Advances in Data on Conflict and Dissent

Kristian Skrede Gleditsch

Abstract In this chapter, I review the role of data in driving innovation in research on conflict. I argue that progress in conflict research has been strongly related to the growth of systematic empirical data. I draw on a series of examples to show how data have served as a source of theoretical innovation. I discuss early models of conflict distributions and their enduring relevance in current discussion of conflict trends and the evidence for a decline in violence. I consider the interaction between theoretical models of conflict and empirical analysis of interstate conflict, as well as the rapid growth in disaggregated studies of civil war and developments in data innovation, which in turn help generate new research agendas. I conclude with some thoughts on key unresolved problems in current conflict research, namely the lack of attention to incompatibilities as the defining characteristics of conflict and accounting for scale and differences in event size.

Keywords Conflict · Data · Models · Distributions · Progress

1 Introduction: The Need for Data in Computational Social Science

Conflict research has a long history, where efforts to record or measure conflict have a central place, but computational approaches to date have been less common. There are some notable exceptions that clearly demonstrate how computational approaches can be very useful in order to more explicitly explore counterfactuals and variation beyond what is available to us through the historical record (Bremer and Mihalka 1977; Cederman 1997). However, computational approaches tend to be the most compelling and effective when they are closely integrated with

K. S. Gleditsch (✉)
University of Essex and Peace Research Institute Oslo (PRIO), Oslo, Norway
e-mail: ksg@essex.ac.uk

E. Deutschmann et al. (eds.), *Computational Conflict Research*,
Computational Social Sciences, https://doi.org/10.1007/978-3-030-29333-8_2

subject-specific theoretical puzzles and informed by empirical data. In this chapter, I review the role of data in driving innovation in research on conflict.

The thesis I advance is that innovations in research on conflict often have followed new data developments. My argument is not that we can simply substitute theory with more data. Indeed, descriptive data rarely speak for themselves, and new and more detailed data will by themselves rarely lead to new theoretical breakthroughs. Indeed, the best data sources are usually based on solid initial theoretical foundations that guide the data collection efforts. However, it is difficult to find good examples where pure theory has had a major transformative impact on conflict research, in the absence of substantial engagement with empirical data. By contrast, data innovations have often helped restate and refine existing research agendas, and open new avenues for theoretical development.

To justify this thesis, I draw on a series of examples of how data have served as a source of theoretical innovation, starting with early models of conflict distributions and their enduring relevance in current discussion of conflict trends, and then how more recent developments in data innovation contribute to new research agendas. I conclude with some thoughts on what I see as particularly important unresolved problems in current conflict research, namely the lack of attention to incompatibilities as the defining characteristics of conflict and accounting for scale and differences in event size.

2 Conflict Research and the Impact of the Early Conflict Data

If we define data rather widely as any empirical observations, then there is of course a long history of data in terms of detailed historical accounts of individual conflicts. Many of these could be highly analytical, as Thucydides' (2000) discussion of the causes of Peloponnesian war (believed to be written around 410 BCE). However, historical accounts tend to be highly case-specific and are rarely comparative or systematic, in the sense of trying to cover a population of conflict or focus on representative cases. Moreover, outside historical accounts, much of the general early research on conflict focused heavily on theory and analyzing conflict in an abstract manner, often detached from descriptive data altogether. Hobbes (1651, p. 78), for example, argued that scholars should try to identify the general conditions that make war possible rather than individual events, just as "foul weather is not based on isolated showers, but inclination to rain." This is in many ways a quite sophisticated anticipation of security dilemmas and efforts to develop more general theory. However, the lack of attention to data and observations also moved us further away from efforts to quantify risk, such as assessing how frequent conflict actually is and how much variation in inclination we see across specific types of conditions. Kelvin (1883) famously equated the quality of science to quantification. Without measuring conflict, we are often left without realistic assessments of risk.

A statement indicating that something "is possible" tells us little more than that probability is above 0 or impossible but less than 1 and certain. Harking back to the weather analogy, the nature and shape of weather distributions certainly play a central place in meteorology. Examining descriptive data on such distributions can help us keep track of how some places have more foul weather than others, and provides a basis for evaluating the possible causes why.

Against this more stringent yardstick, comparative data on conflict are a relative recent development in the long history of conflict research. It remained until recently largely a fringe activity, perhaps in part as a result of policy orientation and aversion to statistics and quantitative methods among many traditional security studies scholars (see, e.g., Fazal 2016). One of the earliest datasets was collected by a sociologist, Sorokin (1957[1937]), who sought to use data to test his theory of conflict as a result of value divergence. With comprehensive information on the dates on key battles and troop sizes since antiquity, Sorokin's data were a major achievement. However, some features also limited their applicability. As the data were restricted to conflict between major powers, they could not speak to conflicts within states or conflicts with smaller powers. There is also no clear delineation of what makes states major powers, and a risk of circularity if influence for conflict is defined based on whether states tend to fight more.

Wright (1942/1965) developed another influential dataset, intended to test a theory of peace as a result of active interstate organization and coordination that served to constraint possible factors that may lead to conflict if left unchecked. Although these data cover a shorter period than Sorokin's, they also included a more comprehensive delineation of states involved in conflict. Wright also devoted a great deal of attention to developing clear inclusion criteria for the data collection efforts. Given his background in law, it is perhaps not surprising that the definitions were skewed towards legal conceptions of war, but his efforts and structured approach had a major influence on subsequent efforts to define war.

The most unusual data pioneer was Richardson (1960), a physicist who sought to identify a dataset of violent events to assist with more fundamental mathematical and statistical models of conflict. Richardson started to work collecting conflict data after World War I, but the data were not published until much later. Richardson's unit of analysis was deadly quarrels, based only on observable deaths. The incidents were classified by their severity in terms of fatalities, binned by "orders-of-magnitude" on a log10 scale. The data were intended to be exhaustive for events above 1.5 (about 32 fatalities). Richardson provides an important first discussion of some of the problems in counting wars from historical records—who are the combatants, when did a war start/end, how many died? In a pithy quote, Richardson (1960, p. 35) concluded that "thinginess fails" when we try to create data on wars as events, and "the concept of a war as a discrete thing does not quite fit all the facts." Moreover, he was the first to explicitly use randomization to consider the sensitivity of his conclusions to decisions about lumping together events as a single war versus splitting episodes within longer wars.

One of the first conflict distribution models analyzed by Richardson (1948) considered the severity and frequency of conflict. He noted that there was a regular

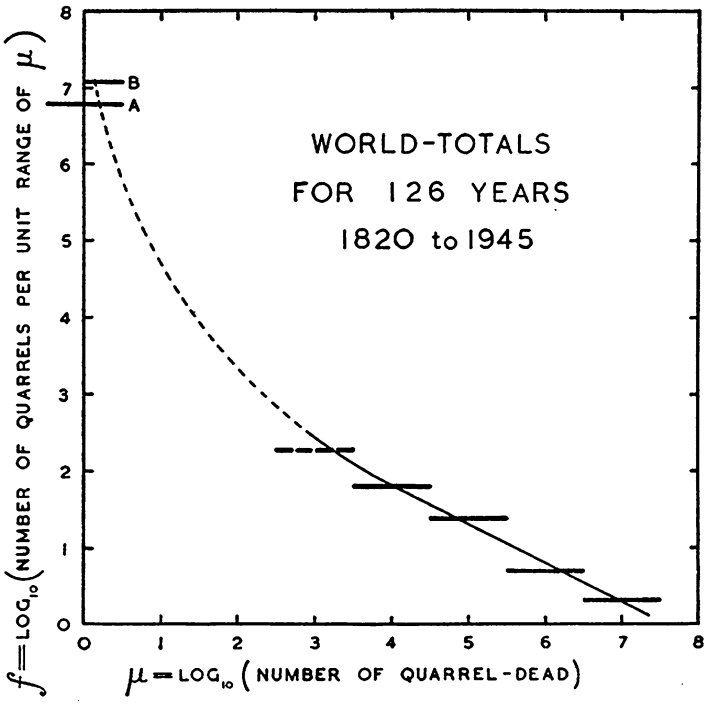

Fig. 1 Quarrel frequency and severity, from Richardson (1948). Richardson's data are binned by severity, hence the horizontal lines

relationship between conflict severity and frequency, where the severity of a conflict in terms of number of people killed x is inversely proportional to frequency. More formally, the frequency of a conflict of severity x scales as $P(x) \propto x^{-\alpha}$, where $\alpha \approx 2$. Richardson's data are displayed in Fig. 1, and provide one of the first empirical examples of a power law. One of the properties is that multiplying severity by a given factor yields a proportional division of the frequency. For example, doubling severity halves frequency. Power laws will appear as roughly linear if displayed on doubly logarithmic axes.

As shown below in Fig. 2, we find a similar relationship for other conflict data sources as well, including more recent data on interstate wars. Indeed, this relationship turns out to be a common feature of many conflict data distributions, including more fine-grained data on individual terrorist attacks (Bohorquez et al. 2009; Cederman 2003; Clauset et al. 2007). However, it is not universal, and it does not hold for all types of conflict. As can be seen in Fig. 2, the fit is much less compelling for civil wars, where we see "too few" severe conflicts in the tail for the observed data to fit well with what we would expect under a power-law distribution.

Skeptics may wonder why this should be regarded as an interesting finding. One way these results can be useful is to assess the expected frequencies of specific types

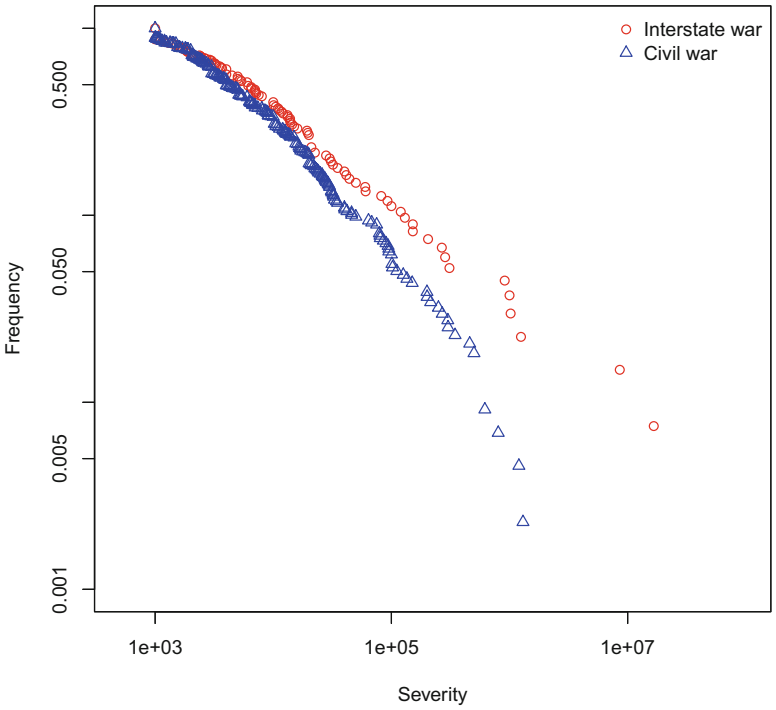

Fig. 2 Frequency-severity (i.e., casualty) distribution for wars, based on the expanded war data from Gleditsch (2004), doubly logarithmic scale

of events. For example, 9/11 is often portrayed as being an unprecedented "Black Swan" event, following the terminology of Taleb (2007). Clauset and Woodard (2013) show that the likelihood of observing an event with the same magnitude as 9/11 since 1968 based on the observed data is as high as 11–35%, depending on the specific assumptions used. The fact that tail events are more likely than many anticipate based on the apparent "typical" conflict is a stark reminder of how major conflicts such as World War I can emerge, even when observers see "no clouds on the horizon." Furthermore, finding that observed events *do not* fit a power law can also be useful to think about possible causes. For example, the poor fit for civil wars suggests that there must be some limiting factors that may prevent civil wars from escalating to more severe conflicts at the same rate as interstate wars (Miranda et al. 2016). For example, non-state actors may have limited resources to increase the war effort or hard constraints on their ability to escalate conflict beyond a certain level.

A second model considered by Richardson pertains to timing of wars. Richardson (1960) found that outbreaks by year were consistent with a Poisson process, a common model for independent random events, where "there is a constant very small probability of an outbreak of war somewhere on the globe on every day" (p. 243). More formally, the number of wars n in an interval such as a year, given a

probability of a war breaking out p, will be $e^{-p}\frac{p^n}{n!}$. Given this formula, as long as p is small, we are most likely to see years without onset, followed by years with a single onset, and the likelihood of seeing a year with n (or more) wars falls quickly the higher the value of n. The idea that conflict outbreaks are random does not sit easily with traditional theories of international conflict. However, many other analyses have found that it is very difficult to reject the simple Poisson model for common conflict data sources (Gleditsch and Clauset 2018; Mansfield 1988).

Again, skeptics may wonder why this analysis is relevant. The Richardson model of timing has become important again in the recent debate over trends in conflict. There has been a great deal of research on the apparent decline in warfare and organized violence, especially after the Cold War. Prominent books by Goldstein (2011) and Pinker (2011), for example, show an observed decline in the number of wars and the number of people killed in war and discuss possible causes. Most people accept that the observed data indeed indicate a decline (see, e.g., Gleditsch and Clauset 2018). However, there is more controversy over whether the observed data provide strong evidence for a trend, or shifts in the underlying distribution of conflict. How do we know that we have not had a spell of good luck, and how confident are we that the number of conflicts would remain low? Under just a slightly different turn of events, for example, the Cuban Missile crisis could have escalated to a severe conflict (Gaddis 2005). Whether we deem trends to endure is of course to a large extent a question on theory, and here I will focus mainly on the statistical aspects of assessing trends. We are used to seeing the historical record as a population, and many find it odd to discuss alternative worlds (Tetlock 1999; Tetlock and Belkin 1996). However, if we think of conflict outbreak as a stochastic process, then it is entirely possible to see a decline of conflict over a period, even if there is no change in underlying frequency of conflict.

Whether we can reject a model of no change based on the independent outbreak and power-law distributions is explored recently in two papers by Cirillo and Taleb (2016) and Clauset (2018). Although there are a number of innovations in analysis and data compared to Richardson, they both consider variants of the timing and frequency-severity models that we have seen. In brief, Cirillo and Taleb argue that we cannot in principle say anything about trends since severe conflicts are so rare. They calculate that for a conflict with five million casualties, the expected waiting time between conflicts would be over 93 years. Based on this, one might argue that one cannot make any conclusions about notable trends just from observing a decline. Clauset also tries to test for evidence of shifts in the distribution after 1945. He finds some evidence that the most severe conflicts may be less common, but not sufficiently strong evidence to reject the no change null hypothesis. Oddly enough, changes such as nuclear weapons, the growth of the number of states, and all types of nonstationary factors we think influence war, such as democracy and trade, appear to have had no impact on the distribution of conflict.

Other scholars have started to examine a broader range of conflicts at the lower end of the distribution, and whether there is evidence of changes in the distribution more recently than 1945 (the only period considered by Clauset), using

change-point detection techniques. For example, Hjort (2018) finds evidence for a break in the distribution in 1965, which coincided with the opposition to the Vietnam War and the hippie movement, so perhaps Woodstock had a longer legacy. Focusing on ethnic civil war, Cederman et al. (2017) find evidence for a change point in the series in the late 1990s, and also provide evidence that the change appears to be due to greater ethnic accommodation. Just as civil wars can be promoted by ethnic exclusion, we are less likely to see an onset of conflicts after changes towards ethnic accommodation and more likely to see conflict termination.

3 Data and Progress in Conflict Research

There have of course been many other important developments in conflict research beyond research on trends. However, one might perhaps also contend that the extent of progress has not been proportional to effort, or at least it has been more limited than the very high aspirations. There has been a great deal of path dependence, where existing data are simply duplicated, without innovation and further refinements. For a long time, there was a dominant tendency to let often ill-defined traditional theories of conflict guide empirical inquiry, and much ink has been spilled on investigating vague notions from the realist school of thought, suggesting that conflict must be some kind of function of the distribution of power across states in the system (Singer 1980). Many analyses have sampled on the dependent variable and just looked at conflict cases, without considering non-war cases or explicit baseline models (Most and Starr 1982).

However, there has undeniably also been a great deal of progress, and much of this has been driven by data developments interacting with theory development (Gleditsch et al. 2014). For example, the early efforts to come up with more explicit list of states allowed defining populations of potential actors, and to derive better explicit models of the opportunities for conflict among individual states or dyads (Bremer 1992). Data on the geography of states has similarly led to a great deal of interesting research on the role of borders, distance, and conflict (Starr and Most 1983). Data on political institutions and economic exchange helped spur the wave of research on liberal peace, or the possible restraining effects of institutions or interdependence on the use of force (Oneal and Russett 2001; Simowitz 1996). This has in turn led to new interest in using network approaches to understand how individual states are embedded in larger networks of interdependence beyond the dyad, as well as new methods for dealing with temporal and spatial interdependence in statistical analyses (Beck et al. 1998; Kinne 2009). Van Holt et al. (2016) conduct a more formal analysis of scientific influence in conflict research based on citation patterns. Their findings are visualized in terms of paths between influential articles and common topics in Fig. 3. It is clear from Fig. 3 that many of the influential articles in the graph on interstate conflict are precisely those that describe new dataset or analytical methods. Notable examples include Jaggers and Gurr (1995, introducing the Polity democracy data prominent in studies of the democratic

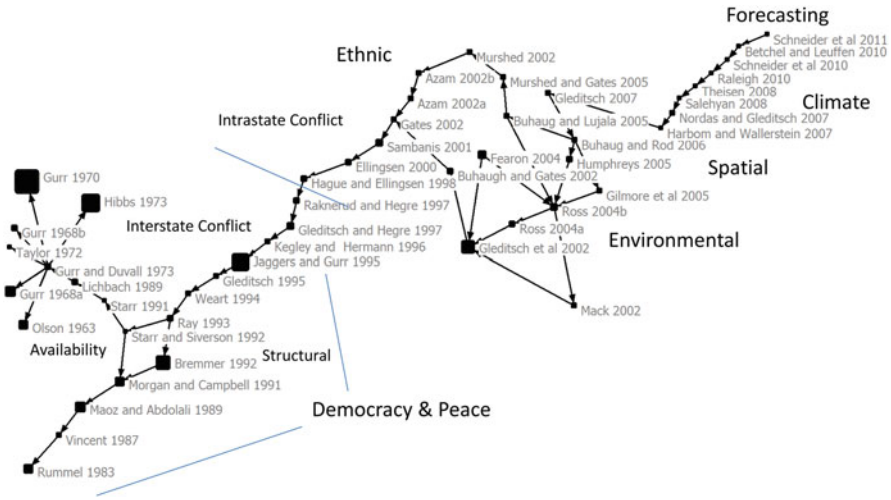

Fig. 3 Critical path and scientific influence in conflict science, reproduced from Van Holt et al. (2016)

peace) and Bremer (1992), which was one of the first articles to propose systematic approaches to dyadic analysis of the onset of Militarized Interstate Disputes (Jones et al. 1996). The article presenting the first version of the extended backdated version of the Uppsala/PRIO armed conflict data prior to 1990 stands out as central in the upper right section of the figure on interstate conflict (Gleditsch et al. 2002).

The specific topics of interest have clearly changed over time. It is notable that the entries in the section of the graph for intrastate conflict in Fig. 3 have much more recent publication dates, and the history of research on civil war differs notably from interstate conflict research. In general, quantitative research on civil war suffered much less from a legacy of traditional theories. In the mid-2000s, there was lot of interest in trying to develop more disaggregated data on civil war, in part promoted by a collaborative network of conflict research in Europe which generated a special issue of the *Journal of Conflict Resolution* (Cederman and Gleditsch 2009). We have seen the development of new data that disaggregate and identify the specific actors involved in conflict (Cunningham et al. 2009), identify more detailed information on specific attributes such as the ties of actors to ethnic groups and more detailed information on ethnic groups (Vogt et al. 2015), and data that provide more detailed information on events within conflicts and their geographical location (Raleigh et al. 2010; Sundberg and Melander 2013). There have also been a number of utilities develop to combining different data sources, such as the geo-spatial cell structure in PRIO-GRID (Tollefsen et al. 2012) and the R package MELTT to match different event data sources by location, time and type (Donnay et al. 2019). Moreover, there has been a great deal of progress in automated coding of information from text sources such as news media reports, which provides an opportunity for real-time

monitoring of events (Gerner et al. 1994, Schrodt and Gerner 1994, see also Maerz and Puschmann, chapter "Text as Data for Conflict Research: A Literature Survey" in this volume). There has also been more attention to out-of-sample forecasting as a better approach to model and theory evaluation (Ward et al. 2010). Importantly, out-of-sample evaluation can help guard against the problem of in-sample overfitting, since it will often be the case that increasingly more complex models that may fit the estimation data well will perform worse out of sample than simpler models. In short, the study of civil war from the mid-2000s has been a period of rapid progress, and much of the progress has clearly been promoted by the development of new data sources and the interaction between theory and data.

4 The Essential Interaction Between Theory and Data in Conflict Research

Although more data have helped take us further, the interaction with theory remains essential. Whereas the development of data tended to follow theories or initial ideas in the early development of conflict research, it is now increasingly common to see more purely data-driven projects, exploiting the vast amount of available data on conflict. Exploratory analysis can often be very helpful and illuminating in its own right, especially if it is guided by new methods that may have advantages over the existing approaches that have commonly been used and help illuminate new aspects. Yet, there are also many cases where we arguably learn less from the analysis conducted, even if they are very competently done from a technical perspective. For example, Zammit-Mangion et al. (2012) use models from geostatistics to model high resolution data on events in Afghanistan, obtained from Wikileaks, on a database of Significant Activities (SIGACTS) compiled by the US Army. They argue that this framework can be helpful for detecting and predicting conflict dynamics such as diffusion and relocation. The model seems to have high predictive ability, but on closer inspection it becomes clear that much of the heavy lifting in the predictive ability is done by the temporal lags. There is also a discernible "ring" in the spatial forecasts of location, which appears to reflect how improvised explosive devices tend to be placed around the Highway 1/Ring Road that circles the country. Ultimately, the model has limited content on the motivation of the actors, and the framework deemphasizes conflict as interaction between antagonists. Moreover, since the SIGACTS data primarily record events by actors perceived as hostile by the US Army, these data do not contain information on the events and actions by coalition forces that we would need to actually study the interaction between the parties and how the conflict evolves as a result of this (Weidmann and Salehyan 2013). Although data can be powerful tools to evaluate and extend theories, we need to avoid putting the data cart in front of the horse, or we risk developing 'weapons of mass distraction' that provide limited insights, no matter how much they appear to be scientific.

5 Key Unresolved Problems in Data for Conflict Research

In closing I would like to flag two important problems in conflict research that I think have not received sufficient attention and remain difficult to consider in existing data sources. The first is the tendency to equate conflict exclusively with violent events, which is very widespread in applied research on conflict. This is not consistent with definitions of conflict that tend to highlight incompatibilities or conflict of interest between actors. Boulding (1963, p. 5), for example, suggests that "[c]onflict may be defined as a situation of competition in which the parties are aware of the incompatibility of potential future positions, and in which each party wishes to occupy a position that is incompatible with the wishes of the other". From this perspective, conflict as an incompatibility could motivate the use of violence, but violence in and of itself is not a defining characteristic of conflict (see also chapters "Advancing Conflict Research Through Computational Approaches"; "Migration Policy Framing in Political Discourse: Evidence from Canada and the USA"; "The Role of Network Structure and Initial Group Norm Distributions in Norm Conflict"; "On the Fate of Protests: Dynamics of Social Activation and Topic Selection Online and in the Streets" of this volume). The requirement that conflict must be perceived by the actors help to demarcate from other very expanded definitions of conflict, such as structural violence that extend the concept of conflict to situations with "objective" interest not necessarily experienced or understood by the actors (Høivik and Galtung 1971). Most and Starr (1983) provide a comprehensive review of other definitions of conflict, most of which have a similar emphasis on conflict of interest as opposed to violent action.

The tendency to equate conflict with manifestations of organized violence has led some researchers to either explicitly or implicitly treat situations without conflict as "peace." This is highly problematic, since we fail to distinguish cases where there are no objective conflicts of interest between actors and cases where conflicts of interest exist, yet do not result in the use of violence. Organized violence requires collective action, and all forms of efforts to initiate collective action may fail for a number of reasons (Sandler 1992). Even when actors have common interests on an issue and would benefit from a change, such as fostering regime change or replacing a government, they do not necessarily have sufficient private incentives to participate in dangerous activities. As such, there will be a temptation to free ride as the benefits of successful dissent would be public and cannot easily be restricted to active participants (Lichbach 1995; Tullock 1971). Moreover, states can deter or raise the costs of collective action by sanctions or retribution. But more fundamentally, conflict may also be waged using means other than violence, including for example demonstrations and strikes (see also chapter "On the Fate of Protests: Dynamics of Social Activation and Topic Selection Online and in the Streets" in this volume). Sharp (1973) and Chenoweth and Stephan (2011) document many instances of important campaigns waged using only non-violent means. Violent and non-violent tactics can be plausible substitutes, where we may not see organized violence used in a conflict because an actor has a comparative advantage in non-violent forms of

contention. For example, over the last couple of years, Venezuela has seen massive mobilization against the Maduro government and proposed institutional changes. On 19 March 2017, a so-called Mother of all Marches of protest mobilized as many as six million participants nationwide, according to estimates by the survey company Meganálisis based on traffic flow and demonstration movement data, an extreme relative level of mobilization in a country of just over 30 million inhabitants (see Lugo-Galicia 2017). Although there have been many instances of violence against protestors as well as occasional violent responses by protestors and riots, we do not have a conventional civil war in the sense of organized armed violence by opposition. Yet, it would be absurd to characterize this as "not a conflict" since we do not see organized violence.

Many studies of civil war have tried to identify potential incompatibilities by focusing on the political and economic status of ethnic groups. From this perspective, all ethnic groups that are disadvantaged in a given state could be seen as potential conflict situations where there exist plausible grievances against the state and motives for dissent. Yet, conflict is a much more general concept than this. First, many violent conflicts are not ethnic, and the share of violent conflicts that are clearly ethnic has arguably fallen. Figure 4 displays the share of ongoing armed civil

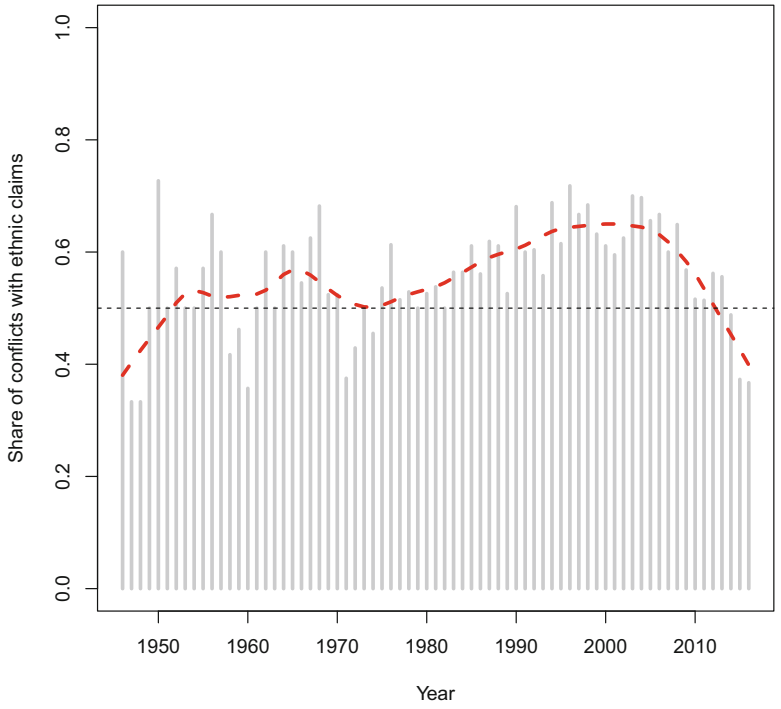

Fig. 4 Share of armed civil conflicts with ethnic claims, based on the ACD2EPR data (Vogt et al. 2015; Wucherpfennig et al. 2012)

conflicts in the Uppsala/PRIO Armed Conflict Data that are deemed to be ethnic, based on the ACD2EPR project linking actors in ongoing armed conflict to ethnic groups in the Ethnic Power Relations data, based on whether organizations make links on behalf of ethnic groups. As can be seen, historically it has been the case that the majority of armed violent conflicts that could be considered ethnic. However, the proportion has recently fallen, and is now less than 50%. One important possible explanation here is that ethnic civil wars have declined precisely since we see less of the ethnic discrimination and exclusion that promotes violence. Cederman et al. (2017) provide evidence that countries with changes toward greater accommodation and inclusion generally have lower rates of subsequent onset and higher likelihood of termination in ongoing conflict than countries that have not seen changes.

If we wish to study potential conflict outside the ethnic realm, the limitations of focusing only on violence become more apparent. The large non-violent campaigns reported in existing data sources tend to be non-sectarian campaigns against authoritarian rulers (Cunningham et al. 2017; White et al. 2015). There are few instances of large-scale direct action involving ethnic groups, although many ethnic groups have relied on various non-violent tactics that do not involve mass mobilization (Cunningham 2013). One might speculate that mass mobilization is more likely to be successful if it can overcome other divisions in a population, as seen in Syria. As such, non-violent forms of contention may be generally less likely to be successful for secessionist aims, and it may be more likely that actors resort to violence precisely when non-violent tactics are the least likely to be effective. Testing these conjectures is very difficult to do adequately with existing data, since they are limited either to violent conflicts or events or only large-scale mobilization. The need to develop better data of incompatibilities and mobilization over these defined independently of the use of violence is one of the major unresolved issues in current conflict research.

Another problem is related to the problem of scale of conflict. Dissent by non-state actors by definition must involve collective action, but the actual level of participation varies dramatically. Yet, many analyses just count events without identifying participation explicitly. For violent conflict, the scale of the conflict is often equated with the number of battle deaths. However, the number of battle deaths is not necessarily a good indicator of participation. For example, one can imagine that a conflict event could be brought to a halt when one antagonist mobilizes superior forces and successfully deters the opponent. Arguably, the number of casualties following the Warsaw Pact invasion of Czechoslovakia in 1968 was limited precisely because Czech First Secretary Dubcek ordered people not to resist the superior invading military forces (see Fraňková 2017).

More generally, participation is an essential feature of interest in its own right, and arguably key to the outcome of the contentious events. There is considerable evidence that activists and organizations seeking to mobilize in dissent see maximizing participation as one of their key objectives—in the words of Popovic (2015, p. 52) "in a nonviolent struggle, the only weapon that you're going to have is numbers".

A standard approach in common research is to count numbers of reported events as a measure of the magnitude of events. The idea is that a situation where we have a higher number of reported events is experiencing a more extensive and significant conflict episode. Although it may well be correct that a conflict with more extensive reporting sees more events, but there is no necessary theoretical or empirical relationship between the number of events and the extent of participation in them. For example, a group that can mobilize a very large number may focus on a single large event, while groups that can only mobilize smaller numbers may carry out many smaller-scale events by the same participants. By counting events, we may erroneously conclude that the latter is a "larger" conflict, even if it involves fewer participants. These examples are not contrived hypothetical examples, but reflect real concerns. Biggs (2018) examines the relationship between the number of strikes and actual participants in strikes, using data from the USA, and shows that the two measures are not highly correlated. Many discussions of event data have been very concerned about the possible selection biases or the problem that smaller events may be omitted due to for example media biases (Weidmann 2016). However, if we scale events by size then it should be easier to get differences in orders of magnitude right, even if there is uncertainty, and we are generally less concerned about the influence of possible noise at the very low end.

In addition to the theoretical problems in using event counts as measure of scale, there are also a number of practical issues arising in delineating what constitutes one "event" as opposed to two or more "separate" events. Many event data projects use different types of "deduplication" efforts to determine whether different reports are to the same or different events, typically considering events to be "the same" if they fall on the same date. However, there is no guarantee that this will work, and it is often the case that report dates may be ambiguous with respect to the day of reporting and the day the events described occurred. Even worse, there is plausibly an inverse relationship between size and granularity in some data sources. The Social Conflict Analysis Data (SCAD) provides a much used event dataset, which extends data on exclusively violent conflict and data limited to large-scale organized non-violent campaigns by providing more detailed event data on social conflict as well as geographical location information (Salehyan et al. 2012). However, many events in the SCAD data are coded as nationwide, where the number of simultaneous events across a country is deemed to be so large that coders can no longer identify exhaustively all the individual events. The nationwide events are likely to have more participants than smaller events that are easy to identify as discrete events, yet analysts counting events may count the former as "less significant" since it is reflected in fewer event counts.

I think these are genuine problems, but in keeping with the theme of theory-data interaction here, I am also relatively optimistic about our ability to find useful approaches to overcome them. With regards to identifying incompatibilities, there is much that can be done to identify conflict constellations using methods such as expert surveys or automated content analyses. For example, recent work on conflict prediction using topic modelling suggests that it may be possible to identify

anti-government claims in news media sources (Mueller and Rauh 2018). Similar types of content analysis techniques could be useful for identifying cleavages or contention more generally, separately from violence or large-scale mobilization. With regards to counting participation, we also have active developments of alternative coding approaches, using photos where we can assess the density or social media data such as twitter via geolocation (Barberá et al. 2015; González-Bailón and Wang 2016; Won et al. 2017). It turns out again that since many things scale, we can use proportionality measure to infer participation. Steinert-Threlkeld (2018) notes that this tracks participation well in so-called women's marches in the USA.

6 Conclusion

In this chapter I have reviewed some examples of the interaction between data development and theoretical progress in the field of conflict research. I hope that I have successfully shown that data in some cases may have preceded theory, but in most cases data have been collected and developed in direct response to initial theoretical beliefs and hunches. However, the availability of data has often led to theoretical re-evaluations and progress; initial hunches may not be fully supported, while other findings lead to new puzzles or research questions. I hope this overview can give some sense of the excitement that I am left with over the progressive nature of interaction between theory and data in conflict research and the maturity of the field. Future central data resources are likely to come from new technologies or sources that have been difficult to use in the past. For example, satellite images are now readily available, and also relatively easy to analyze on a standard computer. Such data can be used to extract information on features for which no meaningful official data exist, such as variation in local income and wealth in countries with poor infrastructure and governance (see Jerven 2013; Weidmann and Schutte 2017). Many sources—including information that was previously classified– can now be extracted from digital sources, rapidly disseminated on the internet, and advances in text analysis and extraction makes it much easier to conduct systematic analysis of such data sources (see, e.g., Biggs and Knauss 2011, Deutschmann 2016). Simulation can provide an important complement to limited observed data, and counterfactual computational analysis can be particularly compelling if it is linked to clear theoretical arguments and grounded in known empirical information (Cederman 1997; Tetlock and Belkin 1996). It is difficult to predict—especially about the future. I make no claim to be able to predict specific new scientific innovations or salient new topics with much confidence, but I am very confident that new data sources and methodologies for data development will figure prominently in a future updated version of a graph of scientific influence in conflict research akin to Van Holt et al. (2016).

A.1 Appendix: Key Contemporary Data Sources, Listed Alphabetically

Armed Conflict Location & Event Data Project (https://www.acleddata.com/). A disaggregated conflict data collection, with dates, actors, types of violence, locations, and fatalities of reported political violence and protest events. The ACLED data are not global, but cover a number of countries in Africa, Asia, and the Middle East.

Correlates of War Project (http://www.correlatesofwar.org/). Provides access to episodic data on interstate wars and militarized interstate disputes. The COW project also collects data on various state-based characteristics such as military capabilities and diplomatic ties between states.

Global Database of Events, Language, and Tone (https://www.gdeltproject.org/). Provides access to machine coded event data from electronic sources from 1979 to the present, using the Conflict and Mediation Event Observations (CAMEO) coding scheme.

Global Terrorism Database (https://www.start.umd.edu/gtd/). Provides access to data on terrorist attacks since 1970, as well as some supplementary data sources on terrorist group profiles.

Integrated Crisis Early Warning System (https://dataverse.harvard.edu/dataverse/icews). Daily event data coded from electronic news sources, with actor, event, and location identifiers. Note that the most recent public version of the data has a 1 year embargo.

Phoenix [Cline Center Historical Phoenix Event Data] (https://clinecenter.illinois.edu/project/machine-generated-event-data-projects/phoenix-data). Event data for the period 1945–2015, machine coded from 14 million news stories from the New York Times (1945–2005), the BBC Monitoring's Summary of World Broadcasts (1979–2015) and the CIA's Foreign Broadcast Information Service (1995–2004).

Phoenix [Real time Phoenix data] (http://eventdata.utdallas.edu/data.html). A real time machine coded event dataset complementing the historical data, available from October 2017.

Non-violent and Violent Campaigns and Outcomes (https://www.du.edu/korbel/sie/research/chenow_navco_data.html). Provides access to an influential dataset that also documents non-violent mobilization over maximalist claims on a government.

Social Conflict Analysis Database (https://www.strausscenter.org/scad.html). Provides access to event data on protests, riots, strikes, inter-communal conflict, government violence against civilians, and other forms of social conflict not systematically tracked in other conflict datasets. SCAD currently includes information on social conflicts from 1990–2017, covering all of Africa and now also Mexico, Central America, and the Caribbean.

Uppsala Conflict Data Program (https://ucdp.uu.se/downloads/). Provides access to data on various types of violent conflicts, including state-based interstate and intrastate conflict, violence against civilians, and non-state/inter-communal conflict, as well as geo-referenced event data.

References

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., et al. (2015). The critical periphery in the growth of social protests. *PLoS One, 10*(11), e0143611. https://doi.org/10.1371/journal.pone.0143611.

Beck, N., Katz, J. N., & Tucker, R. M. (1998). Taking time seriously: Time-series cross-section analysis with a binary dependent variable. *American Journal of Political Science, 42*(4), 1260–1288.

Biggs, M. (2018). Size matters: Quantifying protest by counting participants. *Sociological Methods and Research, 47*(3), 351–383.

Biggs, M., & Knauss, S. (2011). Explaining membership in the British National Party: A multilevel analysis of contact and threat. *European Sociological Review, 28*(5), 633–646.

Bohorquez, J. C., Gourley, S., Dixon, A. R., Spagat, M., & Johnson, N. F. (2009). Common ecology quantifies human insurgency. *Nature, 462*, 911.

Boulding, K. E. (1963). *Conflict and defense: A general theory*. New York: Harper and Row.

Bremer, S. A. (1992). Dangerous dyads: Conditions affecting the likelihood of interstate war, 1816-1965. *Journal of Conflict Resolution, 36*(2), 309–341.

Bremer, S. A., & Mihalka, M. (1977). Machiavelli in Machina: Or politics among hexagons. In K. W. Deutsch, B. Fritsch, H. Jaguaribe, & A. S. Markovits (Eds.), *Problems of world modeling: Political and social applications* (pp. 303–337). Cambridge, MA: Ballinger.

Cederman, L. E., & Gleditsch, K. S. (2009). Special issue on 'disaggregating civil war'. *Journal of Conflict Resolution, 53*(4), 487–495.

Cederman, L.-E. (1997). *Emergent actors in world politics: How states and nations develop and dissolve*. Princeton, NJ: Princeton University Press.

Cederman, L.-E. (2003). Modeling the size of wars: From billiard balls to Sandpiles. *American Political Science Review, 97*(1), 135–150.

Cederman, L.-E., Gleditsch, K. S., & Wucherpfennig, J. (2017). Predicting the decline of ethnic civil war: Was Gurr right and for the right reasons? *Journal of Peace Research, 54*(2), 262–274.

Chenoweth, E., & Stephan, M. J. (2011). *Why civil resistance works: The strategic logic of nonviolent conflict*. New York: Columbia University Press.

Cirillo, P., & Taleb, N. N. (2016). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications, 452*(15), 29–45.

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances, 4*(2), eaao3580. https://doi.org/10.1126/sciadv.aao3580.

Clauset, A., & Woodard, R. (2013). Estimating the historical and future probabilities of large terrorist events. *Annals of Applied Statistics, 7*(4), 1838–1865.

Clauset, A., Young, M., & Gleditsch, K. S. (2007). On the frequency of severe terrorist events. *Journal of Conflict Resolution, 51*(1), 1–31.

Cunningham, D. E., Gleditsch, K. S., González, B., Vidovic, D., & White, P. B. (2017). Words and deeds: From incompatibilities to outcomes in anti-government disputes. *Journal of Peace Research, 54*(4), 468–483.

Cunningham, D. E., Gleditsch, K. S., & Salehyan, I. (2009). It takes two: A dyadic analysis of civil war duration and outcome. *Journal of Conflict Resolution, 53*(4), 570–597.

Cunningham, K. G. (2013). Understanding strategic choice: The determinants of civil war and non-violent campaign in self-determination disputes. *Journal of Peace Research, 50*(3), 291–304.

Deutschmann, E. (2016). Between collaboration and disobedience: The behavior of the Guantánamo detainees and its consequences. *Journal of Conflict Resolution, 60*(3), 555–582.

Donnay, K., Dunford, E. T., McGrath, E. C., Backer, D., & Cunningham, D. E. (2019). Integrating conflict event data. *Journal of Conflict Resolution, 63*(5), 1337–1364.

Fazal, T. M. (2016). An occult of irrelevance? Multimethod research and engagement with the policy world. *Security Studies, 25*(1), 34–41.

Fraňková, R. (2017). *Historians pin down number of 1968 invasion victims*. Radio Praha.

Gaddis, J. (2005). *The cold war: A new history*. London: Penguin.

Gerner, D. J., Schrodt, P. A., & Francisco, R. A. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly, 38*, 91–119.
Gleditsch, K. S. (2004). A revised list of wars between and within independent states, 1816-2002. *International Interactions, 30*(4), 231–262.
Gleditsch, K. S., & Clauset, A. (2018). Trends in conflict: What do we know and what can we know? In W. Wolforth & A. Gheciu (Eds.), *Oxford handbook of international security*. New York/Oxford: Oxford University Press.
Gleditsch, K. S., Metternich, N. W., & Ruggeri, A. (2014). Data and progress in peace and conflict research. *Journal of Peace Research, 51*(3), 301–314.
Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., & Strand, H. (2002). Armed conflict 1946-2001: A new dataset. *Journal of Peace Research, 39*(5), 615–637.
Goldstein, J. S. (2011). *Winning the war on war*. Hialeah, FL: Dutton/Penguin.
González-Bailón, S., & Wang, N. (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks, 44*, 95–104.
Hjort, N. L. (2018). Towards a More Peaceful World [Insert '!' Or '?' Here]. Vol.: https://www.mn.uio.no/math/english/research/projects/focustat/the-focustat-blog%21/krigogfred.html
Hobbes, T. (1651). *Leviathan*. London: Andre Crooke. Retrieved from https://socialsciences.mcmaster.ca/econ/ugcm/3ll3/hobbes/Leviathan.pdf.
Høivik, T., & Galtung, J. V. (1971). Structural violence: A note on operationalization. *Journal of Peace Research, 7*, 73–76.
Jaggers, K., & Gurr, T. R. (1995). Tracking democracy's third wave with the polity III data. *Journal of Peace Research, 32*(4), 469–482.
Jerven, M. (2013). *Poor numbers: How we are misled by African development statistics and what to do about it*. Ithaca, NY: Cornell University Press.
Jones, D. M., Bremer, S. A., & David Singer, J. (1996). Militarized interstate disputes, 1816-1992: Rationale, coding rules, and empirical applications. *Conflict Management and Peace Science, 15*(2), 163–213.
Kelvin, L. [Thompson, W]. (1883). Electrical units of measurement. In *Popular lectures* (Vol. I, pp. 73–136). Cambridge: Cambridge University Press.
Kinne, B. J. (2009). *Beyond the dyad: How networks of economic interdependence and political integration reduce interstate conflict*. (Doctoral Dissertation, Department of Political Science, Yale University).
Lichbach, M. I. (1995). *The Rebel's dilemma*. Ann Arbor, MI: Michigan University Press.
Lugo-Galicia, H. (2017). *El país gritó: "Maduro, no te queremos*". Retrieved from http://www.el-nacional.com/noticias/politica/pais-grito-maduro-queremos_178023
Mansfield, E. (1988). The distribution of wars over time. *World Politics, 41*, 21–51.
Miranda, L. C. M., Perondi, L. F., & Gleditsch, K. S. (2016). The evolution of civil war severity, 1816-2005. *Peace Economics, Peace Science and Public Policy, 22*(3), 247–276.
Most, B. A., & Starr, H. (1982). Case selection, conceptualizations and basic logic in the study of war. *American Journal of Political Science, 26*(4), 834–856.
Most, B. A., & Starr, H. (1983). Conceptualizing "war": Consequences for theory and research. *Journal of Conflict Resolution, 27*(1), 137–159.
Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review, 112*(2), 358–375.
Oneal, J. R., & Russett, B. M. (2001). *Triangulating peace: Democracy, interdependence, and international organizations*. New York: Norton.
Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. New York: Viking.
Popovic, S. (2015). *Blueprint for revolution: How to use Rice pudding, Lego men, and other nonviolent techniques to galvanize communities, overthrow dictators, or simply change the world*. New York: Random House.
Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research, 47*(5), 651–660.

Richardson, L. F. (1948). Variation of the frequency of fatal quarrels with magnitude. *Journal of the American Statistical Association, 43*(244), 523–546.

Richardson, L. F. (1960). *Statistics of deadly quarrels*. Chicago IL/Pittsburgh, PA: Quadrangle/Boxwood.

Salehyan, I., Hendrix, C. S., Hamner, J., Case, C., Linebarger, C., Stull, E., & Williams, J. (2012). Social conflict in Africa: A new database. *International Interactions, 38*(4), 503–511.

Sandler, T. (1992). *Collective action: Theory and applications*. Ann Arbor, MI: University of Michigan Press.

Schrodt, P. A., & Gerner, D. J. (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982-92. *American Journal of Political Science, 38*, 825–854.

Sharp, G. (1973). *The politics of nonviolent action*. Boston: Porter Sargent.

Simowitz, R. (1996). *Scientific Progress in the Democracy-War Debate*. Paper Presented at the Annual Connvention of the International Studies Association, San Diego, CA.

Singer, J. D. (Ed.). (1980). *The correlates of war ii: Testing some realpolitik models*. New York: Free Press.

Sorokin, P. A. (1957[1937]). *Social and cultural dynamics*. London: Owen.

Starr, H., & Most, B. A. (1983). Contagion and border effects on contemporary African conflict. *Comparative Political Studies, 16*, 92–117.

Steinert-Threlkeld, Z. C. (2018). *Twitter as data*. Cambridge: Cambridge University Press.

Sundberg, R., & Melander, E. (2013). Introducing the UCDP georeferenced event dataset. *Journal of Peace Research, 50*(4), 523–532.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Tetlock, P. E. (1999). Theory-driven reasoning about possible pasts and probable futures in world politics: Are we prisoners of our preconceptions? *American Journal of Political Science, 43*(2), 335–366.

Tetlock, P. E., & Belkin, A. (Eds.). (1996). *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton, NJ: Princeton University Press.

Thucydides. (2000). *The history of the Peloponnesian war*. London: Penguin.

Tollefsen, A. F., Strand, H., & Buhaug, H. (2012). Prio-Grid: A unified spatial data structure. *Journal of Peace Research, 49*(2), 363–374.

Tullock, G. (1971). The paradox of revolution. *Public Choice, 11*(1), 88–99.

Van Holt, T., Johnson, J. C., Moates, S., & Carley, K. M. (2016). The role of datasets on scientific influence within conflict research. *PLoS One, 11*(4), e0154148.

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P., & Girardin, L. (2015). Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family. *Journal of Conflict Resolution, 59*(7), 1327–1342.

Ward, M. D., Greenhill, B., & Bakke, K. M. (2010). The perils of policy by P-value: Predicting civil conflicts. *Journal of Peace Research, 47*(5), 363–375.

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

Weidmann, N. B., & Salehyan, I. (2013). Violence and ethnic segregation: A computational model applied to Baghdad. *International Studies Quarterly, 57*(1), 52–64.

Weidmann, N., & Schutte, S. (2017). Using night lights for the prediction of local wealth. *Journal of Peace Research, 54*(2), 125–140.

White, P., Vidovic, D., Gonzalez, B., Gleditsch, K. S., & Cunningham, D. (2015). Nonviolence as a weapon of the resourceful: From claims to tactics in mobilization. *Mobilization: An International Journal, 20*(4), 471–491.

Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM International Conference on Multimedia 2017*. New York: ACM.

Wright, Q. (1942/1965). *A study of war*. Chicago, IL: University of Chicago Press.
Wucherpfennig, J., Metternich, N., Cederman, L.-E., & Gleditsch, K. S. (2012). Ethnicity, the state, and the duration of civil wars. *World Politics, 64*(1), 79–115.
Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., & Sanguinetti, G. (2012). Point process modelling of the Afghan War Diary. *Proceedings of the National Academy of Sciences, 109*(31), 12414–12419.