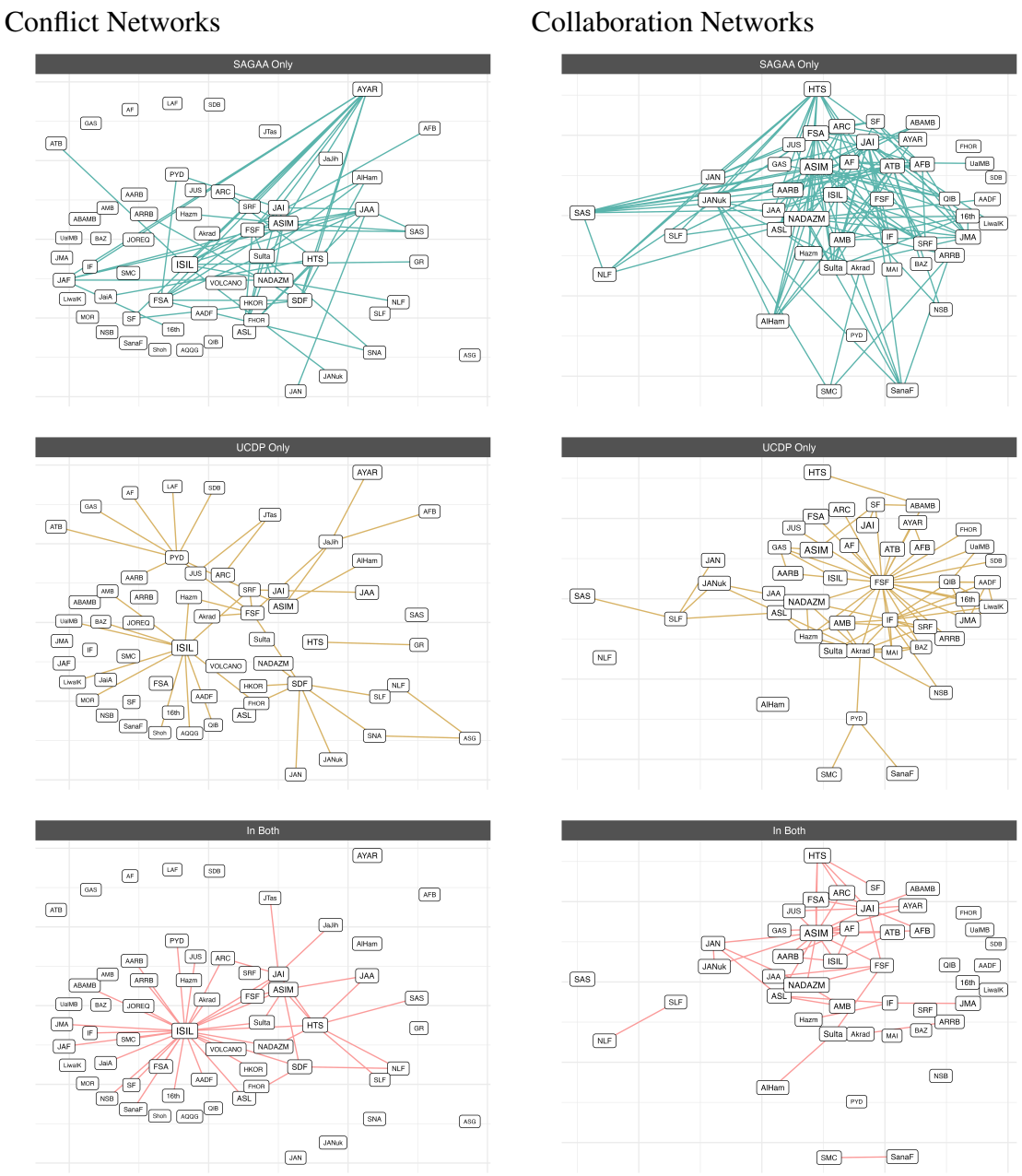

Figure 5: *Comparison of conflict (left) and collaboration (right) networks between SAGAA and UCDP.*

4.1.2 Predictive Comparison

We assess the quality of the information in the SAGAA and UCDP datasets by examining their ability to reconstruct each other, taking into account network dependencies. We begin with the assumption that both datasets provide incomplete representations of the conflict (Demarest and Langer 2022, Donnay et al. 2019). One reason traditional media sources may not capture all events is the lack of attention given to some groups. We observe this in the Syrian context: most conventional news reports focus on actions involving the Syrian State or ISIS, providing at best an incomplete representation of the full conflict system. SAGAA is likely to suffer from some level of incompleteness as well. However, if each dataset contains enough information, we can capitalize on the underlying structural properties of the system that we do observe. The key element that enables the reconstruction of these systems, when we only have a fragment, lies in the inherent interdependencies between actors and the patterns of connectivity among them. For example, if we observe in the raw data that an actor i is collaborating with actors j and k, then j and k are more likely to collaborate with each other than some fourth actor l that none of them share a relation with. Additionally, actors that are more central in the network should be more likely to have an interaction with anyone else than those that are less.

To reconstruct the system, we apply a network-based latent variable approach to infer the missing values; here, the Additive and Multiplicative Effects Network (AMEN) model (Minhas et al. 2021, Minhas, Hoff and Ward 2019). The AMEN model provides a probabilistic representation of the likelihood of a tie between a pair of actors based on the average density of the network, the tendency of certain nodes to form more or less relationships, and the relationship the pair has with other actors in the system. We formulate the model as follows:

$$y_{ij} \;=\; f(\theta_{ij}), \text{ where} \tag{1}$$

$$\theta_{ij} \;= \beta_0 + a_i + \mathbf{u}_i^T \Lambda \mathbf{u}_j + \epsilon_{ij} \tag{2}$$

$$\tag{3}$$

a_i captures variability in actor ties, and the $\mathbf{u}_i^T \Lambda \mathbf{u}_j$ specifies a matrix decomposition approach to embedding the actors in a lower-dimensional social space based on the third-order connections that actors might have. We apply this approach to both the UCDP and SAGAA data separately and analyze whether SAGAA can actually predict ties in the UCDP dataset and vice-versa. This allows us to evaluate the added value that the SAGAA dataset might have over UCDP or vice versa. If the SAGAA dataset can recover information from the UCDP dataset—which we consider factually accurate but likely incomplete—this should give confidence that the other information also in the SAGAA dataset but *not* in UCDP has a similar level of accuracy.

More formally, we denote the SAGAA and UCDP observed networks as Y^S and Y^U, respectively. In order to, "fill in" missing values we use the latent variable model described in Equation 3, where we supply as input the observed network and then draw out predicted values. We employ a 10-fold cross-validation procedure to evaluate how well the predicted networks of one dataset match the observed values of the other. This process proceeds as follows:

- **Divide the data points**: Randomly divide the $n \times (n-1)$ data points of both Y^S (SAGAA) and Y^U (UCDP) into $K = 10$ sets of roughly equal size. Assign each pair ij to a set, denoted as k_{ij}.

- **Predict Y^U using Y^S**:

  - For each $k \in 1, \ldots, K$, perform the following steps:

    * Estimate the model parameters based on $y^S ij : kij \neq k$, which is the data on

pairs not in set k.

* For pairs lm in set k, compute the predicted value of y_{lm} as $\hat{y}_{lm} = E[y_{lm}|y^S ij : kij \neq k]$. This prediction is obtained using data not in set k.

– Generate a sociomatrix of out-of-sample predictions for Y^S, denoted as $\hat{Y}^S$.

– Evaluate the performance of $\hat{Y}^S$ in predicting Y^U by computing metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the Area Under the Precision-Recall Curve (AUC-PR).

- **Predict other combinations**:

  – Repeat the steps 2a to 2c, but this time use Y^U to predict Y^S.

  – Repeat the steps 2a to 2c, but this time use Y^S to predict itself.

  – Repeat the steps 2a to 2c, but this time use Y^U to predict itself.

- **Compare performances**: Assess the predictive performances of SAGAA and UCDP by comparing the AUC-ROC and AUC-PR scores obtained in steps 2c and 3c. This comparison will help determine the added value of one dataset over the other.

By following this procedure, we can assess the ability of each dataset to predict itself and the other in an out-of-sample context. The analysis is conducted for both conflict and collaboration networks, and the performance results are presented in Figures 6.

The performance analysis generates several important findings. First, methodologically, this approach shows potential: we successfully use the networks to predict themselves out-of-sample even without using any exogenous covariates. The high AUC-ROC and AUC-PR scores for SAGAA-SAGAA and UCDP-UCDP indicate that we can effectively predict the original data using this network reconstruction approach. More interestingly, SAGAA is able to predict UCDP almost as well as UCDP can predict itself for conflict and *better* than UCDP can predict itself for collaboration. The network-reconstructed version of UCDP performs notably worse in reproducing the information in SAGAA in both conflict and collaboration cases. We suggest

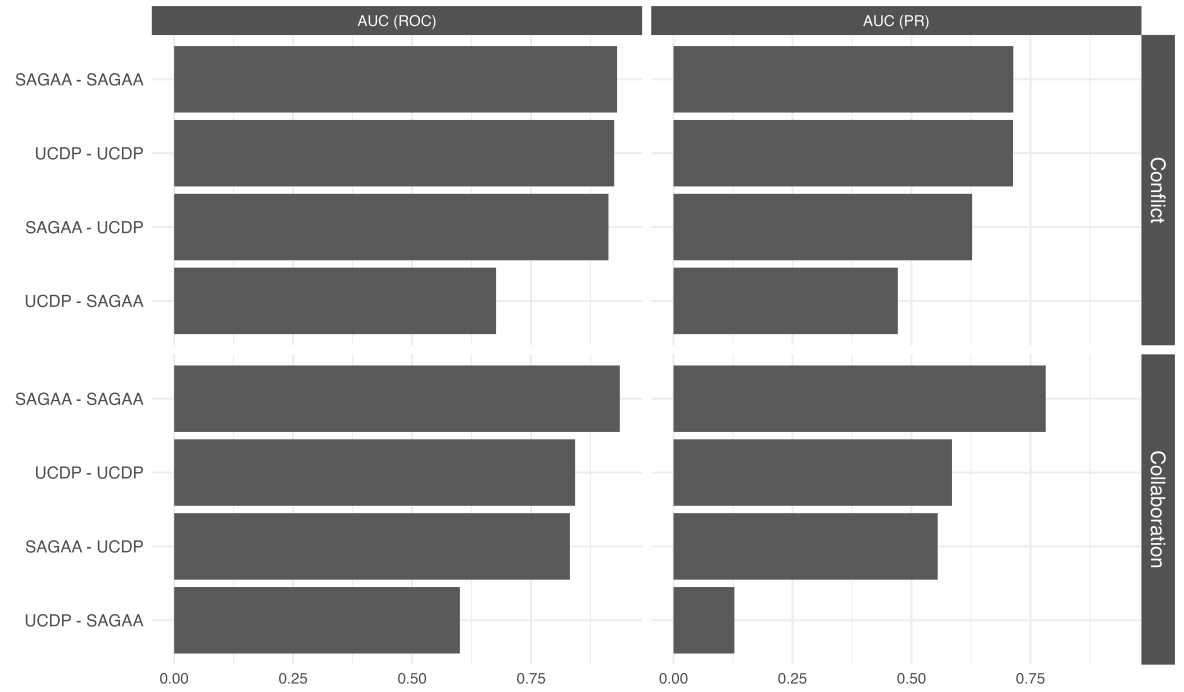

Figure 6: *Predictive performance of conflict (top) and collaboration (bottom) from both data sources in predicting themselves and each other in a cross-validation context.*

this is because SAGAA contains more network information (the network is denser and more complex) than UCDP, which enables better predictions of relations between any pair of actors.

We note that the additional networked information captured in the SAGAA is predominantly information about *smaller* rebel actors. Indeed, networked information about smaller rebel groups seems to aid in understanding the behavior of larger actors, more so than the behavior of larger actors alone; in the UCDP data, smaller actor data is missing. This finding is of significant substantive importance: existing conflict literature has a major bias towards larger conflict actors in multiparty civil wars. Our results suggest that it would behoove scholars to attend to the behavior of secondary actors in these conflicts in order to best understand how and why conflicts unfold the way they do.

4.2 Qualitative Check: Patterns in the Syrian Civil War, Idlib Case

To contextualize these findings, we qualitatively investigate the patterns detailed above. We select a prominent rebel group, Hay'at Tahrir al-Sham (HTS), and systematically consider the

similarities and differences between news reports on the group's activity with the rebel's own social media posts. Our findings align with the conditions under which rebels are likely to lie and be fully forthright (§2.3), highlighting both the promise and pitfalls of using rebel media as event data. Additionally, our investigation uncovered that rebels capture details that are lost in the news; in particular, the SAGAA data set—and, we believe, rebel media more generally—will provide more detailed data about the micro-events of rebel influence, power, and fighting and collaborative relationships.

We selected four major events involving HTS *covered in the news*—introducing selection bias which is unfavorable to SAGAA. Table 2, below, lays out the findings.[12] The qualitative comparison reveals a logical conclusion: rebel media tends to lie when there are incentives to do so. But even in these cases, groups tend to be both reliable and forthcoming regarding location, nature, and timing of events, as well as with whom they fought and collaborated. For example, in each event, the dates, locations, and adversaries described by HTS were corroborated by news outlets (even while the group was deceptive about other elements). Additionally, HTS-generated news and news reports were in agreement about the circumstances of the 2nd and 3rd events investigated. A disconnect between the media types only occurred in the 1st and 4th events, where HTS had ulterior motives for concealing the truth. HTS's denial of attacking Shi'ite civilians demonstrates rebels' motivations to lie about targeting civilians. The 3rd and 4th events exposed HTS portraying themselves as victors regardless of the actual result, illustrating that rebel media is an inconsistent source of battle outcomes. We suggest that rebel media both contain important information that is absent from news data and that rebel media is best used with judicious attention to the types of data being collected. This data may be most beneficial to evaluate specific topics and/or when used in tandem with traditional data resources.

[12]For additional information concerning HTS as well as methodology and source selection please see Appendix A.2.

Event Description	Event Date	News Coverage	HTS-produced Media
Event 1: HTS targeted a Damascus Shi'ite pilgrimage site	March 11th 2017	• Alleged HTS responsibility for the bombing • Directed International scrutiny onto HTS for targeting civilians	• Initially claimed "responsibility for twin attacks in Damascus targeting 'Iranian militias' and National Guard forces," • Then denied "any connection to the Damascus blasts," and claimed to target "only the security branches and military barracks of the criminal regime and its allies."
Event 2: Syrian Government Forces (SGF) launched an extremely brutal chemical attack on HTS-held territory	April 4th 2017	• Described as the "deadliest chemical attack in Syria since the August 2013 assault" • Portrayed the assault and victims' suffering • Provided casualty estimates • Detailed international responses and criticisms	• Disseminated accounts of the event • Focused on the experiences of those harmed • Showcased regional political reactions, highlighting, "a protest in the city of Khan Shaykhun in the Idlib countryside denouncing the chemical massacre committed by the criminal regime
Event 3: HTS and Ahrar al-Sham (ASIM) engaged in military clashes, which culminated in a ceasefire	July 23rd 2017	• Reported on the ceasefire • Documented the reallocation of previously ASIM-held territory into HTS control, including the strategic Bab al-Hawa crossing • Framed the incident as an HTS triumph	• Announced ceasefire and described its' conditions • Recorded territorial acquisitions, including taking "control of the 106th region overlooking the Bab al Hawa crossing after Ahrar al-Sham fighters withdrew from it." • Portrayed as an HTS victory
Event 4: HTS experienced military decline after the Syrian Liberation Front (SLF) successfully attacked HTS strongholds	February 2018	• Chronicled various skirmishes between HTS and SLF • Monitored battle outcomes and documented territorial losses on both sides • Still, the focus remained on the challenge to HTS and on HTS's continuous retreat across various regions of Idlib and Aleppo	• Rarely acknowledged defeats and/or excused them • Conceded the loss of "Tarmala and Ma'arrat an Nu'man in the southern Idlib countryside" • Yet, withheld announcing other territorial turnovers • Claimed they were "forced to retreat from some regions to preserve civilian life." • Most statements depicted their own victories and land

4.3 Applications

Rebel-produced media could be used to answer significant questions of concern to conflict researchers. For example, how do changes in rebel endowments influence attack behavior? When are large attacks followed by a lessening in violence versus a retaliation or escalation? How does weather—or religious holidays—influence attack type and frequency and how does this differ among rebel groups? How are new types of attacks adopted and do they spread throughout existing networks, suggesting socialization or mimicry? When powerful new rebel groups or foreign actors enter the conflict, does this drive increased fractionalization or cooperation among smaller groups? Do geographical features like mountains and rivers impede cooperation among like-minded groups?

5 Conclusions

Using both a unique dataset and novel methodological approach, we demonstrate that rebel-produced media is particularly valuable for understanding the timing and location of rebel violence and rebel relationships with other groups in complex conflicts; indeed, it outperforms existing gold-standard news-based datasets in its ability to do this. Rebel media also offers researchers an in-depth look at day-to-day militant group cooperative and conflictual behavior over time, allowing for rich theorization about and testing of the factors that lead to militant group alliances or fragmentation, the impacts rebel relationships have on the broader trajectories of conflict, the impact of state use of force on these rebel-rebel relationships, or rebel tactical choices on the ground. We expect these findings to generalize both to other conflict contexts and, potentially, to cases of ongoing protest. As complex and unfolding protest movements sweep the globe and populism continues to rise, considering how and under what conditions scholars and policy practitioners alike can collect and leverage data coded from non-state actor media to fill voids left by mainstream news is a timely and critical question.

Efforts to build data about the Syrian conflict, and particularly about the relationships

between groups, faces significant challenges. Rebel groups indeed (as Ostovar and McCants 2013 suggest) changed their names, merged, broke apart, disappeared, formed "joint operation rooms," fought one another while fighting the state, and formed alliances that dissolved soon after. We also faced both translation and transliteration issues, as groups' names are discussed differently in different contexts, referred to using slang terms: e.g. *Daesh* for the Islamic State; *Shabiha*—literally "ghosts"—and *Assad's dogs* for pro-government militia; or *Nusayri* for Syrian armed forces. Many groups also had similar names, including sub-groups and small battalions taking on their own names despite not being independently operational. While we carefully double-checked this dataset (Appendix §A) and verified our dataset (§4), our ontology (Appendix §C) is complex, detailed, and nuanced, meaning that there are certainly errors, remaining issues with translation and transliteration, etc. As well, while this dataset includes information about attacks between state actors and militant groups, it does not capture frequent attacks by Syria, Russia, the US and allies, Iran, Turkey, or other outside groups, especially those which target civilians. Lastly, while the complexity and length of the Syrian conflict render it an excellent candidate for study, more examination of the micro-dynamics of other recent conflicts would enable richer theorization about the structural factors that shape militant conflict and cooperation. As is best practice with any observational dataset, we suggest downstream users run sensitivity analyses on primary results.

With expanding social media use across the globe and the continued presence of low-level armed conflict, non-state actor media in general—and rebel-produced media in particular—offers an important complement to news data for researchers. Understanding the conditions under which actors are likely to lie (§2.3) and where they are likely to be transparent and cleaving to those contours can help ameliorate concerns about data reliability. As cell phone access and social media use continues to rise, and as the targeting of foreign journalists makes many conflict spaces wholly inaccessible, we suggest that non-state actor-produced media can offer significant advantages (when used judiciously) to understand behavior in complex, contentious, and unfolding political contexts.

Bibliography

a Noufal, Waleed Khaled and Tariq Adely. 2018. "Two of the largest factions in Syria's northwest merge, challenge HTS dominance.".
URL: *https://web.archive.org/web/20180723193235/http://syriadirect.org/news/two-of-the-largest-factions-in-syria%E2%80%99s-northwest-merge-challenge-hts-dominance/*

Abboud, Samer Nassif. 2016. *Syria*. Polity.

Abrahms, Max and Justin Conrad. 2017. "The strategic logic of credit claiming: A new theory for anonymous terrorist attacks." *Security Studies* 26(2):279–304.

Aday, Sean, Henry Farrell, Marc Lynch, John Sides, John Kelly and Ethan Zuckerman. 2010. "Blogs and bullets: New media in contentious politics." *United States Institute of Peace* 65:1–31.

Al-Sayed Hussein, Adnan. 2013. "Internationalization of the Syrian Crisis and Settlement Challenges (Arabic).".

Al-Zaraee, Nisreen and Karam Shaar. 2021. "The Economics of Hayat Tahrir al-Sham.".
URL: *https://www.mei.edu/publications/economics-hayat-tahrir-al-sham*

Anderson, Terry, D Randy Garrison, Walter Archer and Liam Rourke. 2000. "Methodological issues in the content analysis of computer conference transcripts.".

Arnon, Daniel, Peter Haschke and Baekkwan Park. 2022. "The Right Accounting of Wrongs: Examining Temporal Changes to Human Rights Monitoring and Reporting." *British Journal of Political Science* pp. 1–20.

Azman, NAB. 2014. "Strategic Communication: ISIL's Race to Dominate the Web." *Counter Terrorist Trends and Analyses* 6(4):9–13.

Bakke, Kristin M, Kathleen Gallagher Cunningham and Lee JM Seymour. 2012. "A plague of initials: Fragmentation, cohesion, and infighting in civil wars." *Perspectives on Politics* 10(2):265–283.

Bang Jensen, Bree, Bernadette Bresee, Sarah K Dreier, Ramin Farrokhi, Emily K Gade, Willa Jeffers, Marcella H Morris, Charitra S Pabbaraju, Kayla Salehian, Ava Sharifi et al. 2023. "The Lab as a Classroom: Advancing Faculty Research Through Undergraduate Experiential Education." *PS: Political Science & Politics* 56(4):455–462.

Barnard, Anne and Michael R. Gordon. 2017. "Worst Chemical Attack in Years in Syria; U.S. Blames Assad.".
URL: *https://www.nytimes.com/2017/04/04/world/middleeast/syria-gas-attack.html*

Baum, Matthew A and Yuri M Zhukov. 2015. "Filtering revolution: Reporting bias in international newspaper coverage of the Libyan civil war." *Journal of Peace Research* 52(3):384–400.

BBC. 2017. "Syria chemical 'attack': What we know.".
URL: *https://www.bbc.com/news/world-middle-east-39500947*

Becker, Megan. 2020. "Importing the laboratory model to the social sciences: Prospects for improving mentoring of undergraduate researchers." *Journal of Political Science Education* 16(2):212–224.

Becker, Megan, Benjamin AT Graham and Kelebogile Zvobgo. 2021. "The Stewardship Model: An Inclusive Approach to Undergraduate Research." *PS: Political Science & Politics* 54(1):158–162.

Behlendorf, Brandon, Jyoti Belur and Sumit Kumar. 2016. "Peering through the kaleidoscope: Variation and validity in data collection on terrorist attacks." *Studies in Conflict & Terrorism* 39(7-8):641–667.

Bellal, Annyssa and Louise Doswald-Beck. 2011. "Evaluating the Use of Force During the Arab Spring." *Yearbook of International Humanitarian Law* 14:3–35.

Berman, Asher. 2012. "Syrian rebel use of social media." *Foreign Policy Research Institute* 10.

Bestvater, Sarah E. and Cyanne E. Loyle. 2023. "Messaging and Mobilization: Rebel Groups, Social Media Communication, and Audience Engagement." *Journal of Peace Research* 0(0).

Blair, Christopher W., Erica Chenoweth, Michael C. Horowitz, Evan Perkoski and Philip B. K. Potter. 2021. "Honor Among Thieves: Understanding Rhetorical and Material Cooperation Among Violent Nonstate Actors." *International Organization* pp. 1–40.

Blaker, Lisa. 2015. "The Islamic State's use of online social media." *Military Cyber Affairs* 1(1):4.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre. 2008. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008. Publisher: IOP Publishing.
URL: *http://dx.doi.org/10.1088/1742-5468/2008/10/P10008*

Bloom, Mia. 2005. *Dying to Kill: The Allure of Suicide Terror*. Columbia University Press.

Bolsen, Toby W, Bailey R Fairbanks, Eduardo E Aviles, Reagan G Pritchett, Justin T Kingsland, Kristina M LaPlant, Matthew D Montgomery and Natalie C Rogol. 2019. "Merging Undergraduate Teaching, Graduate Training, and Producing Research: Lessons from Three Collaborative Experiments." *PS: Political Science & Politics* 52(1):117–122.

Braithwaite, Jessica Maves and Kathleen Gallagher Cunningham. 2020. "When organizations rebel: introducing the foundations of rebel group emergence (FORGE) dataset." *International Studies Quarterly* 64(1):183–193.

Brown, Heather, Emily Guskein and Amy Mitchell. 2011. "The Role of Social Media in the Arab Uprisings." *Pew Research Center* . https://www.pewresearch.org/

journalism/2012/11/28/role-social-media-arab-uprisings/#_ftn6.

Cafarella, Jennifer and Genevieve Casagrande. 2015. "Syrian Opposition Guide.".
URL: *http://www.understandingwar.org/sites/default/files/Syrian%20Opposition%20Guide$_0$.pdf*

Cafarella, Jennifer and Genevieve Casagrande. 2016. "Syrian Armed Opposition Power Brokers.".
URL: *http://www.understandingwar.org/sites/default/files/Syrian%20Armed%20Opposition0Powerbrokers$_0$.pd*

Chenoweth, Erica. 2010. "Democratic Competition and Terrorist Activity." *The Journal of Politics* 72(1):16–30.

Chenoweth, Erica, Jonathan Pinckney and Orion Lewis. 2018. "Days of rage: Introducing the NAVCO 3.0 dataset." *Journal of Peace Research* 55(4):524–534.

Christia, Fotini. 2012. *Alliance formation in civil wars*. Cambridge University Press.

Chulov, Martin and Kareem Shaheen. 2017. "Syria chemical weapons attack toll rises to 70 as Russian narrative is dismissed.".
URL: *https://www.theguardian.com/world/2017/apr/04/syria-chemical-attack-idlib-province*

Cohen, Adi. 2016. "The ugliness of volunteering to fight for Isis.".
URL: *https://www.vocativ.com/321997/volunteer-anti-isis-fighters-join-up-on-facebook/index.html*

Colgan, Jeff D. 2019. "American bias in global security studies data." *Journal of Global Security Studies* 4(3):358–371.

Conrad, Justin and Kevin Greene. 2015. "Competition, Differentiation, and the Severity of Terrorist Attacks." *The Journal of Politics* 77(2):546–561.

Couttenier, Mathieu and Raphael Soubeyran. 2015. "A Survey of the Causes of Civil Conflicts: Natural Factors and Economic Conditions 1." *Revue d'économie politique* (6):787–810.

Crenshaw, Martha. 1985. "An Organizational Approach to the Analysis of Political Terrorism." *Orbis-A Journal of World Affairs* 29(3):465–489.

Crenshaw, Martha. 1987. "Theories of Terrorism: Instrumental and Organizational Approaches." *The Journal of Strategic Studies* 10(4):13–31.

Cunningham, Kathleen Gallagher. 2011. "Divide and conquer or divide and concede: How do states respond to internally divided separatists?" *American Political Science Review* 105(2):275–297.

Cunningham, Kathleen Gallagher, Kristin M Bakke and Lee JM Seymour. 2012. "Shirts today, skins tomorrow: Dual contests and the effects of fragmentation in self-determination disputes." *Journal of Conflict Resolution* 56(1):67–93.

Darwish, Housamedden. 2020. "On revolution and civil war in Syria.". Syria Untold, March 18, 2020. Available at https://syriauntold.com/2020/03/18/on-revolution-and-civil-war-in-syria/.

Davenport, Christian. 2009. *Media bias, perspective, and state repression: The Black Panther Party*. Cambridge University Press.

Demarest, Leila and Arnim Langer. 2022. "How events enter (or not) data sets: the pitfalls and guidelines of using newspapers in the study of conflict." *Sociological Methods & Research* 51(2):632–666.

Desmarais, Bruce A, Jeffrey J Harden and Frederick J Boehmke. 2015. "Persistent policy pathways: Inferring diffusion networks in the American states." *American political science review* 109(2):392–406.

Donnay, Karsten, Eric T Dunford, Erin C McGrath, David Backer and David E Cunningham. 2019. "Integrating conflict event data." *Journal of Conflict Resolution* 63(5):1337–1364.

Dorff, Cassy. 2017. "Violence, kinship networks, and political resilience: Evidence from Mexico." *Journal of Peace Research* 54(4):558–573.

Dorff, Cassy, Max Gallop and Shahryar Minhas. 2020. "Networks of violence: Predicting conflict in Nigeria." *The Journal of Politics* 82(2):476–493.

Dorff, Cassy and Shahryar Minhas. 2017. "When do states say uncle? Network dependence and sanction compliance." *International Interactions* 43(4):563–588.

Drakos, Konstantinos and Andreas Gofas. 2006. "The devil you know but are afraid to face: Underreporting bias and its distorting effects on the study of terrorism." *Journal of Conflict Resolution* 50(5):714–735.

Earl, Jennifer, Andrew Martin, John D McCarthy and Sarah A Soule. 2004. "The use of newspaper data in the study of collective action." *Annu. Rev. Sociol.* 30:65–80.

Eck, Kristine. 2012. "In data we trust? A comparison of UCDP GED and ACLED conflict events datasets." *Cooperation and Conflict* 47(1):124–141.

Fariss, Christopher J. 2014. "Respect for human rights has improved over time: Modeling the changing standard of accountability." *American Political Science Review* 108(2):297–318.

Fariss, Christopher J. 2018. "Are things really getting better? How to validate latent variable models of human rights." *British Journal of Political Science* 48(1):275–282.

Findley, Michael G. and Joseph K. Young. 2012. "More Combatant Groups, More Terror?: Empirical Tests of an Outbidding Logic." *Terrorism and Political Violence* 24(5):706–721.

Fortna, Virginia Page. 2015. "Do terrorists win? Rebels' use of terrorism and civil war outcomes." *International Organization* 69(3):519–556.

Gabbay, Michael. 2008. "Mapping the Factional Structure of the Sunni Insurgency in Iraq." *CTC Sentinel* 1(4):10–12.

Gabbay, Michael and Ashley Thirkill-Mackelprang. 2010. Insurgent operational claims and networks. In *APSA 2010 Annual Meeting Paper*.

Gade, Emily Kalah, Michael Gabbay, Mohammed M Hafez and Zane Kelly. 2019. "Networks of cooperation: Rebel alliances in fragmented civil wars." *Journal of Conflict Resolution* 63(9):2071–2097.

Gade, Emily Kalah, Mohammed M Hafez and Michael Gabbay. 2019. "Fratricide in rebel movements: A network analysis of Syrian militant infighting." *Journal of peace research* 56(3):321–335.

Greene, Kevin T and Caleb Lucas. 2020. "Once more, with feeling: using sentiment analysis to improve models of relationships between non-state actors." *International Interactions* 46(1):150–162.

Grinberg, Emanuella and Merieme Arif. 2017. "Islamist rebels claim responsibility for twin blasts in Damascus.".
URL: *https://www.cnn.com/2017/03/12/middleeast/syria-violence/index.html*

Gururangan, Suchin, Dallas Card, Sarah K Drier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer and Noah A Smith. 2022. "Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection." *arXiv preprint arXiv:2201.10474* .

Hafez, Mohammed M. 2017. "Fratricidal Rebels: Ideological Extremity and Warring Factionalism in Civil Wars." *Terrorism and Political Violence* pp. 1–26.

Harmon, Rachel, Daniel Arnon and Baekkwan Park. 2022. "TIP for Tat: Political Bias in Human Trafficking Reporting." *British Journal of Political Science* 52(1):445–455.

Hoffman, Aaron M. 2010. "Voice and silence: Why groups take credit for acts of terror." *Journal of Peace Research* 47(5):615–626.

Human Rights Watch. 2017*a*. "Death by Chemicals: The Syrian Government's Widespread and Systematic Use of Chemical Weapons.".
URL: *https://www.hrw.org/report/2017/05/01/death-chemicals/syrian-governments-widespread-and-systematic-use-chemical-weapons*

Human Rights Watch. 2017*b*. "Syria: Events of 2017.".
URL: *https://www.hrw.org/world-report/2018/country-chapters/syria*

Jones, Benjamin T and Eleonora Mattiacci. 2019. "A manifesto, in 140 characters or fewer: Social media as a tool of rebel diplomacy." *British Journal of Political Science* 49(2):739–761.

Jones, Seth G. 2017. *Waging insurgent warfare: Lessons from the Vietcong to the Islamic state*. Oxford University Press.

Kalyvas, Stathis N. 2003. "The ontology of "political violence": action and identity in civil wars." *Perspectives on politics* 1(3):475–494.

Kalyvas, Stathis N. 2012. "Micro-level studies of violence in civil war: Refining and extending the control-collaboration model." *Terrorism and Political Violence* 24(4):658–668.

Kearns, Erin M, Brendan Conlon and Joseph K Young. 2014. "Lying about terrorism." *Studies in Conflict & Terrorism* 37(5):422–439.

Kinne, Brandon J. 2013. "Network Dynamics and the Evolution of International Cooperation." *American Political Science Review* 107(4):766–785.
URL: *https://www.cambridge.org/core/product/identifier/S0003055413000440/type/journal*$_{a}rticle$

Klausen, Jytte. 2015. "Tweeting the Jihad: Social Media Networks of Western Foreign Fighters in Syria and Iraq." *Studies in Conflict & Terrorism* 38(1):1–22.

Krause, Peter. 2017. *Rebel power: why national movements compete, fight, and win*. Cornell University Press.

Kydd, Andrew H. and Barbara F. Walter. 2006. "The Strategies of Terrorism." *International Security* 31(1):49–80.

Larson, Jennifer M. 2016. "Interethnic conflict and the potential dangers of cross-group ties." *Journal of Peace Research* 53(3):459–471.

Laub, Zachary. 2017. "Who's Who in Syria's Civil War.".
URL: *https://www.cfr.org/backgrounder/whos-who-syrias-civil-war*

Laub, Zachary. 2021. "Syria's Civil War: The Descent Into Horror.". Council on Foreign Relations, March 17, 2021. Available at https://www.cfr.org/article/syrias-civil-war.

Lister, Charles. 2016. "The Free Syrian Army: A Decentralized Insurgent Brand.".
URL: *https://www.brookings.edu/research/the-free-syrian-army-a-decentralized-insurgent-brand/*

Loyle, Cyanne E, Kathleen Gallagher Cunningham, Reyko Huang and Danielle F Jung. 2021. "New Directions in Rebel Governance Research." *Perspectives on Politics* pp. 1–13.

Loyle, Cyanne E and Samuel E Bestvater. 2019. "# rebel: Rebel communication strategies in the age of social media." *Conflict Management and Peace Science* 36(6):570–590.

Manger, Mark S., Mark A. Pickup and Tom A. B. Snijders. 2012. "A Hierarchy of Preferences: A Longitudinal Network Analysis Approach to PTA Formation." *Journal of Conflict Resolution* 56(5):853–878.
URL: *http://journals.sagepub.com/doi/10.1177/0022002712438351*

Maoz, Zeev. 2012. "How network analysis can inform the study of international relations." *Conflict management and peace science* 29(3):247–256.

Maoz, Zeev, Lesley G. Terris, Ranan D. Kuperman and Ilan Talmud. 2007. "What Is the Enemy of My Enemy? Causes and Consequences of Imbalanced International Relations, 1816–2001." *The Journal of Politics* 69(1):100–115.
URL: *https://www.journals.uchicago.edu/doi/10.1111/j.1468-2508.2007.00497.x*

Mapping Militant Organizations. 2021. "Hay'at Tahrir al-Sham.".
URL: *https://cisac.fsi.stanford.edu/mappingmilitants/profiles/hay'-tahrir-al-sham*

Mapping Militant Organizations. 2022. "Ahrar al-Sham.".
URL: *https://cisac.fsi.stanford.edu/mappingmilitants/profiles/ahrar-al-sham*

McHugh, Mary L. 2012. "Interrater reliability: the kappa statistic." *Biochemia medica* 22(3):276–282.

Mearsheimer, John J. 2011. *Why leaders lie: The truth about lying in international politics*. Oxford University Press.

Metternich, Nils W, Cassy Dorff, Max Gallop, Simon Weschle and Michael D Ward. 2013. "Antigovernment networks in civil conflicts: How network structures affect conflictual behavior." *American Journal of Political Science* 57(4):892–911.

Milton, Daniel. 2020. "Truth and lies in the Caliphate: The use of deception in Islamic State propaganda." *Media, War & Conflict* p. 1750635220945734.

Minhas, Shahryar, Cassy Dorff, Max Gallop, Howard Liu, Juan Tellez, Margaret Foster and Michael D Ward. 2021. "Taking dyads seriously." *Political Science Research and Methods* pp. 1–19.

Minhas, Shahryar, Peter D Hoff and Michael D Ward. 2019. "Inferential approaches for network analysis: AMEN for latent factor models." *Political Analysis* 27(2):208–222.

Moussaoui, Rana. 2017. "Jihadists take control of major Syrian city.".
URL: *https://www.yahoo.com/news/jihadists-control-syrias-idlib-*

rebel-pull-monitor-174640264.html?guccounter=1guce$_referrer$ =
$aHR0cHM6Ly9lbi53aWtpcGVkaWEub3JnLwguce_referrer_sig$ =
$AQAAAMAKqv3tn2vO6Bpor0scmGwSiHVYQIDhX9bKM5OCO6Ok1Zb7yq7CMeHXLVfpCoM$
$upTsfbj3Rh8bSGeOXie4uLbEobP3iqbQa5I2pUCJgzRcyrZ76$

Möller, Frida. 2011. *News Reports versus Written Narratives*. Taylor Francis pp. 74–88.

Nemeth, Stephen. 2014. "The Effect of Competition on Terrorist Group Operations." *Journal of Conflict Resolution* 58(2):336–362.

Nieman, Mark David. 2016. "The return on social bonds: Social hierarchy and international conflict." *Journal of Peace Research* 53(5):665–679.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz and Emre Kıcıman. 2019. "Social data: Biases, methodological pitfalls, and ethical boundaries." *Frontiers in Big Data* 2:13.

Ostovar, Afshon and Will McCants. N.d. "Why Syria's Armed Opposition Has Failed to Unify." *CENTER FOR NAVAL ANALYSES ARLINGTON VA*. Forthcoming.

Parkinson, Sarah Elizabeth. 2013. "Organizing rebellion: Rethinking high-risk mobilization and social networks in war." *American Political Science Review* pp. 418–432.

Pearlman, Wendy. 2011. *Violence, nonviolence, and the Palestinian national movement*. Cambridge University Press.

Perry, Tom. 2017. "Death Toll Climbs To 74 In Damascus Bombings Targeting Shi'ites.".
URL: *https://www.huffpost.com/entry/scores-killed-in-damascus-bombings-targeting-shiites*$_{n5}8c4029de4b054a0ea6afb1f$

Petkova, Mariya. 2018. "Where is al-Qaeda in Syria?".
URL: *https://www.aljazeera.com/news/2018/3/1/where-is-al-qaeda-in-syria*

Seib, Philip. 2017. *As Terrorism Evolves*. Cambridge University Press.

Shapiro, Jacob N and Nils B Weidmann. 2015. "Is the phone mightier than the sword? Cellphones and insurgent violence in Iraq." *International Organization* 69(2):247–274.

Shaver, Andrew and Political Violence Lab. 2022. "News Media Reporting Patterns and our Biased Understanding of Global Unrest.".
URL: *https://static1.squarespace.com/static/5a7d0ffaa9db09345b7ef6ba/t/6332067102ef64645272b6b1/1664*

Small, Melvin and J. David Singer. 1982. *Resort to Arms: International and Civil Wars, 1816-1980*. SAGE Publications.

Snell, James. 2018. "New rebel coalition brings fresh threat to Syria's splintering war.".
URL: *https://thearabweekly.com/new-rebel-coalition-brings-fresh-threat-syrias-splintering-war*

Staniland, Paul. 2014. *Networks of rebellion: Explaining insurgent cohesion and collapse*. Cornell University Press.

Stewart, Frances, Douglas Holdstock and Antonio Jarquin. 2002. "Root causes of violent conflict in developing countriesCommentary: Conflict—from causes to prevention?" *bmj* 324(7333):342–345.

The Carter Center. 2018. "Weekly Conflict Summary: February 15-21, 2018.".
URL: *https://www.cartercenter.org/resources/pdfs/peace/conflict*$_r esolution/syria - conflict/weekly - conflict - summary - 2018.02.15 - 21.pdf$

United Nations Human Rights Council. 2017. "Report of the Independent International Commission of Inquiry on the Syrian Arab Republic.".
URL: *https://www.ohchr.org/Documents/HRBodies/HRCouncil/CoISyria/*$A_H RC_3 6_5 5_E N.docx$

Walter, Barbara F. 2017. "The New New civil wars." *Annual Review of Political Science* 20:469–486.

Warner, Jason and Stephanie Lizzo. 2021. "The "Boko Haram Disaggregation Problem" and Comparative Profiles of Factional Violence: Challenges, Impacts, and Solutions in the Study of Africa's Deadliest Terror Group (s)." *Terrorism and Political Violence* pp. 1–30.

Weidmann, Nils B. 2015*a*. "Communication, technology, and political conflict: Introduction to the special issue." *Journal of Peace Research* 52(3):263–268.

Weidmann, Nils B. 2015*b*. "On the accuracy of media-based conflict event data." *Journal of Conflict Resolution* 59(6):1129–1149.

Weidmann, Nils B. 2016. "A closer look at reporting bias in conflict event data." *American Journal of Political Science* 60(1):206–218.

Welch, Tyler. 2018. "Theology, heroism, justice, and fear: An analysis of ISIS propaganda magazines Dabiq and Rumiyah." *Dynamics of Asymmetric Conflict* 11(3):186–198.

Whiteside, Craig. 2020. "Lying to win: The Islamic State media department's role in deception efforts." *The RUSI Journal* 165(1):130–141.

Yacoubian, Mona. 2020. "Syria Timeline: Since the Uprising Against Assad." *United States Instutıte of Peace,[Erişim Tarihi: 10.11. 2020, https://www. usip. org/syria-timelineuprising-against-assad]* .

Zech, Steven T. and Michael Gabbay. 2016. "Social Network Analysis in the Study of Terrorism and Insurgency: From Organization to Politics." *International Studies Review* 18(2):214–243.

Zeitzoff, Thomas. 2017. "How Social Media Is Changing Conflict." *Journal of Conflict Resolution* 61(9):1970–1991.
URL: *https://doi.org/10.1177/0022002717721392*

Öberg, Magnus and Sollenberg Margareta. 2011. *Gathering Conflict Information Using News Sources*. Taylor Francis pp. 47–73.

Appendix:

Supporting Information for "Re-imaging data collection: the promise and pitfalls of using non-state actor media for large-scale event datasets"

Contents

A Overview

A.1 Background on Syrian Conflict

The Syrian conflict arose from a combination of political, social, and economic factors, rooted in a pre-existing lack of institutions, which are common preconditions to civil war (Couttenier and Soubeyran 2015, Stewart, Holdstock and Jarquin 2002). The Syrian Civil War, particularly, was sparked by the wave of Arab Spring protests that spanned the region. The initial spark that instigated the Arab Spring movement was attributed to Mohammed Bouazizi, a street vendor from Tunisia who committed suicide by self-immolation in protest against the poor economic and political conditions in Tunisia. This act of resistance in 2010 culminated in a wave of uprisings throughout the region, with the help of social media which was essential in "communicating to the rest of the world what was happening on the ground during the uprisings" (Brown, Guskein and Mitchell 2011). In February 2011, Syria's southern city of Dara'a began to experience protests (Abboud 2016).

A common phrase throughout the Arab Spring was "The people want the downfall of the regime." This slogan played a significant role in mobilizing protesters within Syria, after a group of 15 teenage boys spray-painted it on a Dara'a school wall, and were subsequently arrested and tortured. Protesters took to the streets to demand the boys' release, but in the following weeks, their protest demands shifted. As demonstrations spread to many of Syria's cities, protesters called for regime change, the implementation of economic and political reforms, and the release of all political prisoners. Protesters were met with repression and "hollow" reforms from the state (Abboud 2016, Laub 2017).

Like his regional peers, Syrian President Bashar al-Assad's regime's response quickly became militarized; dissenters faced arrest, extreme force (including tear gas and live rounds), and torture, much of which was indiscriminate and directed at unarmed civilians (Bellal and Doswald-Beck 2011). By April 2011, Syrian security forces had killed over 100 protesters and the Syrian army had begun assaults on the cities of Dara'a and Homs (Abboud 2016, Lister 2016).

The Assad Regime's violent crackdown on protesters spurred the organized defection of many of its own Syrian Security Forces, fueling the beginning of the insurgency throughout the summer and fall of 2011. As a result, the Free Syrian Army (FSA) emerged with the

initial goals of countering Assad's regime militarily and protecting peaceful protesters. Despite achieving some success during the early years of the conflict, FSA never managed to centralize its control over its various member brigades and was considered essentially obsolete by 2014 (Laub 2017, Lister 2016).

The introduction of international forces into the Syrian conflict was an "almost immediate aftereffect" to the 2011 protests and the Assad regime's response. Regional and foreign powers as well as non-state actors played a role in the conflict, altering the nature of the conflict and calling into question its popular label as a "civil war" (Abboud 2016, Al-Sayed Hussein 2013, Darwish 2020). Attempting to apply Small and Singer (1982)'s definition of. a "civil war" to the case of Syria, Darwish (2020) found the conflict to be a "proxy war, an international war, an organised slaughter as well as a massacre," simultaneously.

The Islamic State's entry into the conflict and successful seizure of nearly a third of Syrian territory, alongside the growth of other extremist militant groups, complicated the focus of the war and provoked some degree of greater international interference. Concurrently, the threat of the Islamic State framed Western powers' perspective on the conflict and lent legitimacy to the Assad regime as a "proxy partner in the fight against ISIS" (Abboud 2016, Yacoubian 2020).

Beginning in 2015 and 2016, Russia, Iran, and Hezbollah expanded their support of Syria (Laub 2017, Yacoubian 2020). Since engaging its air force in September 2015, Russia has been a supporter of Assad's regime. It has supplied arms, blocked anti-regime actions at international fora, and continued to trade consistently with Syria. Both Iran and Hezbollah deployed military support to Syria, while Iran also provided military training and counsel and billions of dollars in support of the regime. (Abboud 2016, Laub 2017, Yacoubian 2020). Russia, Iran, and Hezbollah's increased involvement in support of the Syrian government, in combination with the United States and other coalition forces trading their previous sponsorship of moderate rebel groups for strategies that focused on eliminating the Islamic State, all but ensured the Assad regime's resurgence.(Yacoubian 2020)

Between mid-2016 and 2018, the Syrian Government Forces recovered and re-established control over the majority of rebel-held territory. Despite Assad's seeming victory, particular Syrian regions still remain under rebel control, acting as a reminder of a potential return to conflict. Moreover, even while the degree of violence has declined, Syria has been devastated by the war and will continue to face consequences in the years to come (Laub 2021).

While the Syrian civil war may seem unique in its scale of destruction, the preconditions leading up to war, rebel groups' fight for popular support from the local population, and foreign intervention are consistent in civil conflict. However, Syria's conflict stands out in one more way: the sheer amount of real-time primary data in a conflict with non-state actors. Social media will continue to be a data source in conflicts, underscoring the urgency to understand when and how it can be used appropriately. The Syrian civil war provides an opportunity to do just that.

A.2 Background on the UCDP Dataset

The Uppsala Conflict Data Program (UCDP) is the world's oldest collection of civil-war related data. UCDP runs string searches through the Dow Jones news aggregator to identify all articles which involve death or injury for each particular conflict and involves some notable differences

in coding protocol. This aggregation produces a pool of roughly 50,000 news articles which are processed by a team of human coders. UCDP triangulates this data collection process through the examination of reports by NGO and international organizations. SAGAA and the UCDP data differ in terms of the number and names of groups tracked because UCDP annually updates their list of groups based on academic publications, and government and IO reports. We derived our additional list of groups from previous academic work on the Syrian Civil War. When our coders encountered a new name within the social media claims, we worked to determine whether this was a new group, or an alias for an existing group, and modestly updated our codebook accordingly. If a leader or prominent figure announced that they were separating from the main group, we recorded it as a splinter. In contrast, UCDP has a threshold of causing 25 battle-related deaths as well as a declared disagreement with the government over territory for determining when a splinter constitutes a new group.

A.3 Qualitative Check: Further Information

This appendix consists of additional information pertaining to the case and source selection, methodology, and findings presented in Section 4.2. This appendix also contains more detailed descriptions of each event examined as well as includes citations for the news and rebel media sources consulted.

A.3.1 Case Selection: Hay'at Tahrir al-Sham

Hayat Tahrir al-Sham (HTS) is a rebel group that fiercely rose to power in the Syrian Conflict, and whose Salvation Government, based in Idlib, was recognized as the largest opposition government in Syria. The group's origins can be traced to one of the war's earliest insurgent organizations, Al-Nusra Front (ANF), which briefly rebranded as Jabhat Fatah al-Sham in 2016, before entering into a merger with several other groups and officially re-establishing itself as HTS the following year (Mapping Militant Organizations 2021). Due to their later emergence in the war, our dataset only captures HTS-produced media during their first two years of existence, between 2017 and 2018. Nevertheless, HTS boasts a prolific social media presence and a sophisticated network of communications administrations. HTS disseminated more self-produced media in the first 2 years after their formation than their predecessors had over the course of their existence. This, in combination with HTS's domination of Idlib and inclusion of numerous member groups and mergers, makes the organization an ideal and representative case for this exercise.

A.3.2 News Outlet Source Selection

For the purpose of this qualitative investigation, we set out to collect a diverse set of news media sources, both major western networks as well as reports from smaller regional outlets. Nevertheless, this goal came second to simply finding relevant news reports that featured combat events which HTS had experienced. Ultimately, however, we did not have to compromise on either of our aims, as we found that our search for these reports naturally produced an array of different news sources. The sources consulted for this exercise range from well-known major media companies, such as the New York Times and BBC, to smaller broadcasters located

closer to the conflict, like the Arab Weekly and Syria Direct.

A.3.3 Methodology and Event Selection

A critical part of this examination was researching the coverage of HTS in the news media and selecting the 4 major events which we compared to their own rebel-produced media. First we employed general searches across think tanks and search engines and combed through HTS profile reports and almost-encyclopedic pieces to determine how to refine our search and what events further investigate. Our selection criteria included the following: the event must have occured between January 2017 and January 2018 as that is the time period for which we have corresponding accounts of HTS-produced media; the event must have been experienced by HTS, in such a way that they would have been considered a main actor in the event; the event must have been combat-related, as opposed to political, social, etc., as our dataset only captures claims of attack from rebel-media. Due to the restrictive parameters very few cases even qualified as potential events for comparison. We attribute this to the fact that many news organizations often do not report on the actions of individual rebel groups, and instead clump many together under the category of 'oppositional rebel groups'. Thus, the four events chosen were the only ones found within the selective criteria which were sufficiently covered by the news media to be appropriate for this investigation.

A.3.4 Detailed Event Descriptions

Event 1

Just months after their formation, HTS came under international scrutiny for targeting a Damascus Shi'ite pilgrimage site. Various media outlets alleged HTS's responsibility for the March 11th bombing (Grinberg and Arif 2017, Human Rights Watch 2017*a*, Mapping Militant Organizations 2021, Perry 2017), yet their self-authored tweets tell a different story. While initially HTS claimed "responsibility for twin attacks in Damascus targeting 'Iranian militias' and National Guard forces," three days later they denied "any connection to the Damascus blasts," claiming they targeted "only the security branches and military barracks of the criminal regime and its allies." Considering that even the OHCHR connected the incident to HTS (United Nations Human Rights Council 2017), it's unlikely their involvement was fabricated. This event exemplifies the incentives that rebel organizations often have to lie about targeting civilians, especially since HTS didn't deny responsibility until after experiencing backlash.

The following claims, which come from our dataset, provided the evidence used above:

At 1427 GMT on 12 March, Twitter user 'Abu-Hamzah" (@SwedSwed0209) tweeted an embedded image of an Arabic-language statement by the Tahrir al Sham Corps titled "Double Operation Against the Iranian Militias in Downtown Damascus," claiming responsibility for twin attacks in Damascus targeting "Iranian militias" and National Guard forces.

At 1813 GMT on 15 March, Twitter user "Al-Qa'qa' #Syria" (@4ixi1) tweeted: "Breaking – copied from the #Tahrir al-Sham page on Telegram: They are denying any connection to the Damascus blasts," along with an embedded image of a "notice" bearing the name of the Tahrir al-Sham Corps in which it denies any connection to the explosions and says "it targets only the security branches and military barracks of the criminal regime and its allies."

Event 2

On April 4th 2017, the Syrian Government Forces (SGF) launched the "deadliest chemical attack in Syria since the August 2013 assault" (Barnard and Gordon 2017). It drew attention from numerous media sources, who focused on describing the assault and the victims' suffering, providing casualty estimates, and detailing international responses and criticisms (BBC 2017, Chulov and Shaheen 2017, Human Rights Watch 2017*b*). Likewise, HTS disseminated accounts of the event, also choosing to chronicle the experiences of those harmed. They showcased political reactions as well, highlighting "a protest in the city of Khan Shaykhun in the Idlib countryside denouncing the chemical massacre committed by the criminal regime". The HTS and news narratives are remarkably similar because the group has no incentive to conceal the attack and even benefits from the negative coverage of their adversaries. The only observed difference in HTS reporting was their focus on regional as opposed to international political backlash, which may indicate their target audience is local.

The following claims, which come from our dataset, provided the evidence used above:

At 1226 GMT: A tweet including an embedded 2-minute 56-second video bearing the Iba' News Agency logo and a link to the same. The video, titled "Dozens of Martyrs and Casualties After Regime Warplanes Targeted Khan Shaykhun Town With Poisonous Gas," features rescue workers clearing casualties, mostly children, from an alleged explosion site; some appear lifeless, others are experiencing convulsions. A White Helmets worker explains that "more than 150 civilians suffered from suffocation, convulsions, and frothing at the mouth." A medical worker states that "the criminal regime's warplanes bombed Khan Shaykhun with bombs containing poisonous sarin gas." Footage depicts medical personnel offering care to affected individuals in a facility as a camera pans over dead children.

On 7 April, Twitter user 'Wikalat Iba' al-Ikhbariyah" (@Agency 1News) official account of the Tahrir al-Sham Corps' Iba' News Agency, began tweeting several messages as follows: At 1913 GMT: A tweet including two embedded images bearing the Iba' News Agency logo, depicting men and children marching and holding signs, captioned: "A protest in the city of Khan Shaykhun in the Idlib countryside denouncing the chemical massacre committed by the criminal regime."

Event 3

Next, we focused on military clashes that arose between HTS and Ahrar al-Sham (ASIM), throughout the Summer 2017 and culminated in HTS victory, by way of a ceasefire, on July 23rd. Both the news outlets (Al-Zaraee and Shaar 2021, Mapping Militant Organizations 2022, Moussaoui 2017) and HTS-produced media documented the ceasefire and reallocation of territory previously under ASIM control as well as framed the event as an HTS triumph. By accurately reporting on their triumph over ASIM, they had the opportunity to portray themselves as military victors. Therefore, in this instance, HTS had an incentive to be transparent.

The following claims, which come from our dataset, provided the evidence used above:

At 0746 GMT on 21 July, Twitter user "Short-News" (@StoShort) tweeted two embedded screenshots of Iba News Agency Telegram account posts that read: "#Iba'Agency — #Breaking: The #Ahraral-Sham Movement burns Al Ma'mal al Azraq, which contains the largest warehouse of rescue packages near Bab al Hawa before withdrawing" and "#Tahriral-Sham taking control of it." "#Iba'Agency — #Breaking: #Tahriral-Sham Corps takes control of the 106th region overlooking the Bab al Hawa crossing after #Ahraral-Sham fighters withdrew from it."

Titter user "Azmiray al-Amawi" (@hmooooodi 12) tweeted an embedded screenshot of the following Tahrir al-Sham Corps Telegram account post: "An agreement has been reached between brothers in #Tahriral-Sham Corps and #Ahrar al-Sham to cease fire, release all those detained by both parties and [arrange] the exit of all factions from the Bab al Hawa crossing, handing it over to a civil administration." It is worthy of note that at 1636 GMT, Muhammad Abu-Zayd, official spokesman of the Ahrar al-Sham Movement posted a "press release" bearing the group's logo on his Twitter account (@AhrarSpokes Man) a "press release" with an identical message.

At 2017 GMT on 23 July, Twitter user "Ahmad Abu-Farhah" (@ahmadabufarha21; user ID 3926183779) tweeted an embedded image of a handwritten statement signed by the leaders of the Ahrar al-Sham Islamic Movement and Tahrir al-Sham Corps updating their agreement of 21 July, on the Bab al Hawa crossing dispute.

At 1506 GMT on 26 July, Twitter user "Abu-Hasan 'Ali al-'Arjani" (@aliahha131) the account of Abu-Hasan al-Kuwayti, an independent cleric and former Tahrir al Sham Corps sharia official, tweeted an embedded image of a handwritten statement signed by "Abu-al-Sa'ad, Ahrar al-Sham Movement" and "Abu-'Ubaydah al-Shami, Tahrir al-Sham Corps," in which both groups "in the Jabal az Zawiyah region and elsewhere" agree to "release detainees on both sides," "withdraw all military troops," "refrain from mobilizing forces," "remove checkpoints that were recently set up," and "end the media escalation."

Event 4

In early 2018, HTS experienced a military decline, after an alliance, known as the Syrian Liberation Front (SLF), successfully attacked HTS strongholds. Various news reports covered the challenge to HTS dominance and their continuous retreat across Aleppo and Idlib (a Noufal and Adely 2018, Petkova 2018, The Carter Center 2018). Yet, HTS's self-authored media was less forthcoming. While HTS publicly conceded the loss of two previously-held towns, they withheld announcing other documented territorial turnovers to SLF, including western Aleppo (Snell 2018). Not only did HTS rarely acknowledge such defeats, but they also often alleged to have reclaimed territory from SLF that had never even been reported as lost in the first place. Additionally, in one of their few statements concerning the territorial concessions, HTS claimed they were "forced to retreat from some regions to preserve civilian life." Nonetheless, the majority of HTS's statements regarding the encounters depicted their own victories and land acquisitions. Altogether these findings highlight HTS's motivation to present themselves in a positive, sympathetic, and triumphant way, even if that comes at the expense of honest reporting.

The following claims, which come from our dataset, provided the evidence used above:

At 2342 listed posting time on 21 February, Ebaa-Agency.com posted a statement bearing the Iba' News Agency logo titled "Ahrar al-Sham and Suqur [al-Sham] Take Part in Skirmishes." The statement cites "Qasurah al-Humsi, a Tahrir al-Sham military commander," as affirming "a prior agreement between the Corps and Ahrar to neutralize the town of Ariha; however, the Movement violated it, encircled the Corps posts, took control of the town, stormed the house of a foreign tighter brother and detained Abu-al-Rabi' al-Jazrawi – as well as 30 brothers." Al Humsi adds that Ahrar al-Sham also took control of "Tarmala and Ma'arrat an Nu'man in the southern Idlib countryside, meanwhile, Suqur al-Sham waged a failed attempt to storm our posts in Khan as Subul." Al-Humsi calls this claimed escalation "provocative."

At 1522 listed posting time on 28 February, Ebaa-Agency.com posted a 2-page statement bearing the Iba' News Agency logo and titled "The Truth About Neutralizing Towns and Hidden Weapons." The statement decries Syria Liberation Front "transgressions" against HTS and says the former did not honor the "neutral status" of towns such as "Batabu, Hazanu, and 'Uqayrabat," but rather went ahead and attacked "civilians" there anyway. It cites "Khalid Idris, a Tahrir al-Sham official and witness of the [neutralization] agreement," as denouncing a "campaign of transgression led by Syria Liberation" that jeopardizes "the lives of residents" and is aimed to terminate "jihad in Syria" in response to "disgraceful conferences" Syria Liberation Front attended, adding HTS was forced to retreat from some regions "to preserve civilian life." Additionally, Idris decries the Syria Liberation Front for Ahrar al-Sham's cache of heavy weapons, allegedly intended to fight the revolution and end jihad" rather than to fight the regime.

B Data collection Protocols

B.1 Data Collection Procedure: Rebel Claims of Attacks

Our team members compiled a list of major rebel groups involved in the conflict based on prior research (Gade et al. 2019, Gade, Hafez and Gabbay 2019, Hafez 2017, Lister 2016), and pulled all documents related to each of these groups from the open source archive. Coders searched the archive for multiple spellings and versions of each group name, documenting these details in a group name disambiguation sheet (see Appendix C) that continued to expand throughout the project as coders identified new variations of rebel group names and previously undocumented battalions or brigades. The group disambiguation sheet served as the actor list referenced throughout the coding process.

Of the 28,000+ posts collected and translated by the US government from Twitter, Facebook, Youtube, and local news sources, some 21,000 constituted claims of attack. Excluded claims include pro-social or governance behavior, commentary, engagement on international events, and retweets/re-posts of other media. Figure 7 details the number of attacks in our dataset by governorate, where governorate placement mirrors a map of Syria, while Figure 2 depicts the targeting portfolio of groups in our data over the course of the conflict.

Coders then extracted variables (Appendix §C) from each claim, including location, date, perpetrator, target, and claim source. The ontology, variable selection, and data collection processes were completed with the motivation of determining the hierarchical, cooperative, and conflictual relationship patterns of the Syrian conflict—widening the range of recorded behaviors as compared with earlier efforts (Gade et al. 2019, Gade, Hafez and Gabbay 2019). Coders consulted an ontology with sample claims, frequently asked questions, and accessed in-depth explanations of coding rules and opinions to inform their decisions. Details of coder training protocol and intercoder reliability assessments are available in Appendix §B.3.

B.2 Group Name Disambiguation

To explore the *hierarchical* relationships, we collected information on sub-groups and parent groups: anytime that a group was tactically engaged on the ground, the relationships they list

to other groups (sub-units/battalions, alliances, joint operations rooms, parent groups, tactical collaboration, fighting relationships) were recorded (Appendix §C). Based on this relational data, we created an alphabetized, hierarchical network of all rebel groups mentioned within our dataset; this 'disambiguation sheet' contains all name variations—spelling variations, aliases, transliteration/translation issues, or slang terms—for each rebel group.

Although our analysis began by collecting claims from the 50 most prominent groups in the conflict, those claims contained descriptions of the activities of nearly 400 active militant groups, battalions, brigades, and operation rooms in Syria. In other words, we documented the life cycle of many Syrian rebel groups—how they were born, merged, splintered, and eliminated. This provided us an uncharacteristically intimate view of how rebel groups form, mutate, and adapt within a civil conflict environment. This document was continually updated throughout the coding and data cleaning process. We track all such decisions in Appendix §C.

B.3 Methodological Approach to Coder Training

In addition to our data collection and data structure innovations, we model pedagogical practices that made this effort possible (Bang Jensen et al. 2023). We build on cutting-edge dataset creation processes from FORGE (Braithwaite and Cunningham 2020) and NAVCO (Chenoweth, Pinckney and Lewis 2018), both of which highlight the value of a 'lab model' for dataset production and student experience (Bang Jensen et al. 2023, Becker 2020, Bolsen et al. 2019). We likewise leverage dataset creation to support a 'stewardship model' within the social sciences (Becker, Graham and Zvobgo 2021); decreasing unpaid labor, giving student coders 'ownership' over the data production process—including writing parts of the ontology and providing authorship opportunities for students who made meaningful intellectual contributions to the project. Through providing students with increased responsibility and ownership, we achieved higher rates of intercoder reliability, developed a more robust coding ontology (diversity of thought led to new variables and coding procedures), and collected data across multiple efforts simultaneously as lead coders developed fluency with data collection protocol and methodology.

B.4 Intercoder Reliability

In order to code the volume of data used to construct the SAGAA data-set, it was necessary to enlist multiple coders. The risk of having a large team of coders is that they will interpret coding instructions differently, reducing the accuracy of the data. Researchers seek to understand the extent to which this is happening through the use of inter-coder reliability scores, which measure coder consistency. For the SAGAA dataset, we conducted inter-coder reliability checks, utilizing Kappa scores, to ensure coders were using the same understanding of terms as they made coding decisions. These checks allowed us to identify areas where individual coders needed more training, as well as codebook provisions which required greater clarity or re-thinking across the board. This section briefly details how these inter-coder reliability checks informed our process, as well as the interpretation of Kappa scores.

B.4.1 Kappa Scores

The Cohen's Kappa score was first developed by Jacob Cohen in 1960 to address the challenge of "percent agreement" as a means of assessing inter-coder reliability, as percent agreement rewarded coders for chance guesses (McHugh 2012). Consequently, percent agreement scores can take values ranging from 0-100, while Kappa scores, like many other measures of inter-coder reliability, can have values between -1 and 1. Values below 0 are worse than we would expect given random chance while values above zero are better than random chance. A score of 0 itself indicates the coders performed the same as we would expect under random chance. The below table relays how to interpret Kappa values. Compared to percent agreement scores, lower Kappa scores correspond to relatively high levels of consistency. For instance, a Kappa score of 0.41 corresponds with 71 % consistency between codes, while a Kappa score of 0.80 corresponds with 90 % accuracy across coders.

When researchers use Kappa scores, it is common to set a minimum, or cut-off score for the inclusion of that coder's work within the data. This cut-off should be determined by the researchers based on what is appropriate to their project rather than pre-determined across projects or fields. Fields such as health research where consistency is more important than scope may choose to set higher cut-off scores than is appropriate in political science. The commonly used standard articulated by Anderson et al. (2000) supports a 70-80 % cut off, which corresponds with a Kappa score of 0.41 to 0.60. For the construction of the SAGAA dataset, our cut-off Kappa score was corresponding with accuracy.

Kappa Score	Interpretation	Equivalence to Percent Agreement
values below 0	no agreement	worse than 50 %
0	no agreement	50 %
0.01-0.20	no to slight agreement	51-60 %
0.21-0.40	fair agreement	61-70 %
0.41-0.60	moderate agreement	71-80 %
0.61-0.80	substantial agreement	81-90 %
0.81-1.00	almost perfect agreement	91-100%

Table A1: **Interpreting Kappa Scores**

B.4.2 Using Inter-coder Reliability Checks to Improve Data Quality

We used inter-coder reliability checks both to improve the clarity of our conceptualization of our variables and to identify coders who needed additional training. First, if Kappa scores were low across coders, that often indicated errors in how a variable was conceptualized or communicated. In such instances, we explored discrepancies between coders and adapted our guidance to improve clarity, and then asked coders to re-code the challenging variable using the new guidance. If Kappa scores failed to improve despite several iterations of this, we evaluated whether the variable was possible to collect with adequate accuracy, and sometimes chose to eliminate it from the data-set.

In addition to an in depth on-boarding procedure, all coders received multiple rounds of practice coding to help them understand each variable. Coders who did not achieve an inter-coder reliability score above 0.65 (averaged across variables) after two rounds were provided with additional training from the lead coder, including "live" coding sessions with her, and then were re-assessed. Coders who did not achieve adequate consistency after multiple rounds were reassigned to different aspects of the project.

B.4.3 Kappa scores for key variables

Variable	Kappa score
governorate	0.81
group id	0.72
planned	0.78
joint statement	0.83
unspecified joint statement	0.67
gov collaboration	0.73
affiliates	0.73
parent group	0.71
operations room	0.67
frenemies	0.38
claim source	0.19

Table A2: **Kappa Scores**

B.5 Double Checking

To ensure the integrity of our data, we performed a thorough double-checking process for 17 of the 27 variables within the data frame. These 17 variables included: Governorate, Group Name, Group ID, Planned, Joint Statement, Unspecified Joint Statement, Government Collaboration, Affiliates, Parent Group, Operations Room, Frenemies, Big Bomb, Target Class, Target Class 2, Target Class 3, Target Class 4, and Claim Source.

Construction of the SAGAA dataset was conducted over three years (2017 -2020). Over this time period, our ontology evolved, informed by our observations of the data sources we were coding. Therefore, we determined that double-checking was necessary to ensure that all claims were coded in accordance with the most recent version of the ontology. To determine which variables would be included in the double-checking process, we examined intercoder reliability scores and weighed the relative importance of each variable. We prioritized double-checking variables based on intercoder reliability as well as importance to our research interests. To go about this endeavor, we created the 'Shiny app,' a website where coders accessed the claims to double check and kept track of their progress. The double-checking process also improved the overall coding ontology because it encouraged collaboration and prompted lab-wide meetings to review in-depth examples of claims and variables that consistently had low intercoder reliability.

As an example, for the WME attack variable, claims of attack were considered to be excessively violent if they met two out of three requirements: significantly destructive, high civilian death count, or involved non-conventional weapon usage. The specific coding rules in each main category evolved and improved throughout the double-checking process, as areas of confusion and ambiguity were collaboratively addressed. For instance, initially the "destructive" category necessitated at least 50 militant casualties and the "high civilian death count" category necessitated 25 civilian casualties to be considered excessively violent. Based on their observations of the data and their growing perception of what a "typical" attack looked like, coders pushed for a lower threshold, prompting the rules to be changed to 40 militant casualties for the "destructive" category and 15 civilian casualties for the "high death count" category. After refining our criteria, we reassessed our coders comprehension of the changes with a loosely-structured quiz, featuring numerous examples of claims that qualified/failed to qualify as attacks meeting the WME threshold. Additionally, we also went through each quiz question together as a lab, answering clarifying questions and integrating input from fellow

coders after the fact. This created improved clarity about what we hoped to measure in the WME attack variable and serves as a model for the double checking process more broadly.

C Full Coding Guide

D Syrian Armed Groups, Alliances and Actions Coding Guide

D.1 Overview

This ontologyconsists of the ontology and coding guide which was the foundation for the process of labeling rebel attack claims from social media data. The following has been abridged below to focus on the ontology and definition of terms rather than the process of coding. More details are available upon request. The dataset includes all reported incidents of attacks claimed by militant groups in Syria from January 2012 – December 2018. The dataset features incident-level observations of attacks by militant organizations against state and non-state actors as well as civilians and infrastructure.

D.2 Coding protocol

Coders combed through these documents, working on one rebel group at a time, identifying claims of attack which were specific enough to code. For the purposes of this project, to count as an attack claim, the social media entry must include a perpetrator, or a perpetrator and a target. For example: "ISIS killed a militant yesterday," would be a codeable claim because a perpetrator is identified. While, "A soldier from ISIS was killed yesterday," would not be a codeable claim because only a target is identified. Furthermore, only attacks initiated by militant organizations and not Syrian Government Forces (SGF) were gathered and coded. Exceptions to this rule included if SGF was using weapons of mass effect (WME) in a claim, if insurgents were reportedly fighting back against SGF, or if militants were working in cooperation with SGF.

D.2.1 Translation Variations

We coded information recorded by an open-source intelligence service funded and translated by the U.S. government. For this reason, coders encountered multiple different variants of translated text. This includes different transliterations of the same Arabic names, words, or phrases, as well as direct translations or inclusions of slang terms. Coders were instructed to keep this in mind at all times while coding claims, especially when looking for and/or creating entries in the Group Disambiguation Sheet (our primary document for untangling militant relationships, name changes, and hierarchies).

D.3 Core variables for event dataset; basis of network and disambiguation/aliases map

Below, we describe the core variables we collected from each claim of attack.

- ***Group Name***—Identification of the militant group that initiated the attack using the the formal name of the group as listed in the Group Disambiguation. If multiple groups joined together to initiate an attack, all names were put within this column separated by semicolons.

- ***Group ID***—Identification of the militant group that initiated the attack using the acronym created for the group in the Group Disambiguation. If multiple groups joined together to initiate an attack, all IDs were put within this column separated by semicolons.

- ***Joint Statement***—When two or more militant groups participated in tactical collaboration (regardless of whether they claimed to be the initiators or targets of an attack), it was noted within this column. An explicit reference (not implied) of tactical cooperation was required. Some examples of language that indicated a Joint Statement were: "in a joint operation with..", "in assisting the...", or "in collaboration with...".

- ***Unspecified Joint Statement***—Because of the potential incentive to avoid explicitly naming groups with which a given group did not wish to be affiliated, we collected information on 'unspecified' collaboration. The existence of such a collaboration was noted in this column. Examples included: "collaboration with the battalions of the area", "collaboration with other factions", "fighting alongside a group of the city", etc.

- ***Affiliates***—Claims often mentioned an "affiliate of a group" but did not make clear whether these "affiliates" were smaller tactical brigades that were part of the same overarching group or bigger independent organizations that were simply working together. Thus, we tied this variable to the word "affiliate" to increase the accuracy of our defined relationships between parent groups and sub-groups. We believe that in most cases, affiliates should be considered sub-groups or units of the larger organization.

- ***Parent Group***—A militant group that contains smaller brigades or battalions under its title becomes a parent group, while the smaller militias become sub-groups of the parent group. It is possible for a sub-group to have multiple parent groups, outlining the vertical hierarchy of the militant organization. For example, "**x** battalion, belonging to **y** brigade, of the **z** front" defines **x** as the sub-group of **y**, and **y** as the sub-group of **z**, making **y** and **z** parent groups (with **z** at the top of the chain).

- ***Government Collaboration***—If the Assad regime was carrying out an attack or was being attacked alongside a militant group, it was counted as an instance of tactical collaboration between rebels and the government. We did not count foreign governments (such as Russia or Turkey) working with militants as government collaboration, only instances involving the Syrian state.

- ***Target entity***—Identified the militant group that was targeted in the attack using the Group IDs created within the Group Disambiguation. If multiple militant groups were attacked, the first encountered target was put in Target Entity and every subsequent target was put in the following numbered Target Entity columns.

- ***Target Language***—The language found within the claim used to determine the identity of the targeted militant group was entered here.

- ***Operation Rooms***—Captured any mentions of operation rooms (a formal, tactical cooperation structure unique to the Syrian civil war) within a claim. Often times, all independent rebel groups that were members of the operation rooms were listed in a claim.
- ***Deaths:***[13]—If the claim mentioned a number of deaths, it would be entered here. To ensure certainty, if the claim lumped deaths and wounded numbers together, nothing was entered into this variable. Furthermore, if the language was vague, the lowest possible number would be entered. For example, "dozens killed" or "at least twenty killed" would be entered in as "12" or "20".
- ***Property***—This column includes explicit targeting of non-military property such as schools, hospitals, refugee camps, markets, mosques, graveyards, cultural heritage sites, apartment buildings, etc. Attacks on governmental units or rebel camps, barracks, and checkpoints did not count as property.
- ***Weapon***—This variable could take one of eight values to identify the type of weapon or method used by the rebel group to carry out an attack. These values were: gun, sniper, IED, missile, suicide, execution, punishment, and arrest. If multiple weapons were used, they were to be listed out in a consecutive, numbered Weapon columns.
- ***Claim Source***— Taking on one of six values, this variable was used to document the source in which the information was coming from. These values include: direct quote (first-hand), indirect quote (second-hand), general news, another militant group, the parent group, or the sub group, based on who is credited with reporting the activities mentioned in the claim.

D.4 Coding Claims

D.4.1 Example of a Typical Claim

"On 17 August, 2017, the official Twitter account of Fath al-Sham Front tweeted: "Yesterday, we killed dozens of regime soldiers with our rockets and guns alongside Jaysh al-Fath in Aleppo. The local battalions of the area helped us in our quest. Although the regime received assistance from Kharijites [derogatory term for ISIL] we still defeated them."

Claim Components:

- governorate: Aleppo
- year: 2017
- source_date: $8/17/2017$

[13]While we choose to include this variable as one indicator of attack severity that future researchers could use to in triangulation with other sources, we have concerns about its reliability. Accurate information on the number of those killed was likely difficult to obtain during the chaos of battle, especially considered many claims are recorded by the attackers themselves. Additionally, militants had incentives to overstate the impact of their attack and claim a higher death count than was real in order to bolster their reputation as powerful and influential. Thus, we suggest using this variable only in very coarse terms and with significant skepticism. We have not been able to externally validate exact deaths estimates.

- actual_date: 8/16/2017
- group_name: Fath al-Sham Front; Jaysh al-Fath
- group_id: FSF; JAF
- joint_statement: yes
- unspecified_jointstatement: yes
- affiliates: no
- gov_collab: yes
- frenemies: no
- WME: no
- target_class: SG
- target_class2: ISIL
- target_language: regime soldiers; Kharijites
- operations_room: no
- planned: no
- deaths: 12
- weapon: missile
- weapon2: gun
- claim_source: direct_quote

D.4.2 Which claims were excluded?

We collected and coded all claims of attacks made by a seed list of 44 militant groups drawn from Gade et al. (2019), Gade, Hafez and Gabbay (2019). We excluded posts that did not include claims of attacks for the purposes of this dataset. However, some claims of attack were insufficiently detailed. Of these, we excluded claims when the following was true:

- There is no information about the perpetrator and target. "Militant groups attacked a checkpoint on May 8th." ← No info on perpetrator, no info on target.
- There is no information about the perpetrator. "A soldier from ISIS was killed by a landmine yesterday." ← We do not know who planted the landmine, so we do not code this claim.

- SGF (the regime) is the initiator (the group that takes the group_id and group_name columns) of an attack. However, exceptions include: if insurgents are fighting back, if there is a gov_collab, or for WMEs.
- Examples:
  - "Today, the regime captured 12 HTS soldiers." This claim should NOT be coded. We do not have social media data for SGF initiated incidents and we have tried to include only group reporting on their own behavior and direct engagements, not on the behavior of other groups or actors in incidents in which they were not involved.
  - "HTS soldiers attack regime forces in retaliation for capturing HTS soldiers." This claim should be coded.
  - "HTS attacked both ISIL and SGF." This claim should be coded.
  - "The regime used chemical weapons." This claim should be coded.

When a claim was not coded for any multitude of reasons, our protocol was to write "skip" in the notes column of that claim so that it was apparent that it was purposefully not coded.

D.4.3 When should a coder use "UBI" (unbranded individual) vs. leave the target blank?

The "UBI" acronym was only used in claims where there was an explicit mention of an individual, however the vagueness of the language still makes said individual impossible to identify. Below examples of this are provided.
"Today, ISIS attacked forces at a checkpoint with guns in Aleppo."

- group_id: ISIS
- target_class: UBI
- weapon: guns
- governorate: Aleppo

"Today, ISIS conducted an attack at a checkpoint with guns in Aleppo."

- group_id ISIS
- target_class: BLANK
- weapon: guns
- governorate: Aleppo

"Today, ISIS conducted an attack."
This claim would not be coded as it is insufficiently.

D.4.4 claim_number (numeric)

Each claim in our data-set has a unique claim_number that was be pre-populated into the worksheets and recorded in the coding process. No two claims have the same claim_number. Claim numbers help us to 1) reference claims and 2) expand claims.

Notes: For example, if the content of one claim is relevant to understanding another, our coders were instructed to include the claim_numbers in the notes columns. Additionally, if two claims were absolutely identical, our coders only coded the first encountered claim and then wrote the claim_number in the second_claim's notes column.

Instructions for Coding: In this case, two different claims are occurring. The first one being ISIS versus SGF (the regime) and the second one being ISIS versus HTS (another rebel group).

Example of a Composite Claim: At 1145 GMT on October 12, Twitter user "high stakes" posted: Last Wednesday, ISIS soldiers used a missile to successfully kill 15 regime soldiers. Yesterday, they killed 2 HTS militia using a machine gun.

In this case, two different claims are occurring. The first one being ISIS versus SGF (the regime) and the second one being ISIS versus HTS (another rebel group).

For our project, each row within our worksheets only includes a singular attack incident. Thus, composite claims were be expanded, i.e. separated from one row into multiple rows. Therefore, in such a situation our coder would:

- Create a new row in their worksheet.
- Copy and paste the second attack incident into the new row's claim_full_text.
- Copy and paste the original claim's document_title, year, and source_date into the new row.
- Proceed to fill in the other variables with information strictly relevant to the new claim_full_text.
- Create a claim_number for new claims by taking the original claim_number and adding decimals afterward. For example, if the original claim is number 100, derived claims would be 100.1, 100.2, 100.3, etc.
- End the original claim with an ellipses (...) and start derived new claims with an ellipse

D.4.5 Group Disambiguation Sheet (not a variable)

Whenever a coder came encountered a new militant group, operations room, etc. in their worksheet, they were required to create a new entry within the Group Disambiguation Sheet. There, they copied the claim_number of the row in which they found the new militant group, operations room, etc. into the entry made within the Group Disambiguation Sheet under "Found Where?". More information on the Group Disambiguation sheet can be found below.

D.4.6 governorate (categorical)

This variable captures the province in Syria where the attack incident reported in the claim occurred. The response could only take the form of one of the fourteen governorates in Syria. Note that coders often came across alternative spellings of the governorate (i.e. Dara'a instead of Daraa or Al-Hasakah instead of Hasaka) but they were instructed to record their responses using the spelling provided in the following list:

- Aleppo
- Raqqa
- As-Suwayda
- Damascus
- Daraa
- Deir ez-Zor
- Hama
- Hasaka
- Homs
- Idlib
- Latakia
- Quneitra
- Rif Dimashq
- Tartus

Governorates & Cities: When the claim provided a city or town name, but not a governorate, coders turned to the location disambiguation document. Here, they could find a list of cities and towns under their residing governorates. If there wasn't an entry within that document, the coder conducted a Google search to locate which governorate that the city or place belongs to. For example, if the claim said an attack took place in "Yalda", they would look up "Yalda, Syria" into the Google search bar and find that it is located in the "Rif Dimashq" governorate. Then, the coder would add the city as a new entry into the appropriate governorate section within the Governorates Cities document. In this case, it would be adding the word "Yalda" at the end of the list within the "Rif Dimashq" table.
Notes: Wilayah is a word that means "region" and thus, coders were instructed not to include this word in their Google searches. Additionally, many Syrian towns' names may be similar to the names of towns in Iraq, so we also instructed coders to be aware of this and cautious to include 'Syria' in the search bar when looking for the governorate's name.

Multiple Governorates: When coders came across a claim that took place in more than one governorate, they entered the name of each governorate and separated the entries with a semi-colon.
Right at the Border: Coders were instructed to not code attacks that took place in other countries unless the claim takes place right at the border of another country and Syria. When this was the case, the governorate variable was left blank but the name of the other country was recorded in the notes column.
Example and Instructions for Coding: Example Claim: On 22 October, HTS claimed to have fired four Grad rockets at the whereabouts of Hezbollah in El Hermel in Lebanon, scoring a hit.

El Hermel is a city close to the border of Syria. When coding Hezbollah claims, these close encounters are common. In this case, the governorate is left blank but Lebanon is written in your notes column.

D.4.7 year (numeric)

In this cell, year that the claim took place, in the YYYY format, was entered. Coders referred to the source_date and actual_date to find the year of the claim and copied it into this column.

D.4.8 source_date (numeric)

In this cell, the date that the claim was posted, in the MM/DD/YYYY format, was entered. Coders referred to the the claim_full_text to find this information. If the claim_full_text did not provide information as to when the claim was posted, coders looked at the document_title.

Range of Dates When a range of dates is provided, coders entered the latest date in the range. For instance, if a claim is described as a "Roundup of Events for 11-18 April 2017", they would enter 04/18/2017 in the source_date cell.

D.4.9 actual_date (numeric)

In this column, coders entered the date that the attack actually occurred according to the text of the claim. This variable took the format of MM/DD/YYYY.

Example Claim and Instructions for Coding: Example Claim: At 1145 GMT on October 12, 2017, Twitter user "high stakes" posted: Yesterday, ISIS soldiers killed 2 HTS militia using a machine gun.

In this case: source_date is 10/12/2017 but actual_date is 10/11/2017.

When to assume: If the claim does not provide an actual_date, then the protocol is to assume the source_date within this column.

However, when to unspecified dates belong to a claim, since a group could only post about an event after it occurred, as a general rule, earlier dates were always assumed as the actual_date.

Range of Dates: When claims reported on battles or other incidents described as occurring over multiple days and provided a range, our coders recorded the latest date in the range.

D.4.10 group_name (text)

In this column, coders entered the name of the group (as found within the Group Disambiguation Sheet) that initiated the attack. Should the claim_full_text not provide information as to who is initiating an attack, coders were instructed to look at the document_title.

Multiple Initiators: If multiple groups join together to initiate an attack, coders put all collaborating group_names in this column separated by semicolons.

D.4.11 group_id (ID)

In this column, coders entered the ID of the group (as found within the Group Disambiguation Sheet) that initiated the attack. This column should be reflective of any groups entered in the group_name column.
Group Disambiguation Sheet: If a coder encountered a group that was not listed in the Group Disambiguation Sheet, they were required to create a new entry.

D.4.12 claim_full_text (text)

In this cell, coders recorded the entire text of the claim. Because of use permissions from the open source translation service, we are not allowed to release the corpus of claims in its entirety. However, if scholars or practitioners have access to US Government open source information already, we can release that data to them.

D.4.13 joint_statement (yes/no)

Coders entered "yes" in this cell if two or more militant groups were either collaborating as initiators or as targets. The focus of this variable is explicitly on tactical collaboration between militant groups. Possible language includes: "in a joint operation with..." or "in assisting the..." or "in collaboration with...".
Do I code it as a yes in joint_statement?

- "ISIS, in a joint operation with Hezbollah, attacked HTS forces today." YES—2 militant groups vs 1 militant group
- "HTS attacked ISIS and Hezbollah forces in xxx area today." YES—1 militant group vs 2 militant groups
- "HTS attacked ISIS and regime forces in xxx area today." NO—1 militant group vs 1 militant & 1 regime **Note:** Put a "yes" in gov_collab, though!
- "HTS attacked ISIS, Hezbollah, and regime forces in xxx area today." YES—1 militant group vs 2 militant & 1 regime **Note:** Put a "yes" in gov_collab and a "yes" in joint_statement.

Explicit Collaboration: If different groups posted on the same day about the same referenced military battle but did not mention each other in their individual tweets, we did not assume joint_statement and required an explicit mention of collaboration for this variable. However, in

this case coders were instructed to mark in the notes columns for every claim that they believe could have potential collaboration with one another based on their interpretation.
Joint_statement Exception: For only the groups listed below, an exception comes into play. These groups are fronts and/or alliances, and thus, they are actually not working singularly even if the context of the claim only brings up a single group_id and makes it seem as though they are. Instead, these group_ids are all-encompassing of the multiple rebel groups that are working together. This rule applies to both perpetrators and targets. Should this exception rule be triggered, coders put a "yes" in joint_statement and wrote "joint_statement exception" in the notes column.

- Jaysh al-Fath (JAF)
- Islamic Front (IF)
- Southern Front (SF)
- Syrian Islamic Front (SIF)
- Syrian Liberation Front (SLF)
- Levantine Front (LF)
- Syrian Islamic Liberation Front (SILF)
- Syrian Revolutionaries Front (SRF)
- Fatah Halab (FHOR)

D.4.14 unspecified_joint_statement (yes/no)

Vague mentions of collaboration made by initiating groups, without any specifying details about who they are collaborating with, take a "yes" in unspecified_joint_statement. Examples of this include: "collaboration with the battalions of the area", "collaboration with other factions", etc.
Double Collaboration: Our coders did not inherently put a "yes" in joint_statement when there was a "yes" in unspecified_joint_statement. They only put a "yes" in both columns if the context fit the requirements of each column.

An example of this includes: "HTS, alongside ANF, used heavy weapons against Al-Assad forces today. The local battalions of Aleppo assisted them to victory."

D.4.15 affiliates (yes/no)

We came across many claims that mention an "affiliate of a group", which generally refers to a tactical brigade within a larger militant organization. However, this relationship is typically not written out. Thus, it is unclear whether mentioned "affiliates" should be assumed as tactical brigades that are part of the same overarching group or as independent organizations that are simply working together for one or more military operations.

This column is strictly language-based. Only when the word "affiliate" appeared in the claim and the context matched the above description, was a "yes" put in this column.
Group Disambiguation Sheet: When a coder put a "yes" in this column, they were also instructed to put the name of the affiliate into the Group Disambiguation Sheet next to the group they are supposedly affiliated to.

D.4.16 parent_group (ID)

If a subgroup or affiliate is tactically engaged on the ground and their parent group is mentioned somewhere in the claim, the coder entered the group_id of the parent group in this column. If either parent groups or subgroups do not have IDs, the coders first created them within the Group Disambiguation Sheet and then coded accordingly.
Examples of Subgroups and Parent Groups (and sometimes Affiliates):

- "Jund Al-Haqq Brigades, an offshoot of the Suqur al-Sham Brigades, claimed..."
- "Ahrar al-Sham Islamic Movement, of the Syrian Islamic Front, claimed..."
- "Al-Murabitin Battalions, belonging to the Liwa al-Islam Brigade, claimed..."
- "Al-Ansar Brigade, an affiliate of the Jaysh al-Islam, claimed..."

Subgroup and Parent Group: If both a subgroup and parent group (within the same organization) are tactically engaged on the ground, coders put both groups in the group_name and group_id column. In addition, they wrote out the parent group's ID again in the parent_group column. This also counts as a "yes" in joint_statement.
Group Disambiguation Sheet: Whether it be the existence of a subgroup or a parent group, coders were instructed to ensure that both group entries within the Group Disambiguation Sheet were reflective of the relationship present in the claim and to update the document as needed.
Multiple Parent Groups: If there were multiple parent groups, coders entered them all into the parent_group column separated by semicolons.

D.4.17 gov_collab (yes/no)

Coders entered "yes" in this cell if the regime (SGF) carried out an attack or is being attacked alongside a militant group. The focus of this variable is explicitly on tactical collaboration between militant groups and the regime. Note that foreign governments collaborating with the regime does not count as a "yes" in gov_collab, such as SGF and Russia (RFF) working together.
Does it count as gov-collab?

- "ISIS, in a joint operation with Hezbollah, attacked HTS forces today." NO 2 militant groups vs 1 militant group
- "HTS attacked ISIS and Hezbollah today." NO 1 militant group vs 2 militant groups
- "HTS attacked ISIS and regime forces today." YES 1 militant group vs 1 militant 1 regime

- "HTS attacked ISIS, Hezbollah, and regime forces today." YES 1 militant 1 regime vs 2 militant groups Note: This becomes a situation where you can put a "yes" in gov_collab and a "yes" in joint_statement.

D.4.18 frenemies (yes/no):

Coders entered "yes" in this cell if a group made an explicit claim about a "make-up" or "break-up" with another group. Note that frenemies statements do not need to include specific claims of an attack. We are just trying to capture any explicit statements that Group A is working with Group B or, alternatively, Group A is no longer working with Group B. Thus, coders left blank the variables (WME, target_class, target_language, deaths, property, weapon, target_class2) only if they are not relevant after entering "yes" in frenemies. This is one of the few instances on our project in which this was allowed.

The difference between frenemies and joint_statement: "HTS denounces ISIS and declares a new alliance with ANF." (YES in Frenemies) "ISIS and HEZ joined forces in an attack against HTS." (NO in Frenemies)

Notes: In some cases, the exchanging of prisoners between militant groups, can be marked as frenemies. The exchanging of prisoners between SGF and militant groups should not be marked as frenemies. However, coders were instructed to document this situation in the notes column.

Only if the claim's context is relevant, dismantling, stealing, or gaining weaponry would it be marked as frenemies. Moreover, the context being relevant is uncommon because dismantling, stealing, or gaining weaponry in and of itself was not considered a code-able claim in the first place. If a claim says that one group had "talks" or "met with" another group, that was also treated as frenemies because they are putting forward an effort to fix, maintain, or improve relations.

Troop defection counted as frenemies as well. Coders specified who the troops defected from and went to. However, this only counted if the troop defection was from rebel group to rebel group.

Regime troops defecting was not counted as frenemies, but coders similarly documented these situations the notes columns.

One rebel group re-tweeting the activities of another group was not counted as frenemies unless additional commentary is shared on the relationship between both groups.

The mentioning of an "affiliate" does not inherently mean frenemies.

D.4.19 WME (yes/no)

Coders entered "yes" in this cell if the claim included at least 2 of the following 3 criteria.

Destructive: If the attack resulted in significant damage or was directed at a key leader of a group. A key leader of a group means a well-known character. If a Google search cannot bring up results when searching the name of a leader that was deliberately killed, it is probably not a qualifier for WME. If $\tilde{4}0$ militants or soldiers (this includes the regime) were killed or wounded, this automatically triggers a "yes" in WME. In these cases, coders wrote "high death count" in the notes column.

Civilian: If the attack harmed a number of civilians. Note that a smaller threshold is required for killed civilians relative to the amount of killed targets to qualify. We require 15 civilians to be killed for it to equally and automatically trigger a "yes" in WME. Coders recorded "high death count" in the notes column should this occur.

Non-conventional: If the attack contained any sneaky, non-conventional aspects.
Unconventional weapons include: Death filmed for distribution, chopping off hands, stoning someone, torture, interrogation, taking hostages, kidnapping, suicide bomber, suicide attack, martyrdom-seeking operation, decapitation, execution, public death, chemical warfare, IED, VBIED, car bomb, sticky bomb, landmines.
Conventional weapons include: Guns, sniper, grenades, explosives, missiles, guided missiles, mortar, guided round, RPG, rockets, heavy artillery, heavy weapons, cannon, shelling, tanks, airstrikes, jets, drones, arresting, capturing, Jahannam gun, SPG-9 gun or rocket, anti-tank gun, guns on trucks, missiles on tanks, Dushka machine gun, grad rockets.

In order to trigger WME, 2 out of 3 criteria needed to be fulfilled. Thus, if a suicide attack (a non-conventional weapon) took place but without a described high level of destruction, it did not count as WME since it only fulfilled 1 out of 3 criteria.

Notes: If an international (Russia, US, Iran, etc.) aircraft, jet, or drone was struck from the sky, this did NOT count as WME. However, coders were instructed to record these events in the notes columns.

Airstrikes dis NOT count as WME. If an airstrike kills a key leader, it still did not count as WME simply because we are more concerned with the activities of rebel groups in relation to WME versus international forces.

When in doubt, we instructed our coders to reach out when confused about this column and to over capture this variable rather than not record such events.

D.4.20 target_class (ID)

In this column, coders entered the ID of the group (as found within the Group Disambiguation Sheet) that was targeted in the attack.

Multiple Targets If multiple groups were targeted within an attack, the first encountered target was put into target_class, and every subsequent target in the proceeding target_class columns. This is unlike the group_name and group_id columns that use semicolons to separate multiple entries.

Group Disambiguation Sheet: f a coder encountered a group that was not listed in the Group Disam- biguation Sheet, they were required to create a new entry.

D.4.21 target_class2 (ID)

If more than one group was targeted, coders entered the second group name here.

D.4.22 target_class3 (ID)

If more than two groups were targeted, coders entered the third group name here.

D.4.23 target_class4 (ID)

If more than three groups were targeted, coders entered the fourth group name here.

D.4.24 target_language (text)

In this column, coders entered the language you used to determine the target_class group. For instance, if the "rejectionist army" is being attacked, a look at the Group Disambiguation Sheet informs us that "rejectionist army" stands for SGF. Thus, we would enter "rejectionist army" in the target_language column, but SGF in the target_class column.

Multiple Targets: If the claim mentions multiple targets, coders entered the language used for each of the relevant groups separated by semicolons (even though each target_class gets their own column).

D.4.25 operations_room (yes/no)

Many claims refer to operations rooms - these are formal, cooperative tactical units. Coders put a "yes" in this column if an operations room is present. They were instructed to look up the operations room within the Group Disambiguation Sheet and create a new entry if there was not an ID already assigned to the unit.
Group Disambiguation Sheet: Coders ensured that all mentioned groups within the claim_full_text that are alleged members of the operations room were included in the Group Disambiguation Sheet under "Independent Members". They did not need to create IDs via new entries for these members unless one of them individually initiated an attack against a target.
Unspecified Operations Room: Should a mentioned operations room list out all alleged members but not provide a name for themselves as an overarching organization, coders still put a "yes" in operations_room. Then, they created a code within the "Unspecified Operations Room" part of the Group Disambiguation Sheet.
Joint_statement: If the operations room is tactically engaged on the ground, coders put a "yes" in joint_statement and a "yes" in operations_room. In these instances, the operations' room ID was also recorded in group_name and group_id variables. If the operations room is mentioned but not tactically engaged on the ground, coders put a "no" in joint_statement but a "yes" in operations_room.

D.4.26 planned (yes/no)

We did not want to double-count an attack incident by coding both the preparations for an attack and the actual attack itself. Coders put a "yes" in this column if the claim mentions rebel groups planning their future attack, fortifying their base in preparation for a future attack, or walking/storming/heading towards their future attack. Coders left blank the variables (WME, target_class, target_language, deaths, property, weapon, target_class2) only if they were not relevant after entering "yes" in planned. Similar to frenemies, this is one of the few instances on our project in which this was allowed. Examples of this include:

- "A video produced by Jaysh al-Islam depicts a Jaysh al-Islam force heading in the direction of the Aleppo countryside to respond to the aggression of ISIL."

- "HTS has launched a fortification campaign in the liberated regions on the eastern Daraa fronts and praises the efforts of civilians who have lent a helping hand."
- "Fath al-Sham Front sets out to break Aleppo's siege. A video features tanks and APCs heading to battle, armed fighters boarding an APC as they vow revenge, as well as fighters on foot and tanks in a desert terrain."

D.4.27 deaths (numerical)

If the claim mentions a number of deaths, coders entered it here. If the language is vague ("dozens") or provides a range ("at least twenty"), coders entered the lowest possible number. In this case, it would be "12" or "20". If the claim lumps deaths and wounded together ("at least four killed or wounded"), coders were instructed to leave the column empty. An empty deaths column is not the number 0, but a period.
Note: Unless specific death numbers were explicitly mentioned, our coders were instructed to, under no circumstances, enter any number into the column, even if they were certain that people died in the instance.

Example claim and instructions for coding:

On 22 November, ISIS claimed to have detonated an IED in a large group of HTS forces, killing all of them.

Even though we are certain people died in this case, we cannot put know the lowest possible number of deaths, and do not assume that it is 1. Therefore, this cell would be left blank.

If the claim explicitly says *several* individuals were killed, coders entered "3" for the number of deaths.

Sniping claims need to explicitly state that the victim died, otherwise coders did not count these victims as a deaths column inputs.

D.4.28 property (text)

If the claim mentions the explicit targeting of non-military property, coders entered it here.

Does it count as property?
Yes: Schools, Hospitals, Refugee Camps, Markets, Oilfields, Apartment Buildings, Mosque, Graveyards, Cemeteries, Cultural Heritage Destruction, Setting Fire to Cigarettes, Pornography, or Alcohol
No: Tanks, Machine Guns, Bulldozers, Airports, Military Camps, Barracks, Checkpoints

Multiple Properties: If multiple pieces of property were attacked, coders documented them within this column separated by semicolons.

Property Assumptions: If a place or property was attacked without specific mention of a target, coders left the target_class blank. Exceptions include attacks on refugee camps, markets, hospitals, apartment buildings, and schools which were assumed as CIV.

However, attacks on governmental units, military camps, barracks, checkpoints, or airports were not given the same assumption and the target must be explicitly mentioned to qualify.

In cases of cities or towns, only if it is shelled did a coder assume an attack on CIV. Otherwise, general attacks on cities or towns did not take a target_class assumption.
Notes: Hospitals were always assumed to mean CIV because even military hospitals are considered civilian locations under the laws of war.

D.4.29 weapon (categorical)

If the claim mentions a weapon or method used to carry out the attack, coders entered it here. This column could only take one of nine values. If a mentioned weapon did not fit into one of these categories, coders entered a period in this cell and wrote the mentioned weapon in the notes column instead.

- gun: guns, gunfire, machine gun, automatic weapons, light weapons
- sniper: note that sniping claims need to explicitly state that the victim died, otherwise it does not count as a deaths column input
- IED: explosive, explosives, sticky bomb, roadside bomb, vehicle bomb, landmine
- missile: guided missile, mortar, guided round, RPG, rocket, heavy artillery, heavy weapons, cannon, shelling, tanks
- suicide: suicide bomber, suicide attack, martyrdom-seeking operation
- execution: decapitation, public death, death filmed for distribution
- punishment: chopping off hands, stoning someone, torture, interrogation
- arrest: arresting, capturing, taking hostages, kidnapping
- airstrike: if the claim mentions airstrikes on behalf of international forces such as Russia, US, Israel, Assad, etc

Airstrikes: When airstrikes occurred, coders did not put the international force that was conducting the airstrike as the group_name and group_id. Instead, they just acknowledged the existence of the airstrike in the weapons column.

If only an airstrike was taking place, coders filled in the weapons column and skipped the rest of the variables as they were no longer relevant to the purposes of this project at the moment.

If an airstrike was mentioned in the presence of another attack incident, coders coded the claim as they normally would but with “airstrike” as an input within one of the other weapons columns (weapon2, weapon3, weapon4).
Double Checking: When coders ran into a claim that is strictly focused on an airstrike during the double checking process, they were instructed to write ”skip” in the notes column.

If an airstrike was mentioned in the presence of another attack incident, double-checkers coded the claim strictly based on the attack incident and simply acknowledged the presence of the airstrike by writing “airstrike” in the notes column.

Guns as Missiles: Please find below a list of weapons often described as "guns" but are actually missiles. Each of the weapons below was coded as a missile in the spreadsheet. Not both "gun" and "missile" – just "missile"!

- Jahannam gun
- 23mm gun
- Spg-9 gun
- 57mm gun
- 106mm recoilless rifle
- Dushka machine gun
- 30mm gun (also referred to as 30mm autocannon)
- BMP gun (missile on a tank)
- B-10 recoilless rifle
- 75mm gun
- Anti-tank gun (almost always missiles)

Notes: "Light weapons" typically mean guns. "Heavy weapons" typically mean missiles. "Medium weapons" is where it becomes vague. Coders were instructed to code for "light weapons" and "heavy weapons" but not "medium weapons".

D.4.30 weapon2 (categorical)

If more than one weapon was used, coders entered the second weapon here.

D.4.31 weapon3 (categorical)

If more than two weapons were used, coders entered the third weapon here.

D.4.32 weapon4 (categorical)

If more than three weapons were used, coders entered the fourth weapon here.

D.4.33 claim_source (categorical)

This variable was used to capture who exactly provided the information about the claim. This variable can take one of five values:

- direct_quote: This is used when the claim is coming directly from official group bodies. For instance, an official Twitter or YouTube account for a group or a group's leader is coded into this category. This would also be used if the group releases an official statement/claim about their activity. Another way to think of this is as a primary source document.
  Because Twitter frequently removes accounts belonging to militant groups, the social media handles representing the "official body" of a group often change. Thus, we do not have a way of verifying whether or not various Twitter handles associated with a particular militant group are their official account. Instead, we must assume that the US Government, the body that originally translated our dataset, only pulled tweets from official accounts. Thus, even if you don't recognize a social media handle (the username that either stands alone or follows an "@" symbol), assume it represents the militant group in question and code as a direct_quote.

- indirect_quote This is used when the claim is attributed to or representative of an official group account but is reported second-hand. For instance, if a news organization is reporting on a recent battle and says, "We interviewed Commander X and he said..." this is still information that is coming from the militants' themselves, but it is a secondary source document. Compare this to a news_general claim, in which a news organization is reporting on a recent battle and does not quote anyone at all.

  If someone is retweeting, it counts as an indirect_quote: The following is a retweet by Jaysh al-Islam of a post originally released by Twitter user "Al-Naqib Captain Islam 'Alloush" (https://twitter.com/islamalloush0), who identifies himself as a Jaysh al-Islam spokesman: "Backup is being sent to the Bala front, which is witnessing a dangerous escalation by the regime, in a flagrant violation of the cease-fire," on 8 March.

  Another example of an indirect_quote: At 1612 GMT on 25 May, Twitter user "Khadim al-Ikhwah" (@khadim388225580) tweeted a link to an ISIS statement titled "A Martyrdom-Seeking Operation Strikes a Nusayri Army Gathering in Tadmur."

- news_general: This is used when general news reporting presents a claim. This does not feature direct quotes from militant organizations, but is just a round-up of news presented by a media outlet. For example, reports issued by the Syrian Observatory for Human Rights would be coded as news_general. This variable is meant to capture who reported the claim. For instance, if SOHR quotes an unnamed source in their article, it would not be coded as other. It would still be coded as news_general because SOHR is the one reporting the information.

  If news_general claims do not identify the initiating group or target class and instead talk about an attack in a very broad and neutral fashion, groups were coded in the order that they were encountered between group_id and target_class. Furthermore, the total amount

of deaths were still documented in these claims even when there is not clarity about who the initiating/target group are. However, coders put in the notes column that they did not know who the initiating group was, let alone who the deaths belonged to.

- other_militants: This is used when a militant group is making claims about the activity of a different group. For instance, Group A is alleging that Group B conducted an attack on a town that killed 3 civilians. Oftentimes, this will be used with allies or affiliates since they do not have a clear hierarchical relationship to one another, unlike subgroups and parent groups.

  When it comes to other_militant sourced claims, coders only marked it as frenemies if they said something positive or negative about the group they are reporting on. If it is just a reference to another group that doesn't indicate support or animosity, then that wouldn't take a response in frenemies.

- parent_group: If a subgroup is the only militant group that is tactically engaged on the ground but their parent group is reporting on the incident, coders marked "parent_group" under the claim_source column.

  Even if an affiliate is doing an attack and the parent_group column is filled in with "the other group's" ID, the claim_source would NOT be parent_group. Instead, it would be other_militants, as we are not double-dipping on the purposes of those choices; other_militants represents affiliate and ally relationships, while parent_group represents only parent_group and sub_group relationships.

- sub_group: If a parent group is the only militant group that is tactically engaged on the ground but their sub group is reporting on the incident, coders marked "sub_group" in the claim_source column.

- other: For any claim which does not fit into the above criteria, coders entered "other" in the claim_source column. This may include quoting of "unnamed sources" that have no links to a group or organization. It also includes tweets from users who are reporting on a situation from an activist or civilian perspective.

D.4.34 notes (text)

This column was used to enter any relevant information, including irregularities in the case, difficulties in coding, a rare or exceptional event, etc. When in doubt, coders were instructed to enter information!

Should there be multiple sets of notes within this column from multiple coders, the entries were separated with semicolons. Individual coders marked their particular note with their initials.

Overall, what things in the coding guide required notes?

- If a claim was skipped.
- If another claim is being referenced.
- If a coder is doubtful of their answers and require additional review.

- If a coder needs to explain how they resolved tricky claim elements.
- If a claim is happening close to Syrian external borders.
- If a coder needs to write “joint_statement exception” when the rule is triggered.
- If there is an exchanging of prisoners between SGF and militants.
- If there is troop defections between SGF and militants.
- When chemical warfare is occurring.
- If a drone, airplane, or jet is struck from the sky.
- When a coder doesn’t know who initiated an attack and who the target was.
- If there is governance behavior on behalf of rebel groups occurring.
- Writing “high death count” because of a WME assumption.
- Any groups that a coder believes could be potentially collaborating due to claim text but collaboration is not explicitly mentioned.
- Reason for over-capturing a WME incident.
- When a mentioned weapon is not a provided option.

Double Checking During the double checking process we also added some additional rules to our note section guide.

- If a claim is strictly focused on an airstrike incident, coders wrote ”skip” in the notes column.
- If a claim mentions an airstrike in addition to a claim of attack incident, coders included ”airstrike” in the notes column.
- If the group_name and group_id columns are pre-filled and the claim language makes it impossible to verify, coders noted ”filled in perpetrator” in the notes column.

D.4.35 checked (yes/no/flag)

This column represents whether or not a claim was double-checked. It can take one of three options: “no”, “yes”, and “flag”. Once all the columns of a particular claim were double-checked for accuracy, coders finished off the claim by marking “yes” in this column.
Flag: If coders only double-checked a few columns or most columns of a claim but ran into a concern that required additional screening, they marked “flag” in this column.

D.4.36 coder_id (text)

Coders entered their initials in this column. Multiple coders initials were separated by semi-colons.

D.4.37 checker_id (text)

This only applies to the double-checking Shiny App. Coders that double-checked or reviewed a claim ensured that their initials were put in this column following the original coder's initials.

D.5 Additional Notes

- If a claim mentions "liberating" or "seizing full control" of a city or village and the group that they are liberating the city from is not mentioned, then coders left the target_class blank. "Liberation" in and of itself is not a specific enough term for us to determine a target or attack classification from.
- "Storming" or "paving the way" does not count as a weapon. Furthermore, it does not count as code-able unless the act of "storming" or "paving the way" is specified to include the targets that were hurt in the process.
- If all that is vaguely said is "militants" or "forces", that is not an indication of SGF. However, if all that is vaguely said is "soldier", "army soldier", "captain", "lieutenant", "officer", "serviceman", or "military convoy", those were assumed to be SGF.
- Dismantling, stealing, or gaining weaponry is not a code-able claim. It may potentially be frenemies and is only code-able if stealing weaponry is happening in conjunction with an actual attack.
- Spies or accused spies were not identified by the interests they represent. For example, an Iranian spy who works for the al-Assad regime should be coded as SGC for the target_class, not IRAN or SGF.
- If nationalities are mentioned within vague generalizations like "pro-regime militia containing Iranians, Lebanese, and Iraqis," that was only be coded as SGF.
- If it says Iranian Militia, Iraqi Forces, Russian Forces, etc., coders put the country group code as the target class (IRR, RFF, etc.)
- However, if it doesn't say forces or militia after the mentioned nationality, then coders record the target class as CIV (if the context made sense to do so).

D.6 Group Disambiguation

Whenever two or more militant groups were collaborating tactically as either initiators of an attack or as join targets on the recieving end of an attack, a 'yes' was recorded under the joint statement variable. In contrast, the government collaboration variable was used for cases in which regime forces were either carrying out an attack or being attacked alongside a militant group. In claims where there was both more than one militant group and regime forces tactically engaged on the ground, a 'yes' was coded for both joint statements and government collaboration. Lastly, when a claim mentioned an operations room by name – a formal and cooperative tactical unit – a 'yes' was entered under the operations room variable. Here,

the militant groups that were the intended targets of the attack in each claim were recorded. The "frenemies" variable, which consists of statements of a 'make up' or 'break up' with another militant group, captures the political and social side of relationship-building among rebel groups (rather than capture attack behavior). Qualifying textual references are neither inherently *conflictual* or *cooperative* but instead reflect rebel group's political behavior separate from tactical cooperation. Thus, this variable can reflect the intent to cooperate– or to stop cooperating– but is not bounded by the opportunity or capacity to cooperate the way the variables above are bounded. For instance, if a number of rebel groups on the verge of collapse committed to cooperation, but then lacked the opportunity to do so, we would capture that intent through this variable.

When we encountered an issue as a team—e.g. in which a coder identified that the same group might have been given two different identifying codes due to a transliteration issue, or that two groups with very similar names might have been given the same code, despite being different groups—we did a deep dive into validation of this potential discrepancy, consulted subject matter experts, and pulled all claims related to those group coded to validate whether or not this issue required re-coding. In several cases, we returned to the original dataset and re-coded 80-100 claims that had been mislabeled.

D.7 Covariate Data for Rebel Groups

D.7.1 Data Construction Process

As a corollary to the network construction discussed in the previous section, we collected information about 80 rebel groups that were most central in the Syrian civil war. Specifically, we collected data on the rebel groups with the highest degree of centrality in our cooperative and conflictual networks. Thus, this dataset details which rebel groups were most critical to the Syrian conflict, as measured by network centrality rather than press coverage or perceived capacity.

We coded variables about Syrian rebel groups traits at both the general and year-group level to capture change over time. Key variable information coded in the general row indicates information that could not be found at the year-group level for each militant group. We selected these variables based on the previous literature on factors that shape the microdynanmics of civil war. This covariate dataset will provides insight into potential drivers of rebel group alliance formation, splintering, merging, and overall collaboration.

A detailed coding guide was created in collaboration of all coders assigned to this project. At all times while working the on covariate data collection, they were instructed to have this document open. In addition to providing detailed instructions for approaching each variable, the coding guide also gave coders access to spreadsheet templates and key resources which they were expected to use. For each of the key groups examined, information was organized on a separate spreadsheet, distinct to that organization. Data was collected at both the general and year level for every key group. Thus, when any variables were collected on a group this was always recorded in the general row. Additionally, if that information provided a specific year or range of years, this was recorded in subsequent rows on the spreadsheet, each corresponding to one year in the groups active life-span. Our data collection only focused on the years between 2012 and 2018, although some groups existed outside of this time frame. In order to standardize

our searches across our multiple coders, we put in place a strict procedure. First, coders were instructed to consult a set of reputable resources collected on the coding guide, before moving to google scholar and search engines if they proved fruitless. Second, we implemented the 20 minute rule. What this means is that every coder had to spend at least 20 minutes searching for each variable before they could assume that there was no information available about that variable and move on. Thus, anytime there is a ”no” in a cell, this represents 20 minutes of search time on part of the coder. Additionally, because of the difficulties of gathering this type of information at the year level, we also had to institute a rule for assumed variables. Coders never assumed the onset of events. However, for assumed variables only, if there was evidence that indicated the start of an event or action, we assumed that behavior continued unless specific information indicated that it stopped in a certain year. The following are our assumed variables: State Sponsorship, Resources SS, Non-State Sponsorship, Resources NSS, Alliance, Group Make up, Leader, Rivals, Economic Ideology, Political Ideology, Religious Ideology, and Territorial Ideology.

D.7.2 Core Variables

- ***Operational location***—Any Syrian governorates where the key group was active. Evidence that a group was “active” in a governorate includes controlling territory, conducting attacks, or having a significant presence in the province (examples include a training camp, facility, or safe zone).

- ***State sponsorship***—Foreign governments (but not private individuals within a foreign country) who have been implicated as sponsors for the key group. Sponsorship can take the form of material resources (money, weapons, etc.), training, verbal support, territory, or intelligence. Once evidence was found that a state actor began sponsoring a key group in a given year, it was assumed that they continued to sponsor that key group in subsequent years, unless further research indicated that they stopped at a certain point. State sponsors were denoted in the covariate datasets using the ISO3 codes for the countries. Notes column describes nature of sponsoring relationship.

- ***Non-state sponsorship***—Non-state actors who have been implicated as sponsors for the key group. Non-state sponsors can include but are not limited to, private citizens from other countries, private sector entities, non-governmental organizations, philanthropic organizations, and international rebel groups. Sponsorship can take the form of material resources (money, weapons, etc.), training, verbal support, territory, or intelligence. Once evidence was found that a non-state actor began sponsoring a key group in a given year, it was assumed that they continued to sponsor that key group in subsequent years, unless further research indicated that they stopped at a certain point. Due to the inconsistent nature of non-state sponsorship, textual information was used to chronicle the non-state actors in the covariate datasets.

- ***Resources***—The forms of sponsorship provided by both state and non-state actors to the key groups. Resources were classified into the following categories:

  - ***Material resources***—Sponsorship of the key group through money, guns, weapons,

drugs (for the purpose of being sold to generate funding), or any other kind of financing.

  - ***Training***—Sponsorship through either extending invitations to training camps or providing experts to train the key group.
  - ***Verbal support***—Sponsors issue official and unofficial statements to publicly legitimize and support the groups.
  - ***Territory***—Sponsors provide the key group with operational control of land or bases.
  - ***Intelligence***—Sponsors provide the key group with essential information, such as the objectives and capabilities of it's rivals.
  - ***Other***—Seldom used to characterize any forms of sponsorship that do not fit into the above categories.

- ***Group Size***— Estimates of the key group's size. These estimates can be highly varied from year to year especially when one group absorbs another or experiences internal splintering. In order to account for the uneven availability of size data, particularly for specific years, if low estimates and high estimates were found, even across sources, they were combined to create ranges.

- ***Splinter***— Splintering is defined as the breakaway from an existing group that is publicly announced or established. A splinter could also include two or more groups that merge, then later have one part which breaks off again. When evidence was found that a group splintered off into separate groups this column was marked with a "yes" and the Group IDs from the Group Disambiguation sheet were used to indicate the name of the group or groups that splintered off during a specific year or year range in which it took place.

- ***Joining Battalions***—Factions that have separated from other battalions and have joined the key group, but have not changed their name or significantly changed their leadership structure. Both the joining and joined groups keep their separate names and retain independent command structures. When evidence was found that a group qualified as a joining battalion, a "yes" was placed in the column and the group IDs as well as the year they joined were also recorded. Additional information about the joining battalion, such as reason for joining, dynamic between the two groups, ideological orientations of the joining group, size of joining group, etc., was also recorded.

- ***Merger***—A merger unifies two or more groups in a partnership where they are then defined under one name and command structure. Neither of the independent groups who make up the merger exist in any form outside of the merger, meaning that individual subordinate groups disappear and resources are shared. When evidence was found that two or more groups merged, the name of the group that resulted from the merger was recorded in the year it was established. Additional information on the reasons the groups merged, how this will change the group's operations, if a group seems to be the dominant group, etc., was also recorded.

- ***Alliances*** —The alliances or operations rooms of which the group was a member. If we found evidence that a group was a member of an alliance, their membership was recorded for all years following the onset of the alliance, unless there was evidence that the group left the alliance. We recorded if any information could be found about the alliance membership, including why this group joined the alliance, the size of the alliance group, and ideological beliefs of the alliance.

- ***Powerbroker/potential powerbroker***—The powerbroker designation comes from a series of reports issued by the Institute for the Study of War (Cafarella and Casagrande 2015, 2016). In these reports, groups may qualify as either a powerbroker or potential powerbroker for each governorate, the definitions of which are included below. The ISW defines a powerbroker as "a group that defines the success of military operations against either the Assad regime or ISIS, is strategically located, and/or plays a leading role in governance" and a potential powerbroker as "a group that could achieve significant battlefield effects against Jabhat al Nusra and/or ISIS in western Syria upon receipt of increased outside support, including securing direct military gains and cohering other smaller brigades into new coalitions." If these reports designated a group as either a powerbroker or potential powerbroker, it was recorded in the "General" row.

- ***Ideal economic polity***— What is the economic order (e.g. type of state economy) that the rebel group aims to create post-conflict? While most groups seldom publicize their economic philosophy, there are some groups who indicate an affiliation to Marxism. A few Kurdish groups, for example, espouse Marxist views. If there was evidence found of a groups economic ideology, a group was categorized as one of the following:

  - ***Marxist***— prioritizes state ownership of the means of production (e.g. FARC, the original IRA).
  - ***Social democracy***— prioritizes a mixed economy, where there is private property with some public services under public ownership.
  - ***Capitalist***— prioritizes the private ownership of the means of production and a free market economy.

- ***Ideal political polity***— What is the political order (e.g. type of government) that the rebel group aims to create post-conflict? How do they envision their government post-conflict? Political ideology can take one of the three values, described below. When evidence was found of a group's political ideology, it was recorded. This variable may change throughout the conflict, as some groups claim to support elections early on but later change their views. When evidence was found of such changes, these were recorded and described in the notes section.

  - ***Full democracy***— prioritizes representative government through elected officials.
  - ***Hybrid regime***— combines features of democratic and autocratic regimes; often holds regular elections but is also characterized by political repression.
  - ***Authoritarian***— prioritizes concentrated and centralized power under a single leader or a small elite cohort.

- ***Ideal Territorial Polity***—: Does the rebel group want to gain territory? If so, what are their aspired boundaries? This variable can take one of four values, although most groups active in the Syrian conflict fall into either a status quo or global territorial ideology.
  - ***Secular but group is not bound by religious goals***
  - ***Secular but religion is important to membership***
  - ***Moderate religious order***— laws have a basis in religious texts but allows flexibility in private conduct
  - ***Strict religious order***— prioritizes order governed by religious law in all aspects of life. E.g. calling for strict Sharia
- ***Ideal religious polity***—: What is the religious order (e.g. involvement of religion with the state) that the rebel group aims to create post-conflict? When evidence was found of a group's political ideology, it was recorded.
  - ***Status Quo***— The rebel group likes the physical borders of the state but wishes for a different governmental order
  - ***Bounded Autonomy***— Group seeking partial political separation from parent state and/or seeking self-determination with the ability to govern themselves under auspices of central authority
  - ***Separatist***— Group seeking full political separation from parent state; seeking self-determination as a fully independent state or seeking to be incorporated within another state
  - ***Globalist***— Are they trying to actively take over new territory and found their own empire?
- ***Internal Conflict***— Internal conflict captures a range of scenarios, including purges of high-ranking officials, major defections, ousting of leaders; tension between group members; disagreement among group leaders; conflicting media statements from branches of the same group. If evidence of one of these scenarios or a was found in the research of a key group, it was recorded and briefly explained.
- ***Notes*** — Notes capturing any details about a group's narrative, relationships with other groups, tactics, and any other notable information that did not fit with a more relevant variable was recorded in this column.

D.8 Covariate data for the Syrian state

Because the state did not maintain a social media presence to document its attacks during the conflict, we collected three types of data from outside sources designed to dovetail with our rebel social media data. While measures of state capacity and presence were imperfect over the conflict, we determined that human rights abuses, state use of WME, and control of territory were three types of available and reliable data that would provide some sense of state behavior over the course of the war. Our data about state use of violence came from an independent

human rights organization's monthly reports dedicated to monitoring and documenting human rights violations in Syria: the Syrian Network for Human Rights (SN4HR). Separately, we recorded events from the SN4HR monthly reports that involved the recorded WME use by the Syrian state. Most of these WME attacks involved the use of chemical weapons or phosphorous.[14] As an additional checkpoint on the changing nature of state power and control, we leveraged Carter Center reports on territorial control over the course of the conflict. Our data captures whether each faction that the Carter Center tracks (Syrian Government, Rebels, Kurdish groups, and ISIS) had control of a Syrian governorate in each month starting from January 2013 to December 2018. The compiled data is an integration of the Conflict Mapping and Analysis Reports from the Carter Center's Support for Peace in Syria project.

D.8.1 State Violence

One critical feature of the Syrian civil war is that the state remained intact as an actor throughout the duration of the conflict, and adopted many of the same strategies as did rebel groups, including targeting civilians, deploying WME attacks, and making and breaking alliances. To model the conflict without the inclusion of state attack activity would offer a warped perspective on conflict dynamics, as both rebel group attacks and alliance formation could be driven by state action, not just the activities of other militant groups. Additionally, collecting data on rebel groups but not the state offers the state an artificial exceptionality, positioning it as above the fray.

A critical source of data on the activity of the Syrian state was the Syrian Network for Human Rights (SN4HR), which is an independent human rights organization dedicated to monitoring and documenting human rights violations in Syria (`https://sn4hr.org/about-us/`). The SN4HR creates detailed monthly reports describing human rights abuses in Syria. When analyzing these reports, we recorded the SN4HR's documented number of several different types of human rights abuses. These abuses included deaths (of civilians, medical personnel, civil defense personnel, and media personnel) death by torture and massacre, cases or arbitrary arrest, and attacks on various different facilities (including but not limited to worship, medical, and educational facilities) also the number of outlawed and indiscriminate attacks as well as use of barrel bombs were measured. The reports often provided names of individuals affected, details about the damage done, information about the perpetrators, as well as relevant images and specifics about the incidents. Due to the detail of the reports, they were read completely to find relevant and pertinent information, and we extracted information related to the key variables above, using the background information to help guide coding decisions as needed. In all cases of a range or unclear number of events, the lesser number or lower value was used.

While the SN4HR is an incredibly comprehensive resource, due to its data collection process, there is some variation in how certain abuses and locations were cataloged over time. For instance, sometimes Damascus and the Damascus suburbs were treated as separate locations, and other times they were collapsed together. When in doubt, our team grouped these claims under Damascus, the city. We also identified inconsistencies related to the classification

[14] Phosphorous is a fast-burning substance that is often used during combat to create smokescreens. It easily sticks to skin or clothing and can cause severe external and internal burns that can go through bone and organs. It is not classified as a weapon under the Chemical Weapons Convention of 1993. It is illegal to use on civilians.

of targets, reported deaths, and whether attacks were tallied by group or aggregated. We detail the coding decisions we made in each of these cases in the appendix.

D.8.2 State and non-state WME Use

Separately, we recorded events reported in the SN4HR monthly reports that involved the reported use of WME by the Syrian state. Most of these WME attacks involve the use of chemical weapons or phosphorous.[15] For each event, we include variables that– where available– capture the date of the attack, the location, the target, the time, the number of deaths, the number of injuries, as well as notes. Where possible, each observation is cross-validated by checking the UN Organization for the Prohibition for Chemical Weapons (UN OPCW) report on Syria, the Syrian Archive (a project led by Syrian human rights activists and lawyers to document human rights abuses during the civil war), and the UN Human Rights' Council's Independent International Commission of Inquiry on the Arab Syrian Republic (UN HRC IICISAR),[16]

We also identified a total of 311 militant WME attacks; attacks must have included two of the three prongs (destructive attack, civilian targeting, unconventional elements) detailed in Table **??** to be coded as involving a WME. Coders were instructed to favor "over-coding" WMEs, flagging any claim that could possibly involve a WME. The study authors then went through this over-captured list to ensure accurate capture of all WMEs.

WME use in Syria most frequently involves explosives, suicide bombings, and executions, rather than nuclear, chemical or biological weapons. Suicide bombings, explosives, and executions are noteworthy because while they may require a dedicated membership, they do not require a great deal of prior knowledge or technology, but their effects are magnified by the internet and social media. Thus, militant group ability to use WME is not contingent upon technological capacity or foreign support. In Syria, the target of WME attacks is most frequently the Syrian regime, rather than other militant groups or civilians.

Type of WME Use in Syria

WME Type	Frequency of Occurrence
explosive	100
suicide	72
execution	58
shelling	39
landmine	17
torture	13
chemical	9
other	7

WME attack frequency is highest around Syria's northeast border and in the Damascus suburbs: Al-Hasakah, in northeastern Syria, and Rif Damashq, in the Damascus suburbs, followed by Latakia, Homs, Damascus and Qunietra. This overlaps with, but is distinct from, the overall violence, as Damascus and Rif Damashq experienced the most attacks. Thus, WME use cannot solely be explained by conflict intensity. Some regions which experience high levels of violence, such as Latakia and Homs, have relatively infrequent attacks involving WMEs.

[15] Phosphorous is a fast-burning substance that is often used during combat to create smokescreens. It easily sticks to skin or clothing and can cause severe external and internal burns that can go through bone and organs. It is not classified as a weapon under the Chemical Weapons Conventional of 1993. It is illegal to use on civilians.

[16] Both UN reports are sometimes criticized by human rights activists because of the role close allies of the Assad regime played in the report construction. While human rights groups may have the incentive to document as many chemical attacks as possible, international bodies may be more cautious in their assumptions.

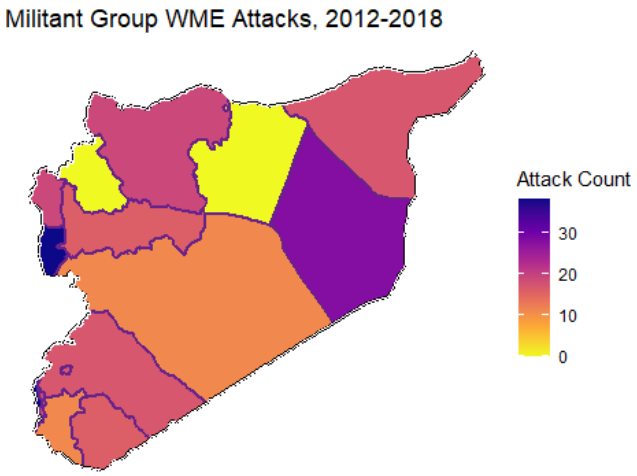

Figure 7: *Militant WME use by governorate (2012-2018)*

The proclivity to use WMEs differs across militant groups, with large and active groups such as ISIL, Tahrir al-Sham, and Al-Nusra Front carrying out the bulk of attacks involving WME. This is not surprising; these groups have existed through multiple years and across regions and have the greatest opportunity to carry out attacks. However, as a proportion of total attacks, a relatively small number of these groups' attacks involve WME. Smaller groups that do not assume territorial control or state building roles, such as Ansar Al-Din Front, Khattab Company and the Talhah Ibn-ubaydallah Brigade, carry out attacks that involve WME in the highest proportions of their total attacks. This suggests WME attacks are not simply a product of group size or strength (which is often considered a measure of group capacity—(Christia 2012)).

WME use by militant Group

Group	WME Attack Count	WME Use as Proportion of Total Attacks (as a percent)
Islamic State	83	2.39
Tahrir al-Sham	33	3.86
Al Nusra Front	26	2.30
Al-Rahman Corps	11	2.06
Ahrar al Sham	11	1.00
People's Protection Units	11	2.32
Jund al-Aqsa	10	7.58
Al-Tawhid Brigade	9	1.82
Fath al-Sham Front	8	2.23
Islamic Front	7	1.06
Al-Furqan Battalion	5	2.55
Jaysh al Islam	5	0.40
Ansar al-Islam	3	7.32
Fistaqim Kama Umirat	3	0.67
Liwa al Islam	3	1.60
Levantine Front	3	0.50
Sa'id Ibn-Zayd Brigade	3	21.43
Southern Front	3	0.50
Al-Faruq Brigades	2	25.00
Ansar al-Din Front	2	40.00
Khattab Company	2	40.00
Talhah Ibn-ubaydallah Brigade	2	40.00
Syrian Coalition	2	13.34
Syrian Democratic Forces	2	1.20

D.8.3 Territorial Control

The Syrian Government Control dataset tracks the territory controlled by the Syrian Government (and other major actors) throughout the Syrian civil war. It captures whether each actor (Syrian Government, Rebels, Kurdish groups, and ISIS) had control of each Syrian governorate

in each month from January 2013 to December 2018. The compiled data is an integration of the Conflict Mapping and Analysis Reports from The Carter Center's Support for Peace in Syria project raw citation in file.

There are fourteen governorates within Syria, and they are the primary administrative unit parallel to a state or a province. Governorates are comprised of 60 districts, the next administrative unit below the governorate. The governorate may represent an overly blunt measure of territorial control, as actors often control a fraction of the territory or are contesting control. However, more discrete data on territorial control was incomplete or unreliable at the time of writing.

Observations in this data set are listed at the governorate-month level. There are three possible values within this dataset: "yes", "no", and ".". The value "yes" identifies that an organization had indisputable control of at least part of a governorate. In many cases, there was split control between multiple actors. In that case, every actor with indisputable control was given a "yes" for that particular governorate. Indisputable control was defined as beyond the front lines of a given militant organization. While attacks could be made within an area of control, such as air strikes, car bombings, and assassinations, the mere presence or success of an attack did not grant control to a perpetrator. Rather, there were several distinct ways an organization could be given a "yes" for a particular governorate. The simplest way to be given a "yes" was if a report included a map of territorial control. Yet the majority of reports before March 2017 did not include territorial maps.

In reports without territorial maps, a "yes" was given if the prose identified indisputable control of a governorate. Often times the Carter Center reports would identify indisputable control in one of two ways: by identifying troop or front line advancement or labeling towns under the possession of an organization. For instance, if a report mentioned that "ISIS advanced their front line in Aleppo, clashing with forces in Regime controlled Afrin", both ISIS and the Regime would be granted a "yes" for control of Aleppo because their front lines (thus area of control) were within the Aleppo Governorate. A report may also say that "ISIS troops clashed with forces in the pro-Regime town of Afrin", only granting a "yes" for the Regime as the report only indicated that the Regime had sufficient control.

The other two values that were assigned were "no" and ".". The value "no" was only assigned in reports that included territorial maps that showed that an organization lacked any governorate. A period indicates that there is not enough data to draw conclusions about a group's territorial control over a governorate. In cases where there is no data about that governorate in a particular month, "no data available" was written in the notes section.

Due to the constantly evolving situation in Syria, no assumptions were made month to month about territorial control. Rather, if it was unclear if an organization had undeniable territorial control, a period was put in place rather than a yes or no. Due to this fact, there are large sections of missing data on Regime and organizational control. The decision to not include assumptions about territorial control was made to avoid inconsistencies when analyzing reports that did not include territorial maps.

This dataset heavily relies on the high standards of the Carter Center Conflict Mapping and assumes that territorial control would not be shown on a map without sufficient evidence. With that said, it is important to keep in mind that this dataset was created using secondary sources and there is a degree of separation from the primary source data.

E Group Disambiguation Sheet

This disambiguation list is a critical input to our coding process but also offers future researchers a clearer lens on the population of militant groups in Syria. All coders recorded new group names as they were encountered alongside relevant network information about those groups (such as if they mentioned hiearchial relationships with other groups). Furthermore, coders added alternate spellings and names to rebel group entries, improving the search capabilities of other coders and reducing duplication of rebel groups. Lastly, if a coder wanted to create a new entry within this repository, they would add their initials and provide an associated claim number so that all rebel groups could be traced to where they were first mentioned in our dataset.

Thus, the "Group Name Disambiguation" repository took on a life of its own, becoming a space for coders to slowly piece together a network of all rebel groups in the Syria civil war, their aliases, and the ways in which they related to one another. In other words, we documented the life cycle of many Syrian rebel groups—how they were born, merged, splintered, and eliminated. This provided us with an uncharacteristically intimate view of how rebel groups form, mutate, and adapt to a civil conflict environment. This document was continually updated throughout the coding and data-cleaning process.