Insurgency and Small Wars: Estimation of Unobserved Coalition Structures

Francesco Trebbi and Eric Weese*

March 2018

Abstract

Insurgency and guerrilla warfare impose enormous socio-economic costs and often persist for decades. The opacity of such forms of conflict is an obstacle to effective international humanitarian intervention and development programs. To shed light on the internal organization of otherwise unknown insurgent groups, this paper proposes two methodologies for the detection of unobserved coalitions of militant factions in conflict areas. These approaches are based on daily geocoded incident-level data on insurgent attacks. We provide applications to the Afghan conflict during the 2004-2009 period and to Pakistan during the 2008-2011 period, identifying systematically different coalition structures. Applications to global terrorism data and identification of new groups or shifting coalitions are discussed.

*University of British Columbia, Vancouver School of Economics, CIFAR and NBER, francesco.trebbi@ubc.ca; Kobe University, weese@econ.kobe-u.ac.jp respectively. The authors would like to thank Eli Berman, Ethan Bueno de Mesquita, James Fearon, Patrick Francois, Camilo Garcia-Jimeno, Jason Lyall, Carlos Sanchez-Martinez, Jake Shapiro, Drew Shaver, Austin Wright and seminar participants at UCSD, Penn, Chicago Harris, Columbia, Stanford, Berkeley, Rochester, UQAM, Kobe, Tokyo, Toronto, CIFAR, and Princeton for useful comments and discussion and the researchers at the Princeton University Empirical Studies of Conflict Project for generously sharing their incident data online. Nathan Canen provided excellent research assistance. We are grateful to the Social Science and Humanities Research Council for financial support.

1 Introduction

Among the many political sources of welfare loss, few compare in magnitude to military conflict and, in the post World War II period in particular, to the losses ascribed to civil war and insurgency [O'Neill, 1990; Collier, 2007]. Insurgency, defined as armed rebellion against a central authority,[1] is also one of the most opaque forms of conflict. Intertwining connections with the population blur the lines between combatants and civilians [Kilcullen, 2009]. The relative strength and even the identity of potential negotiating counterparties are often unclear, and in the words of Fearon [2008] "*there are no clear front lines.*" Such forms of conflict have disproportionately affected poor countries and are gaining central status in the literature on the political economy of development.[2]

In this paper we provide methods of identifying the unknown structure of insurgent groups and then use these methods to detect changes in group structure beyond what is reported in the best publicly available datasets. When faced with violent incidents in multiple regions, we show how to recover unknown insurgent groups' number and location from microlevel attack data. We demonstrate how these methods translate into applications such as the detection of shifts in alliances and the emergence of divisions among insurgent groups over time, the early identification of unknown rebel groups, and the validation of geographical information on group boundaries.

Empirically, we focus on the costly insurgencies of Afghanistan and Pakistan.[3] On the U.S. side alone, Afghan operations cost the lives of more than $1,800$ troops between 2001 and 2011, and more than \$444 billion in military expenses. Statistics for Afghan civilians appear less certain, but the adverse effects are painfully obvious even to the casual observer. In Pakistan, sectarian insurgencies have led to $18,583$ dead and $19,356$ wounded between 2012 and 2015 alone [Pak Institute for Peace Studies, 2016], and have diverted resources away from development assistance programs and public goods provision.

In Pakistan the ethnic basis of insurgent groups is well known. In Afghanistan,

[1] According to O'Neill [1990] "*Insurgency may be defined as a struggle between a nonruling group and the ruling authorities in which the nonruling group consciously uses political resources (e.g., organizational expertise, propaganda, and demonstrations) and violence to destroy, reformulate, or sustain the basis of one or more aspects of politics.*"

[2] See Blattman and Miguel [2010], Berman and Matanock [2015], and König et al., [2017].

[3] Unfortunately, we are not aware of suitable data for Syria. In Appendix A we further discuss the case cases of Iraq, Syria, and Libya, all instances where our methodologies could be potentially of use.

however, there is a serious disagreement about whether the Taliban was a unified organization or whether it was rather an umbrella coalition of heterogeneous forces during 2004-2009. Some policymakers and researchers were skeptical of the degree of control that Taliban leader Mullah Mohammed Omar exerted over the powerful Haqqani Network and the Dadullah Front.[4] Other observers, however, promoted an opposite view. In an insightful qualitative essay Dorronsoro [2009] states: "*The Taliban are often described as an umbrella movement comprising loosely connected groups that are essentially local and unorganized. On the contrary, this report's analysis of the structure and strategy of the insurgency reveals a resilient adversary, engaged in strategic planning and coordinated action".* Evidence in support of this position includes the existence of the *Layha* (a centralized code of conduct for Mujahidin), as well as the strong centralizing tendencies of the *Obedience to the Amir* (a manual endorsed by Mullah Omar).[5]

A quantitative researcher faced with this disagreement might think of turning to a standard database, such as the Worldwide Incidents Tracking System (WITS) or the Global Terrorism Database (GTD), as both sources report (when possible) the group that perpetrated each attack. For 2004-2009, both WITS and the GTD code most attacks in Afghanistan as undifferentiated "Taliban". The question, however, is not whether the Taliban exist – this is not in dispute – but rather whether they represent multiple agents, several independent decision makers loosely connected in name but without a unitary strategy. Although WITS data ends in 2009, the GTD continues to code attacks in 2010-2016 simply as "Taliban". Using our method, we show that while the 2004-2009 phase of the insurgency is best described by a unitary Taliban, around 2013 the internal structure of the Taliban fragments into multiple groups.

[4] Christia [2012] indicates 4 warring groups in Afghanistan for 2002-2012. Smith [2005] believes that the Taliban are "not a single monolithic movement, but a series of parallel groupings." Christia and Semple [2009] state explicitly that "*the Taliban is not a unified or monolithic movement*" and Thruelsen [2010] writes that "*the movement should not be seen as a unified hierarchical actor that can be dealt with as part of a generic approach covering the whole of Afghanistan.*"A UN report [2013] stated that *"despite what passes for a zonal command structure across Afghanistan, the Taliban have shown themselves unwilling or unable to monopolize anti-State violence. The persistent presence and autonomy of the Haqqani Network and the manner in which other, non-Taliban, groupings like the Lashkar-e-Tayyiba are operating in Afghanistan raises questions about the true extent of the influence exerted by the Taliban leadership."* Brahimi [2010] reports a statement by Ashraf Ghani, current Afghan president, in a lecture for the Miliband Programme at LSE indicating *"The Taliban are not a unified force – they are not the SPLA in Sudan or the Maoists in Nepal"* while Giustozzi [2009] states that "*the Taliban themselves are not fully united and the insurgency is not limited to the Taliban."*

[5] These are available in English translation as Munir [2011] and Ludhianvi [2015], respectively.

We do not employ data on the identity of perpetrators coded in the standard sources, but instead use the informational content of certain types of attacks to reach our conclusions. Insurgent groups with the ability to launch simultaneous and geographically separated attacks appear to do so: we make use of the specific covariance structure that arises from these attacks. Our model thus relies on conclusions from the existing literature about the propaganda and costly signaling value of launching complex coordinated attacks – a mechanism that finds tragic support in the psychological impact of events such 9/11, Mumbai, or the Bataclan.

We present our econometric model in Section 2. A country experiencing an insurgency is described as a set of districts in which violent incidents can occur each day. Attacks on the same day in different districts will occur with greater-than-random frequency if the same insurgent group is operating in these areas. Using a variety of assumptions regarding what the reference cross-district covariance in attacks would be in the case where there were no organized groups, we calculate which sets of districts are more correlated than would be expected by chance alone. We then use this information to estimate the cluster of districts in which each group operates. Our proposed method complements and advances the literature on community detection in networks [Copic, Jackson, and Kirman, 2009]. It is also related to the identification of spatial patterns of interaction, such as the Ellison and Glaeser [1997] study of economic agglomeration.[6]

We present estimation methods that allow for a single district to be contested by multiple groups, or by a single group, or by no groups. Our methods provide the number of insurgent groups operating, the geographic area of each of these, and the intensity of each group's activity in each district. The methods we present can accommodate slow-moving trends in violence over time, and are robust to aggregate shocks (such as weather, seasonality, or U.S. troop movements) that might affect insurgent activity in many areas simultaneously.

We apply this method to data – described in Section 3 – on attacks in Afghanistan

[6] Ellison and Glaeser [1997] develop an index to evaluate whether a given industry is more geographically clustered than would be expected had plants been located at random. They then present an index that can be used to consider whether two or more industries "coagglomerate" more than would be expected by random chance. The availability of information about supply chain relationships across industry pairs (districts in our case), a crucial input in their approach, is a major difference from our approach (where such latent relationships are not available and have to be estimated). Duranton and Overman [2005] use permutation tests to assess the statistical significance of spatial clustering that they observe in a Ellison and Glaeser [1997] type model, in their case considering individual industries in the UK.

and Pakistan. We present the main results of this analysis in Section 4. We find that for the 2004-2009 period insurgent activity in Afghanistan is best represented by a single organized group rather than several independent groups. This result is robust to constraining the analysis to districts with a number of incidents above specific thresholds, to controlling for religious holidays, and to limiting the analysis to incidents explicitly claimed by the Taliban. We also conduct an analysis of Pakistani insurgent attacks, using multiple data sets for the period 2008-2011. In the case of Pakistan, our methodology detects multiple insurgent groups (four, in fact) and is consistent with the extant qualitative literature on insurgency in the country. We confirm our results for both Afghanistan and Pakistan using GTD data.

Our approach can detect changes in insurgent organization not visible in the group coding reported in standard publicly available sources. In Section 5 we show evidence that one insurgent group in Pakistan (the Sindhudesh Liberation Army) can be identified by our methods as a coordinated entity almost one year before GTD indicates any presence of coordinated attacks in Sindh. Using our tests, we also report quantitative evidence of increasing fragmentation of the Afghan Taliban in the 2010-2016 period. This change is not visible in standard database coding, such as perpetrator group entries in GTD data. Our findings in this case potentially offer a simple explanation for the disagreement in the qualitative literature discussed above: early on the Taliban are basically a unified fighting force, but they become increasingly fragmented in the aftermath of the death of their historical leader, Mullah Omar.

According to our econometric model, random attacks that occur without any coordination should be approximately Poisson distributed. In Section 6, we look for overdispersion relative to this Poisson in order to determine the extent and importance of insurgent coordination in attacks. We extend this analysis to the full panel of 162 countries with geocoded attacks available in the GTD.

An increasing amount of attention has been devoted within the fields of development economics and political economy to the study of armed conflict within countries, in particular civil wars and insurgency. Both Political Science and Economics have provided some of the most recent and novel insights in the study of insurgency.[7] As underlined by Blattman and Miguel [2010], a remarkable characteristic of this recent wave of research has been a strong empirical bent and an increasing attention to mi-

[7] These include Berman [2009], Berman et al. [2011], Condra and Shapiro [2012], and Bueno de Mesquita [2013]. Economists have been interested in the analysis of violence and conflict at least as far back as Schelling [1960] and Tullock [1974].

crolevel (typically incident-level) information. The use of geocoded micro data in this area is a departure from more established “macro” empirical approaches, which were based on country-level information or aggregate conflict information.

This paper is one in the new “micro” style, with a specific emphasis on the analysis of insurgency and small wars. Economic and statistical evidence on the role of anti-government guerrilla activities is still sparse, even though such activities cause substantial damage worldwide and appear from a quantitative perspective to be the predominant form conflict in civil wars since 1945 [Fearon, 2008; Ghobarah et al., 2003]. Insurgents’ strategies are generally not well understood, and neither are the subtleties of their interactions with the noncombatant population [Gutierrez-Sanin, 2008; Kilcullen, 2009], nor their economic incentives [Berman et al. 2017]. A particular incentive for further study is that insurgent activity is also often linked to terrorist activities, and thus there is a connection with the growing literature on the economics of terrorism [Bueno de Mesquita and Dickson, 2007; Benmelech, Berrebi, Klor, 2012]. Some of this recent wave of research has also taken the direction, which we share, of focusing on the network structure of conflict. One such recent example is the work on the complex alliances in the post-Mubutu Congo wars studied by König et al. [2017], while Horowitz and Potter [2014] focus on terrorist alliances and effects of attacks. Importantly, both these papers take in their main analysis the organization of groups’ alliances and their structure as given,[8] while our work explicitly does not. In fact, we show how our methods can be used as an instrument of validation of data on alliances.

Our work is most related to the conflict studies literature focused on the internal organization of insurgent and terrorist groups. Berman [2009] offers an analysis of the internal management of defection risk within terrorist groups. Alliance formation in 1978-1998 Afghanistan is studied in Christia [2012].[9] What seems clear from the conflict literature is that studying group organization may offer a useful perspective in understanding insurgency. This is the main goal of the paper.

[8] König et al. [2017] include an extension where robustness to endogenous link formation is provided.

[9] With specific emphasis on Afghanistan, Christia [2012, ch.3-5] studies in detail the dynamics of ethnic alliances pre 9/11. Staniland [2014, ch.5] discusses the Taliban organization in the periods 1994-2001 and post 9/11. While the author discusses the Taliban as an “integrated organization” with a defined centralized structure and unique leadership (p. 136-137), the presence of the Haqqani Network and of Hezb-i Islami is also considered.

2 Model

This paper relies on the fact that an insurgent group launches simultaneous attacks when it is possible for it to do so. Simultaneous attacks are challenging to coordinate, particularly because perpetrators risk detection by government agents as they communicate regarding the attack: Shapiro [2013] details at length the trade-offs that arise in limiting communications in order to avoid government forces. We thus believe that these attacks are propaganda designed to signal an insurgent group's strength. In Appendix B we consider a setup where the strength of insurgent groups is unobserved, and show how simultaneous attacks correspond to the signals in a Spence [1973] type signalling model.

Below, we take as given the fact that insurgent groups launch simultaneous attacks, and present an empirical model of insurgent activity designed to identify a set of key parameters regarding the latent organization of insurgent groups. First, we discuss how to decompose the covariance matrix of insurgent attacks into a form useful for estimation and the rationale for doing so. Next, we describe two estimation approaches: the first assumes that exactly one insurgent group is present in each district, while the second relaxes this assumption. We conclude by demonstrating how we can avoid potential bias from long-term trends in attacks across districts. Avoiding bias from weather, seasonal fluctuations, and slow moving time trends is important when applying our methods to actual data, but we discuss these details last in order to simplify the exposition.

2.1 Insurgent Attacks

Let districts be indexed by i, and let there be a total of N districts in which attacks occur. Violent occurrences in i can be of two types: unorganized or organized by an insurgent group. We make a distinction between attacks initiated by unorganized local militants and those initiated by members of an organized group, because we wish to allow for the possibility that there are no organized insurgent groups present in a district even though attacks may be observed there.

Let $\ell_i \geq 0$ be the number of unorganized local militants in district i. Let organized insurgent groups be indexed by j, and let $J \geq 0$ be the total number of such organized groups active anywhere in the country. Let $\alpha_{ij} \geq 0$ be the number of members in district i belonging to organized group j. Time is discrete and indexed by t. In our analysis below, the time periods used will be days. This high-frequency attack data is useful because it reduces the number of attacks that are simultaneous simply by random chance, a feature whose importance will be clear in what follows.

In each time period, the probability that an unorganized local militant launches an attack is η, which does not change across time (this assumption is relaxed in Section 2.5). The decision by unorganized militants to attack is independent of the decision of anyone else (unorganized militant or group member). The expected number of attacks by local militants in district i at time t is thus $\eta\ell_i$, and the variance within district i is $\eta(1-\eta)\ell_i$. The covariance in these attacks between two districts i and i' is zero: the attack decisions are made independently, and the probability of an attack is constant.

In contrast to unorganized militants, members of an organized group are more likely to attack on some particular days than on others. Let ϵ_{jt} be the probability that a member of group j will attack at time t. This probability is the same for all members of group j, and whether any given member attacks is independent of other attack decisions after conditioning on the attack probability ϵ_{jt}. As ϵ_{jt} does not vary across districts at t, this process will induce coordination in group j attack behavior.

Across time, the covariance of attacks between two members of the same group is $\text{Var}(\epsilon_{jt})$. We assume that this variance is constant across groups, and will refer to it as σ^2.[10] Assume that for any other group j', ϵ_{jt} is uncorrelated with $\epsilon_{j't}$. Thus, the covariance of attacks between two members of different groups is zero. This is also to say that, if two groups j and j' can coordinate in their attacks and do so *systematically at daily frequency*, we will consider them *de facto* the same organized group.[11] If the extent of such coordination decreases over time or disappears entirely (because groups split), our approach will be able to pick this up, provided sufficient microlevel data is available.

Consider the members of group j. Define x_{it} as the total number of attacks at time t in district i. If there are α_{ij} members in district i and $\alpha_{i'j}$ members in district i', then the covariance in attacks over time between these two districts, due to the presence of

[10] This is with limited loss, as district specific heteroskedasticity will not be separately identifiable from variation in group-specific parameters α_{ij}.

[11] Besides the organizational complexity of matching day after day the attack behavior of another group, there is also a potential loss of signalling value of the group identity, as noncombatants will be uncertain about whether it is j or j' launching the attacks. In one of the few quantitative studies of terrorist alliances Horowitz and Potter [2014] emphasize how the main gains from alliances stem from shared capabilities and technology for attack (for improvised explosive devices or suicide attacks), and they do not mention the possibility of using dog-whistles to engage in simultaneous attacks. We do not exclude that this may happen occasionally, only that it does not happen constantly.

members of group j, is $\sigma^2\alpha_{ij}\alpha_{i'j}$. Summing over members of all groups, the covariance in attacks between districts i and i' will be $\mathrm{Cov}(x_{it}, x_{i't}) = \sigma^2 \sum_j \alpha_{ij}\alpha_{i'j} \geq 0$.

The setup just presented is clearly a stylized representation of the attack behavior of insurgent groups, and the covariance structure imposed is not without loss of generality. A particularly strong assumption made in the model is that the members of an insurgent group do not move between districts: a given group j has a certain membership α_{ij} in district i, and those members will either be encouraged to attack in a given period (a high ϵ_{jt}), or not (low ϵ_{jt}).

A very different model would be one in which members of an insurgent group are mobile, and in any given period have the choice of attacking in one of many districts. This latter model implies that organized groups should lead to negative covariances between districts, as insurgent group members who attack in district i could not also be attacking in district i' in the same period. In contrast, the model presented above suggests that $\mathrm{Cov}(x_{it}, x_{i't})$ should be positive if at least one insurgent group j has members in both i and i', as attacks in both i and i' will be higher in periods when ϵ_{jt} is high and lower in periods when ϵ_{jt} is low; and zero otherwise. Consistently with these assumptions, in the data the observed covariances $\mathrm{Cov}(x_{it}, x_{i't})$ are systematically non-negative.[12] The qualitative research of Deloughery [2013] and others also suggests that a model without substantial substitution in attacks across districts appears most appropriate.

We now present our estimation methods for parameters describing the number of insurgent groups and their presence across locations. We first present a decomposition of the covariances just discussed. We then assume that insurgent groups are non-overlapping, and use this decomposition as part of an algorithm that involves hierarchical splits of our set of districts. We then consider the case where groups are potentially overlapping, and show how our covariance decomposition can be used as the starting point for a non-negative matrix factorization algorithm. The approach based on clustering and that based on matrix factorization are largely complementary in that they rely on different assumptions. This provides a form of cross-validation for our results, which is important given the novelty of these methodologies within the field of political economy.[13]

[12] Permutation tests of the sort discussed later indicate that the mean covariance is positive at any reasonable confidence level. Results available upon request.

[13] In the working paper version of this paper we used an approach based on spectral clustering (see Luxburg [2007]). We discuss this approach, and reasons why our current approach may be preferable,

2.2 Covariance Decomposition

Let Γ be the covariance matrix for the attacks discussed in Section 2.1, where the entry in row i and column i' gives the covariance in attacks across time for these two districts. Our analysis will be based on this matrix, and others created from it. The covariance matrix Γ can be decomposed as $\Gamma = \Gamma_D + \Gamma_L$, where Γ_D is a diagonal matrix and Γ_L is a low rank matrix of the form

$$\Gamma_L = \sigma^2 \begin{bmatrix} \sum_j \alpha_{1j}\alpha_{1j} & \sum_j \alpha_{1j}\alpha_{2j} & \cdots & \\ \sum_j \alpha_{2j}\alpha_{1j} & \sum_j \alpha_{2j}\alpha_{2j} & & \\ \cdots & & \sum_j \alpha_{ij}\alpha_{ij} & \\ & & & \cdots \end{bmatrix}. \tag{1}$$

This decomposition is considered because the diagonal entries of the covariance matrix are a sum of variance from unorganized militants and variance from organized groups, and only the latter is of interest.[14]

As a normalization, we set $\sigma^2 = 1$.[15] We do not observe Γ_L but we can proceed by using an estimate $\hat{\Gamma}_L$: details regarding this estimate are provided in Appendix C. Because of this decomposition, we effectively do not use pure random uncoordinated violence as part of our procedure for the detection of groups: random violence acts only as noise and is not used as part of the estimator. Thus, areas where some group adherents may be present, but that group has no organizational capacity will not be detected by our method.[16]

2.3 Non-overlapping Insurgent Groups

We desire both an estimate $\hat{J}$, the total number of organized insurgent groups, as well as an estimate $\hat{\alpha}_{ij}$ for each district i and group j, giving the number of insurgent members of the group operating in that district. The set of estimates $\{\hat{\alpha}_{ij}\}$ will have a

in Appendix D.

[14] The diagonal entries of Γ do not in general have a useful form. The situation even with mixture Poisson distributions does not appear to be simple: see Ashford and Hunt [1973] for the Poisson-Gamma distribution, and Karlis and Xekalaki [2005] for mixture Poisson distributions in general.

[15] See Appendix E for a discussion of potential issues if σ_j^2 is in fact different for different districts or groups.

[16] For example, ISIS may have successfully radicalized some individuals in North America and convinced them to conduct attacks, but these were all lone wolf style attacks, because ISIS does not have the infrastructure to safely coordinate attacks in North America.

total of $N \times \hat{J}$ elements. It turns out to be easiest to first produce the $\{\hat{\alpha}_{ij}\}$ estimates for various values of $J \in \{1, ..., J_{\max}\}$, and then choose a $\hat{J}$ based on examining this set of estimates.[17] We will thus begin by assuming that J is known, and consider how to compute estimates $\{\hat{\alpha}_{ij}\}$ given J. After this, we will then consider how to choose $\hat{J}$.

Estimation via standard clustering techniques requires an additional assumption different from those that will be needed for Section 2.4. Specifically, it is necessary to assume that the various insurgent groups present do not have overlapping territories. That is, there is one organized group j present in any given district i.[18] Based on this assumption, reordering the districts i allows Γ_L to be written as a block-diagonal matrix:

$$\Gamma_L = \begin{bmatrix} \Gamma_L^1 & ... & 0 \\ ... & \Gamma_L^j & ... \\ 0 & ... & \Gamma_L^J \end{bmatrix}. \tag{2}$$

Here there are a total of J organized groups, and each block Γ_L^j has the form given in Equation 1. To produce estimates $\{\hat{\alpha}_{ij}\}$ we will first determine which organized group is present in each district, and then we determine the strength of this group in the district.

To determine which organized group is present in each district, we will follow a modified k-means type approach. Begin by constructing a scaled version of Γ_L:

$$\Gamma_L^{\text{cor}} = D(\sum_j \alpha_{.j}\alpha_{.j})^{-1/2}\Gamma_L D(\sum_j \alpha_{.j}\alpha_{.j})^{-1/2}, \tag{3}$$

where $D(.)$ indicates a diagonal matrix with the specified vector on the diagonal. This process is occasionally referred to as "sphering" and it often improves the quality of the clustering. By assumption, for each district i, $\alpha_{ij} = 0$ for all but one group j, and thus Γ_L^{cor} is constructed by dividing row i and column i of Γ_L by the value of α_{ij} for the single group j that is present in i.

The "cor" superscript is used because Γ_L^{cor} is positive semi-definite with all diagonal entries equal to one, and thus has the form of a correlation matrix. However, observe

[17] The exact choice of $J_{\max}$ is not important.

[18] This is a direct consequence of the standard assumption that factors should be orthogonal, combined with the fact that insurgent prevalence α must be non-negative.

that each off-diagonal entry $\gamma_{ii'}$ has now been divided by $\alpha_{ij}\alpha_{i'j}$ if the same insurgent group is present in districts i and i'. Thus, in exactly the same way as (2), after suitable rearrangement Γ_L^{cor} is a block diagonal matrix with entries consisting only of zeros and ones:

$$\Gamma_L^{\text{cor}} = \begin{bmatrix} 1_{N_1} & ... & 0 \\ ... & 1_{N_j} & ... \\ 0 & ... & 1_{N_J} \end{bmatrix}, \tag{4}$$

where 1_{N_j} is an N_j by N_j matrix consisting entirely of ones and corresponds to the N_j districts that have group j present in them.

Running a k-means type clustering algorithm on Γ_L^{cor} would be trivial, but only the finite sample version is available. Let $\hat{\Gamma}_L^{\text{cor}}$ be the correlation matrix associated with the finite sample covariance matrix $\hat{\Gamma}_L$. This $\hat{\Gamma}_L^{\text{cor}}$ will have off-diagonal entries that are neither zero nor one. However, given J any reasonable clustering algorithm should be able to recover which insurgent group is present in which district, provided enough data is available. Once districts have been clustered into groups, estimates for $\{\alpha_{ij}\}$ can be obtained. Appendix F provides further details for these steps.

We now consider how to produce an estimate $\hat{J}$ of the number of insurgent groups present. A major difficulty we face in determining the number of groups is that it is not obvious how to construct a null distribution for potential test statistics. For example, suppose that we wished to test for $J > 2$ versus the null hypothesis that $J = 2$. The distribution of plausible test statistics under the null would in general depend on whether one of these two groups is very large compared to the other, or whether they are of roughly equal size.

The only case where this difficulty is avoided is in the test of $J > 1$ versus the null hypothesis that $J = 1$, because under the null there are no nuisance parameters, as there is only one group present and thus every district must be assigned to that group. In this special case where the null hypothesis is $J = 1$, we will show that a permutation test can be constructed due to the simplicity of the group structure under the null. Our approach will thus be based on the repeated splitting of groups, as the only test we have available is one that asks whether a group should be split in two. The technique we use turns out to match that of Bruzzese and Vistocco [2015], except that in their case they assume that their data has a hierarchical form, whereas in our case we are dealing with a block diagonal correlation matrix that does not have any hierarchy.

We begin by looking at the set of all districts $\mathcal{N}$. Run a standard clustering procedure using distances based on Γ_L^{cor} to split these districts into two clusters, $\mathcal{N}_1$ and $\mathcal{N}_2$. Let $Q(\mathcal{N}_1, \mathcal{N}_2)$ be some test statistic that takes a high value when the division of $\mathcal{N}$ into $\mathcal{N}_1$ and $\mathcal{N}_2$ looks like a "good" division. To determine whether we actually want to split the $\mathcal{N}$ districts into the two groups $\mathcal{N}_1$ and $\mathcal{N}_2$, we will use a permutation test to calculate a cutoff value for Q.

If we do not split $\mathcal{N}$ into two groups, we are done and our estimate of the number of groups is $J = 1$. If we do split $\mathcal{N}$ into two groups, we apply our method recursively. That is, let our new set of districts be $\mathcal{D} = \mathcal{N}_1$, compute a clustering of these districts into clusters $\mathcal{D}_1$ and $\mathcal{D}_2$ using the relevant portion of Γ_L^{cor}, and then test whether we should actually split $\mathcal{D}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$. If we do split, we continue the recursion downwards. If not, we move to considering $\mathcal{D} = \mathcal{N}_2$. At the end of this procedure, we will have a partition of $\mathcal{N}$ into groups, where each of these groups should not be split further according to a permutation test.

The standard choice of permutation test would be to follow Bruzzese and Vistocco [2015], and consider permutations of attacks that generate different Γ^{cor} matrices. The test statistic Q would be based on how much of the covariance matrix can be explained by splitting the districts being considered into two groups, rather than leaving them as a single group. This standard approach runs into problems with the calculation of the null distribution of Q, because a reference distribution for $\hat{\Gamma}_L^{\text{cor}}$ needs to be calculated.[19] This calculation appears to be extremely complicated, because the answer depends on the finite sample behavior of $\hat{\Gamma}_L^{\text{cor}}$, which is not well understood. We avoid this problem by modifying the Bruzzese and Vistocco [2015] approach, and use a Q defined with respect to a set of auxiliary covariates Z, rather than Γ^{cor}.[20]

To see why this simplifies the problem, note that in the model the only source of correlation in insurgent attacks across districts is through ϵ. In particular, our model assumes that if the same insurgent group is present in both districts i and i', the correlation in attacks between districts will not depend on the relationship between any other covariates of i and i'. For example, it does not matter whether i is geographically adjacent to i', or geographically distant.

[19] Specifically, we would need to know how much adding a second group should improve model fit, if there is only actually one group in the data.

[20] The standard approach is presented in Appendix Table I.1. The results are the same as in the main text.

We will now add one additional assumption. Suppose that the districts where a given insurgent group is present are less dispersed in terms of these auxiliary covariates Z than a set of randomly chosen districts. For simplicity, we will focus specifically on geography, with Z_i being a vector indicating which other districts are geographically adjacent to i, but our approach is potentially more general.

Let $Z_{ii'} = 1$ if districts i and i' are geographically adjacent. Let $Q(\mathcal{D}_1, \mathcal{D}_2)$ describe the geographic dispersion of the insurgent group territories when the set $\mathcal{D}$ of districts are being split into two groups, according to the following formula:

$$Q(\mathcal{D}_1, \mathcal{D}_2) = \sum_{i \in \mathcal{D}_1} \sum_{i' \in \mathcal{D}_1} Z_{ii'} + \sum_{i \in \mathcal{D}_2} \sum_{i' \in \mathcal{D}_2} Z_{ii'} \tag{5}$$

That is, the test statistic Q is a simple calculation regarding whether $\mathcal{D}_1$ and $\mathcal{D}_2$ represent distinct geographic regions (in which case Q should be high, as there are a great many adjacencies), or whether the districts in $\mathcal{D}_1$ and $\mathcal{D}_2$ are randomly interspersed (in which case Q should be low). We do not worry about normalizing Q with respect to the total number of adjacencies, because the threshold value of Q will be computed using exactly the same set of districts $\mathcal{D}$.

The intuition for this test statistic is that, if all the districts in $\mathcal{D}$ are really part of the same group, then a split of these districts into two groups $\mathcal{D}_1$ and $\mathcal{D}_2$ will be based on finite sample noise, which is by assumption uncorrelated with geography. Thus, the additional groups should not be correlated with geography, and thus values of Q should be quite low, as the group labels are randomly assigned. A threshold value is easy to generate using Montecarlo permutations of the group structure: randomly permute the identities of all the districts, thereby forcing group membership to be unrelated to geographic location.

One does not have to exclusively rely on the Q statistic to estimate the number of groups J. In fact, this parameter is also recoverable using an entirely different approach, one based on the spectral properties of Γ_L. Notice that each Γ_L^j in (2) has rank 1, implying the rank of Γ_L is J.[21] The rank of Γ_L can be then consistently estimated by applying the intuition of Ahn and Horenstein [2013], using the eigenvalues of Γ_L.

Ahn and Horenstein's "eigenratio" approach proceeds as follows. Suppose that we were interested in estimating the rank of Γ_L. Let $\hat{\lambda}_k$ be the $k-$th largest eigenvalue of

[21] This is because the vectors $\alpha_{\cdot j}$ and $\alpha_{\cdot j'}$ describing insurgent group presence are orthogonal for $j \neq j'$.

$\hat{\Gamma}_L$. Asymptotically, the first J of these eigenvalues will be positive and bounded away from zero, while the remaining $N - J$ will go to zero. Ahn and Horenstein consider the "eigenratio"

$$\text{ER}_k = \hat{\lambda}_k / \hat{\lambda}_{k+1}. \tag{6}$$

Asymptotically, ER_k will converge to some positive value c_k for $k < J$. However, it will diverge to infinity for $k = J$, as the denominator becomes increasingly close to zero while the numerator remains bounded away from zero. A simple estimate for $\hat{J}$ can then be obtained by choosing the $\hat{J}$ that gives the highest value for $\text{ER}_{\hat{J}}$.[22]

The ER estimator has a finite sample tendency to estimate $\hat{J} = 1$, because the eigenvalues of random matrices are generally distributed so that the first few eigenvalues are spaced further apart than most of the remaining eigenvalues.[23] This effect has been noted previously by Ferson and Kim [2012], and Guo-Fitoussi and Darne [2014] perform an extensive simulation-based analysis.[24] The attack datasets that we consider in this paper, however, are noisier than the data generally used by researchers studying factor models in macro or finance. We are thus particularly interested in the finite sample properties of the estimator when the signal-to-noise ratio is very low. In Appendix G, we provide figures illustrating the behavior of the eigenratio estimator as the signal vanishes. The simulations that we perform are similar to those conducted in Guo-Fitoussi and Darne [2014], as well as the original Montecarlo exercises of Ahn and Horenstein [2013]. To the best of our knowledge, however, the figures that we produce have not previously appeared in the literature. This includes Appendix Figure G.3d, showing the distribution of the eigenratio estimator under the null hypothesis that there is no group structure.

2.4 Potentially Overlapping Insurgent Groups

We now relax the assumption that insurgent groups do not overlap. As in the approach discussed above, we will begin by assuming that J is known, and estimate

[22] Ahn and Horenstein [2013] require that there be some exogenous maximum number of possible factors, $J_{\max}$. We follow this, and use $J_{\max} = 80$ for this paper. Simulations are provided in Appendix G.

[23] Classic references here include the Wigner [1955] semi-circular distribution and the Marchenko-Pastur [1967] distribution.

[24] As Mirza and Storjohann [2014] point out, the effect is visible in the original Ahn and Horenstein [2013] simulations.

$\{\alpha_{ij}\}$. We then produce an estimate $\hat{J}$ based on a comparison of these estimates for different values of J. The estimates for $\{\alpha_{ij}\}$ will be based on non-negative matrix factorization, and the $\hat{J}$ estimate will be based on a modification of the ER_k discussed above.

We first construct an estimator for the $\{\alpha_{ij}\}$, given an assumed number of groups J. Consider choosing $\hat{\alpha}_{ij}$ for each district i and group j to satisfy the set of restrictions

$$\hat{\gamma}_{ii'} = \sum_j \hat{\alpha}_{ij}\hat{\alpha}_{i'j}.$$

where $\hat{\gamma}_{ii'}$ is the relevant entry in $\hat{\Gamma}_L$, estimated in (10) in Appendix C. If there are N districts, there are $N(N+1)/2$ restrictions: one for each off-diagonal element in one half of the symmetric covariance matrix, plus the diagonal elements. If there are J groups, there are $N \times J$ parameters to be estimated: one $\hat{\alpha}_{ij}$ for each district i and group j. A necessary condition for identification is thus that $(N+1)/2 \geq J$.

In the data the number of districts is large relative to plausible numbers of groups, and thus this inequality holds strictly and a penalty function is required. An obvious estimator for $\{\alpha_{ij}\}$ would then be the squared Frobenius norm

$$\underset{\hat{\alpha}_{ij}\geq 0}{\operatorname{argmin}} \sum_i \sum_{i'} \left(\hat{\gamma}_{ii'} - \sum_j \hat{\alpha}_{ij}\hat{\alpha}_{i'j} \right)^2 . \tag{7}$$

Unfortunately, solving this optimization problem directly by searching the space of $\{\alpha_{ij}\}$ is challenging because the problem as stated is non-convex in $\{\alpha_{ij}\}$. A variety of algorithms have been proposed for solving this problem. We will use the "Procrustes rotation" algorithm of Huang, Sidiropoulos, and Swami [2014]. This algorithm does not attempt to minimize (7), but instead solves a related optimization problem based on a spectral decomposition of $\hat{\Gamma}_L$. Huang and Sidiropoulos [2014] show that this algorithm is effective at solving (7), despite the fact that this objective is not used as part of the algorithm.[25]

We now consider how to produce an estimate $\hat{J}$. In Section 2.3, the rank of Γ_L was J, because the vectors $\alpha_{.j}$ and $\alpha_{.j'}$ describing insurgent group presence would be

[25] See Appendix H for more details. A previous version of this paper optimized (7) directly, using the algorithm of Birgin, Martinez, and Raydan [2000]. The (qualitatively identical) results from this direct approach are available upon request. Huang, Sidiropoulos, and Swami [2014] is orders of magnitude faster, converging in seconds or minutes rather than hours or days.

orthogonal for $j \neq j'$. This is no longer true if groups have the potential to overlap. Instead of the rank of Γ_L, we thus base our estimate $\hat{J}$ on the completely positive rank of Γ_L: that is, the rank of A, where $\Gamma_L = AA^{\mathrm{T}}$, and all entries of A are non-negative. Without further assumptions this decomposition is not identified. For example,

$$\Gamma_L = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{8}$$

could be decomposed either into $A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ or $\tilde{A} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Huang and Sidiropoulos [2014] summarize assumptions under which the non-negative factorization of Γ_L becomes unique for practical purposes:[26] each factor must have at least $J-1$ non-zero entries, and the non-zero entries of one factor must not be a subset of the non-zero entries of any other factor. In the example above, the second assumption is violated by matrix $\tilde{A}$. We will assume that the Huang and Sidiropoulos [2014] assumptions are satisfied. Thus, faced with the covariance matrix in (8), we would conclude that $\hat{J} = 1$ based on the factorization employing matrix A.

If Γ_L were known, the number of organized groups could thus be calculated immediately by producing a non-negative factorization of Γ_L. However, only the finite sample $\hat{\Gamma}_L$ is actually available, and in general this matrix will not have a non-negative factorization due to finite sample variation.

To address this problem, we will use a modification of the eigenratio approach. The intuition behind the Ahn and Horenstein [2013] approach appears very general. Consider a rank k approximation to an $N \times N$ matrix. The first k eigenvectors can be used to create such an approximation. How much better is a rank $k+1$ approximation? If the $k+1$th eigenvalue is very small relative to the kth eigenvalue, then considering a rank $k+1$ matrix instead of a rank k matrix does not improve the approximation very much, and ER_k will thus be very high. We can apply this intuition to the case of the group structure of $\hat{\Gamma}_L$. Let A_k be an approximate non-negative factorization of $\hat{\Gamma}_L$ with k factors. How much better would A_{k+1} be as an approximation to $\hat{\Gamma}_L$? Asymptotically, if Γ_L was produced by k groups, the improvement will be zero.

[26] Conditions theoretically guaranteeing the uniqueness of the factorization are more complicated: see the references in Huang and Sidiropoulos [2014].

A ratio equivalent to Ahn and Horenstein's "eigenratio" can then be expressed as

$$\text{NNR}_k = \frac{||\hat{\Gamma}_L - A_k A_k^{\mathrm{T}}||_F^2 - ||\hat{\Gamma}_L - A_{k-1} A_{k-1}^{\mathrm{T}}||_F^2}{||\hat{\Gamma}_L - A_{k+1} A_{k+1}^{\mathrm{T}}||_F^2 - ||\hat{\Gamma}_L - A_k A_k^{\mathrm{T}}||_F^2} \tag{9}$$

where $||.||_F$ is the Frobenius norm. The intuition for NNR_k is exactly that of the ER_k: if Γ_L has a completely positive rank of k, then the $k+1$th factor should not help explain Γ_L, and thus NNR_k should diverge to infinity. In contrast, values of NNR_k for $k < J$ will converge to finite values.

The finite sample behavior of the eigenratio estimator is shared by estimators using NNR. It will thus be important to check whether the values of NNR obtained might have arisen by random chance from data with no actual group structure. Consider the value of $\max_{k<J_{\max}} \text{NNR}_k$. We wish to compare this test statistic to its distribution under the assumption that there is no actual group structure, obtaining appropriate p-values.[27]

To do so, we consider a "reference distribution" where there are no organized groups. We randomly generate attack data based on this distribution, calculate an equivalent to $\hat{\Gamma}_L$ based on this randomly generated data, calculate a value for NNR_k based on this matrix, and then repeat this process 100 times. We consider three different reference distributions: details are provided in Appendix I.

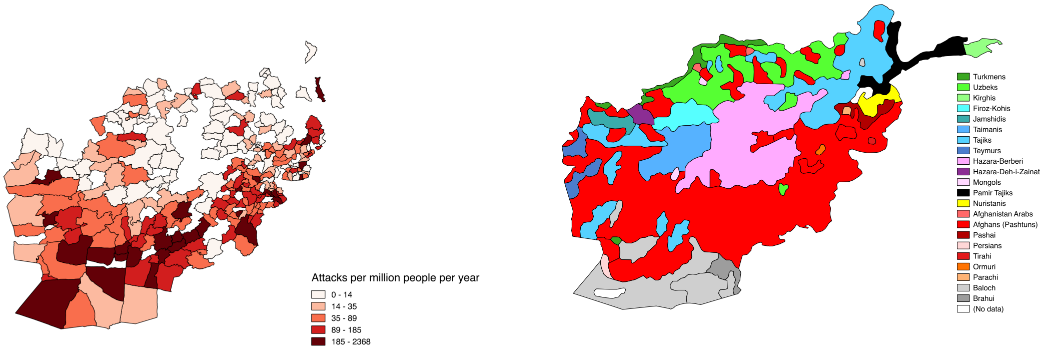
2.5 Robustness: Potentially changing district environments

Both the non-overlapping and overlapping approaches just described assume that the covariance in attacks by group members across districts remains the same even across long periods of time. In the observed data, however, it may be the case that in earlier years certain districts are the focus of many attacks, while in later years activity shifts to other districts. These sorts of long-term changes can be accounted for by considering only the covariance in attacks across districts within shorter time windows.

Let Γ_m be calculated the same as Γ from Equation (1), but using only daily attack data from month m. As the number of days of data used to calculate estimates of Γ_m does not increase asymptotically for any given month m, estimation based on a

[27] Other hypothesis tests are difficult to perform: the distribution of eigenvalues resulting from random variation in finite samples is not obvious. We thus do not report confidence intervals for $\hat{J}$. For similar reasons, we also do not report confidence intervals for $\{\hat{\alpha}_{ij}\}$ below.

Figure 1: Afghanistan data



(a) Total attacks per capita

(b) Ethnic Groups

single Γ_m would be inconsistent. Aggregating across months, however, results in a consistent estimator that is robust to changes in attack probabilities between districts at monthly frequency.

Specifically, assume that the probability of an attack in district i in month m, either from unorganized militants or an organized group, now changes with a parameter ζ_{im}. That is, the probability of an attack from a unorganized militant is now $\zeta_{im}\eta$, and the probability of an attack from member of organized group j is now $\zeta_{im}\epsilon_{jt}$. Let $D(\cdot)$ again indicate a diagonal matrix with the given entries on the diagonal. If ζ were known, the standardized matrix $\tilde{\Gamma}_m = D(\frac{1}{\zeta_m})\Gamma_m D(\frac{1}{\zeta_m})$ could be summed to create $\tilde{\Gamma} = D(\sum_m \zeta_m)\tilde{\Gamma}_m D(\sum_m \zeta_m)$. $\tilde{\Gamma}$ could then be used to estimate $\{\alpha_{ij}\}$. In reality, ζ is unobserved; however, dividing by the observed number of attacks creates a feasible estimator, with $\{\alpha_{ij}\}$ identified up to scale.

This approach can be employed with both estimation based on clustering and that based on non-negative matrix factorization. Appendix J provides further details.

3 Data

Incident-level information is the main input to our empirical analysis. Both Afghanistan and Pakistan were covered by the Worldwide Incidents Tracking System, a discontinued U.S. government database [Wigle 2010].[28] Data is available for the location, date,

[28] The data remains accessible online courtesy of the Empirical Studies Of Conflict (ESOC) project at Princeton University.

and type of violent incidents from January 2004 to September 2009.[29]

In some cases the identity of the group responsible for the attack is also reported. While we do not use this identity information as an input to our analysis, we will be interested in comparing the identity information with the output of our analysis in order to see what value our methods can add.

The violent incidents catalogued in the WITS data are episodes of violence initiated by insurgents, or acts of random violence. The data does not include violence directly connected to military counterinsurgency operations, such as for instance a U.S. military attack on a Taliban safe house or the bombing of a fortified compound. Appendix K provides details.

The location reported for an attack in WITS is given as latitude and longitude coordinates. This would seem to suggest that attacks could be analyzed as some sort of spatial point process. Closer inspection, however, reveals that the latitude and longitude coordinates reported are not those of the actual location of the attack, but rather the coordinates of a prominent nearby geographic feature. Sometimes this is a city or village, but for the vast majority of incidents the location given is that of the centroid of the district in which the incident occurred.

In Afghanistan, the district is the lowest-level political unit, and we will use it as our geographic unit of analysis. A few districts have been split in recent years: this paper uses 2005 administrative boundaries, which specify 398 districts. The WITS data effectively provides panel data at the district-day level, with $N = 398$ and $T = 2082$. District-level geographic locations are also used for the Pakistan WITS data.

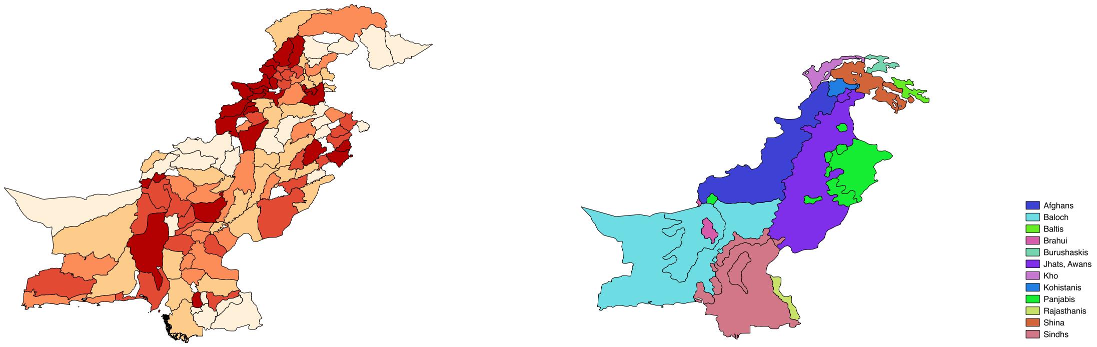
According to the data, there are some days where as many as 64 different districts in Afghanistan are affected by simultaneous insurgent attacks. However, there are also 123 districts with no reported incidents over the entire 2004-2009 time period. The identity of the insurgent group launching the attack is provided 55% of the time. Overwhelmingly (98% of the time) these attacks are reported as being launched by

[29] The following two examples illustrate the typical form of incident descriptions:

"*On 27 March 2005, in Laghman, Afghanistan, assailants fired rockets at the Governor House, killing four Afghan soldiers and causing minor damage. The Taliban claimed responsibility for the attack.*"

"*On 19 February 2006, in Nangarhar, Afghanistan, a suicide bomber detonated an improvised explosive device (IED) prematurely near a road used by government and military personnel, causing no injuries or damage. No group claimed responsibility.*"

Figure 2: Pakistan data



(a) Total attacks per capita

(b) Ethnic Groups

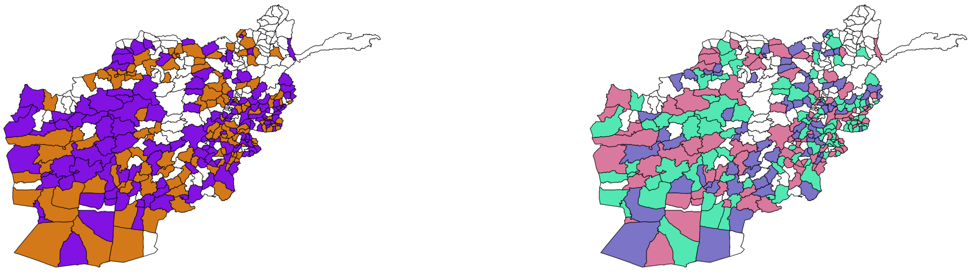
"Taliban".[30] We do not use this group identity data as part of our analysis. Instead, in Section 4 we conclude based solely on the pattern of simultaneous attacks that there is only a single group active in Afghanistan during this period, and we thus concur with the coding given in WITS: in Section 5 we consider later data and show that we disagree with a similar coding there.

For Pakistan, the BFRS dataset [Mesquita et al. 2015] is also available. This is similar to WITS, in that it provides daily data on violent incidents, including geographic information. BFRS data is available until 2011, and over the WITS time frame of 2004-2009, BFRS contains approximately twice as many incidents as WITS. Because of the greater number of attacks recorded, we prefer the BFRS data to the WITS data.

We only make use of BFRS data from mid-2008 until the end of the sample in 2011, because qualitative evidence suggests that the structure of insurgent groups during this 3.5 year period was relatively stable. In early 2008 national elections took place in Pakistan, producing a new executive after the resignation of General Pervez Musharraf: we use this important political break as the starting point for our analysis. The Taliban also strengthened considerably around 2008 [Iqbal and De Silva 2013], and there is qualitative evidence of an active organization in Sindh (the Sindhudesh Liberation Army) and in Balochistan (the Baloch Republican Army).

[30] The WITS attributes 54% of attacks to the Taliban, 0.6% to Hizb-i Islami, 0.3% to al Qa'ida, and 45% to unknown, and the GTD attributes 59% of attacks to the Taliban, 0.5% to Hizb-i Islami, 0.1% to al Qa'ida, 0.8% to the Haqqani network, and 39% to unknown (the rest is split among another seven groups).

Figure 3: Afghanistan groups via spherical k-means



(a) 2 Groups

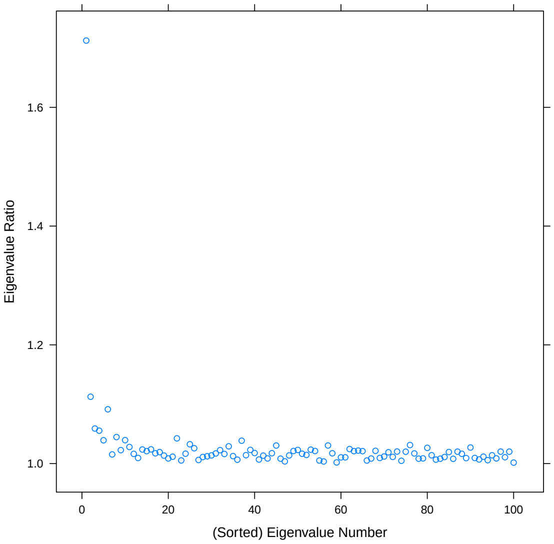
(b) 3 Groups

For geographic data on ethnicities, we use the Soviet *Atlas Narodov Mira* data.[31] In Figure 1 we show the pattern of attacks by district in Afghanistan and the distribution of ethnicities. The concentration of attacks in the ethnic Pashtun areas is evident. In Figure 2 we report the same information for Pakistan. This data forms the basis for most of the analysis that will be performed immediately below in Section 4.

In Section 5 we use more recent data to show how our approach provides information on group structure beyond what is already available in input datasets. For this exercise we use the Global Terrorism Database, which has an ongoing data collection effort. As of the writing of this paper, data is available through to the end of 2016. The GTD provides data on the location and date of attacks at the same resolution as WITS and the BFRS. Unfortunately, the GTD includes only a fraction as many attacks: for Afghanistan, WITS reports 7846 attacks whereas the GTD only reports 1692 attacks over the same period. A major advantage of the GTD data, however, is that it provides a coding of "related" attacks based on expert opinion, which effectively identifies exactly the simultaneous attacks that we are interested in analyzing with our method. Thus, although the GTD dataset we use in Section 5 is smaller, it includes much less noise: we can focus on only simultaneous attacks that were judged by analysts to truly be organized related attacks, thereby omitting most of the cases where individual attacks simply happened to occur on the same day through random chance.

[31] The version used is the "Geo-referencing of ethnic groups" data set of Weidmann et al. [2010].

Figure 4: Eigenratios, Afghanistan



4 Results

We first analyze attack data from Afghanistan, and then consider the case of Pakistan. In both cases, we begin with the spherical k-means clustering and splitting approach outlined in Section 2.3, and then proceed to the non-negative matrix factorization approach of Section 2.4. For all these analyses, we will use attack covariance matrices calculated using only within-month variation, as described in Section 2.5, unless noted otherwise. This is because it is important to avoid contamination by long-term trends, as well as seasonal variation in conflict.

4.1 Afghanistan

To illustrate attack data from Afghanistan, Figure 3 shows clustering based on spherical k-means, as outlined in Section 2.3. Qualitatively, the clusters shown in the figure appear indistinguishable from random noise.[32] We now test this hypothesis formally using the statistic proposed in (5).

Column I of Table 1 shows the results of this analysis. We begin by considering the null hypothesis that all districts are associated with the same insurgent group,

[32] These figures are calculated based on a "within month" covariance matrix, as described in Section 2.5. Results do not change with other approaches.

Table 1: Estimation of $\hat{J}$ based on hierarchical splits

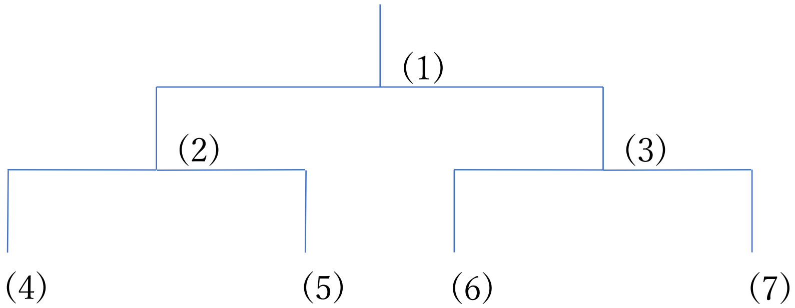
		Afghanistan I	Pakistan II
Split at (1)?	Randomly shuffled data (mean)	284.99	137.67
	Std. dev.	11.09	8.52
	Actual data	289.00	159.00
	p-value	0.38	0.01
Split at (2)?	Randomly shuffled data (mean)		44.42
	Std. dev.		4.30
	Actual data		64.00
	p-value		0.00
Split at (3)?	Randomly shuffled data (mean)		34.30
	Std. dev.		4.61
	Actual data		47.00
	p-value		0.01
Split at (4)?	Randomly shuffled data (mean)		18.22
	Std. dev.		3.01
	Actual data		16.00
	p-value		0.71
Split at (5)?	Randomly shuffled data (mean)		14.01
	Std. dev.		2.90
	Actual data		19.00
	p-value		0.08
Split at (6)?	Randomly shuffled data (mean)		12.63
	Std. dev.		2.32
	Actual data		15.00
	p-value		0.23
Split at (7)?	Randomly shuffled data (mean)		9.21
	Std. dev.		1.87
	Actual data		10.00
	p-value		0.44

Each column computes a test statistic Q as described in Section 2.3, based on a within-month covariance matrix as described in Section 2.5. Figure 5 shows the order of the potential splits. Columns differ in the underlying attack data used:

Column I uses the full Afghanistan WITS dataset.

Column II uses the Pakistan BFRS dataset for May 2008 - October 2011.

Figure 5: Hierarchical Splits

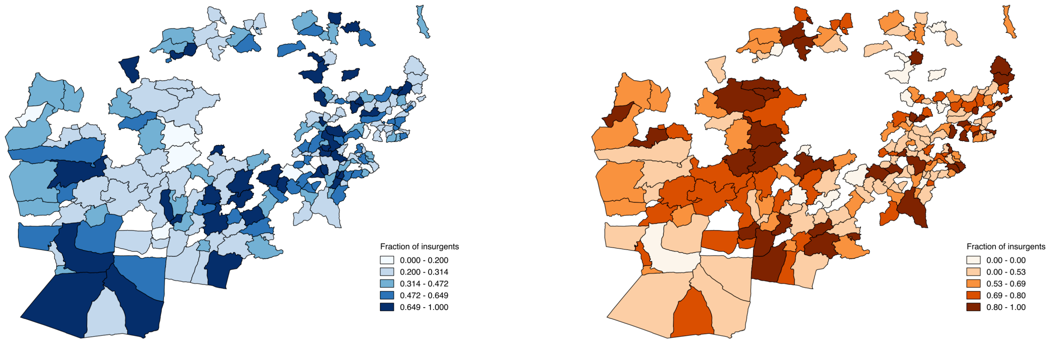


and ask whether we should instead split the districts into two groups. In Figure 5, this first potential split is indicated by (1). We calculate how geographically distinct these split groups would be, and also calculate how geographically distinct we would expect them to be under the null hypothesis that there is actually only one group. These calculations are shown in Table 1 on the rows following "Split at (1)?" The results displayed in column I of Table 1 show that for Afghanistan the actual data leads to groups that are no more geographically distinct than would be expected by random chance. Thus, for the Afghanistan data we stop at one group. The eigenratio approach of Ahn and Horenstein [2013] produces an identical result, as illustrated in Figure 4, where one large eigenratio at 1 is strikingly evident. Thus, both approaches suggest that the Taliban do not appear divided into multiple organized groups.

The group membership shown in Figure 3 involves a discrete partition of districts into insurgent groups. Some districts, however, might have more than one active insurgent group. The model presented in Section 2.4 allows for this possibility. An additional advantage of this model is that it provides a test against the null hypothesis that $J = 0$, and all attacks are the result of disorganized local actors. In contrast, the model used in Section 2.3 assumes that there is exactly one organized group present in each district, and thus this model cannot be used to test the hypothesis that there are actually no groups.

The non-negative matrix factorization approach of Section 2.4 gives very similar results to those just discussed. In contrast to the previous approach, multiple insurgent groups may now be active in any single district, and thus it is no longer easy to display the estimated insurgent group structure on a single map. Instead, we produce one map for each group. Figure 6 provides a visualization of the factorization in the case where $J = 2$. Again, there is no discernible pattern to the estimated insurgent groups.

Figure 6: Afghanistan, 2 groups via NNMF



(a) Group 1

(b) Group 2

Table 2: Estimated number of groups via NNR, Afghanistan

	Not by Month		By Month		
	I	II	III	IV	V
Afghanistan (WITS, Jan 2004 - Sept 2009, weighted)	4	4	1	1	1
(p value, vs. no group structure)	0.57	0.41	0.01	0.01	0.03
Afghanistan (WITS, Jan 2004 - Sept 2009, unweighted)	1	1	1	1	1
	0.02	0.02	0.02	0.03	0.06

Each row presents two estimates of $\hat{J}$, the number of groups present. Columns I and II show the first estimate, described in Section 2.4.

Columns III through V show the second estimate, based on the within-month covariance matrix as described in Section 2.5.

In each column, the p values presented are a test of the null hypothesis that there is no group structure. Other tests (e.g. $J = 1$ vs. $J = 2$) appear difficult to construct.

Columns I and III compute p values by comparing to a reference distribution where the time of the attacks within each district has been permuted. See Appendix I for a description of this and other reference distributions.

Column IV is the same as Column III, but the time of attacks is permuted only within each month.

Columns II and V consider only permutations that keep constant the total number of attacks in each district and on each day.

We are particularly interested in whether we can reject the null hypothesis that $J = 0$, i.e. that there are no organized insurgent groups present at all. We begin by calculating the ratio described in Equation (9), and choose $\hat{J}$ so as to maximize this ratio. We then consider the distribution that this ratio would have if there were actually no organized groups. To do this, we use the permutation approach described in Section 2.4 and Appendix I.

Table 2 shows the results of this analysis for the Afghanistan data. There are four estimates of J provided. Beginning with the first two columns of the first row, $\hat{J} = 4$ in the case where districts are weighted proportionally to the number of attacks in the district. Continuing to the next three columns of the first row, $\hat{J} = 1$ if, in addition to this weighting, the covariance matrix is calculated considering only within-month variation in attacks using the approach described in Section 2.5. The second row provides estimates without weighting districts, and results in $\hat{J} = 1$ regardless of whether the approach in Section 2.5 is used or not.

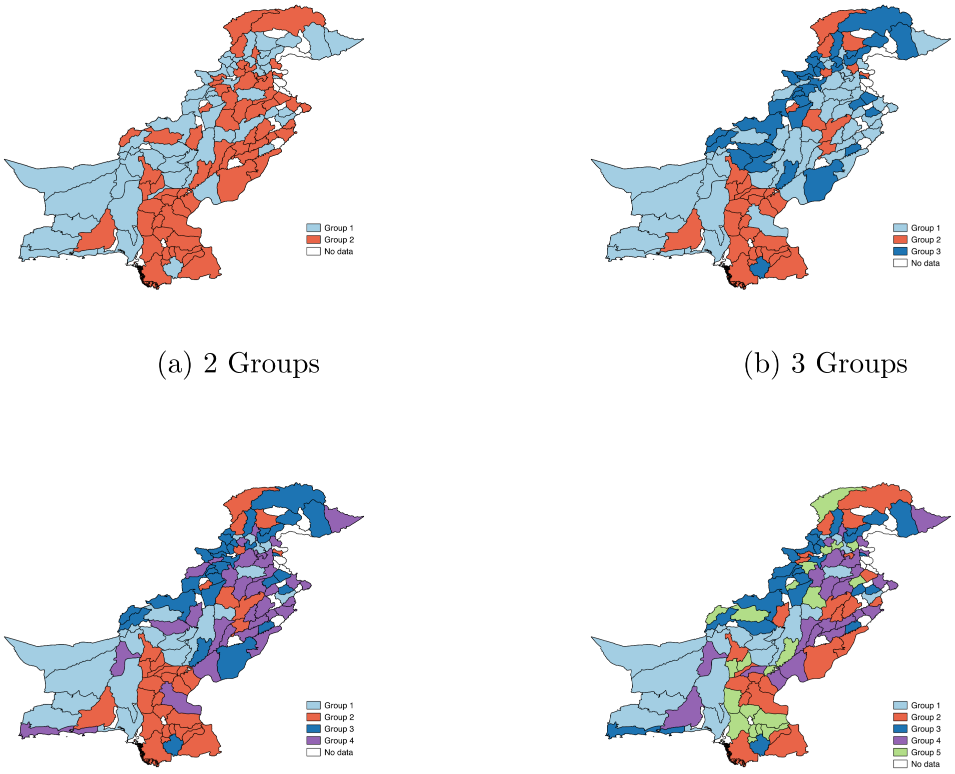
Below each $\hat{J}$ estimate a p value is shown, corresponding to a test of the null hypothesis that in fact there is no group structure, with $J = 0$. We see that in general the null is rejected at the 95% level. The exception is the case where our estimate was $\hat{J} = 4$. With this specification, the model appears to have low power. This analysis supports the results obtained in Table 1, in that there appears to be one organized group of insurgents, rather than more than one. Furthermore, the observed NNR_1 values, calculated according to Equation 9, appear to be more extreme than would be the case if there were no organized groups at all. Table 2 shows that the observed data appears to be inconsistent with $J = 0$, a conclusion that we were not able to draw from the results shown in Table 1. Overall, in the Afghan case all our methodologies point to a single, organized Taliban insurgent group during the 2004-2009 period.

4.2 Pakistan

Clusterings based on the Pakistan attack data are shown in Figure 7. Unlike the results for Afghanistan shown in Figure 3, our clusterings for Pakistan, computed on the basis of the attack covariance matrix, result in groups that appear to be clustered geographically. For a more formal analysis, we consider the Q statistic results in column II of Table 1.

Unlike the case with Afghanistan in Column I, we do not stop immediately with an estimate of $\hat{J} = 1$. Instead, for Pakistan, the first set of rows in Table 1 shows

Figure 7: Pakistan groups via spherical k-means



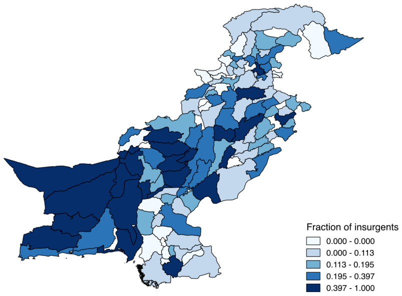
(a) 2 Groups

(b) 3 Groups

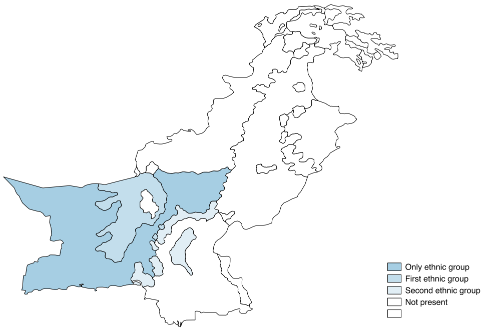
(c) 4 Groups

(d) 5 Groups

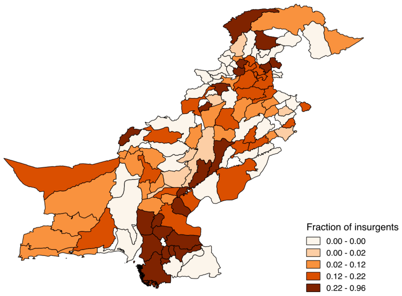
that if the set of all districts is split into two groups, these groups are substantially more geographically distinct than would be expected if there were no actual group structure. We thus split the set of districts into these two groups, and continue recursively, asking for each of these two groups whether the group should be further divided. These questions are indicated by (2) and (3) in Figure 5, and the next two sets of rows in Table 1. In each of these cases the potential splits appear to be more geographically distinct than would be expected by random chance, and so in each case the group is split, leading to a total of four groups. Continuing recursively, we consider whether any of the four groups we now have should be further split. These questions correspond to the final four sets of rows in Table 1. We see that none of these splits generate groups that are more geographically distinct than would be expected by random chance, and thus we do not split any of these groups. At this



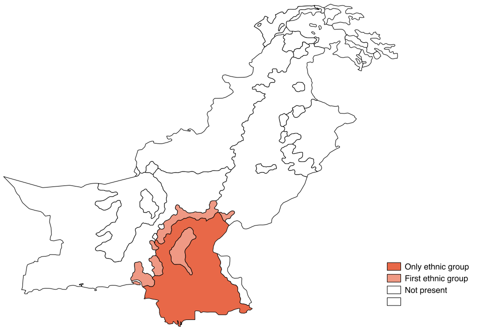
(a) Group 1



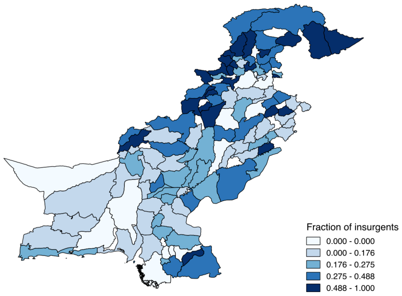
(b) Balochis



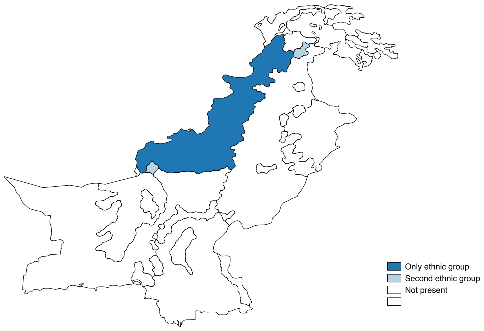
(c) Group 2



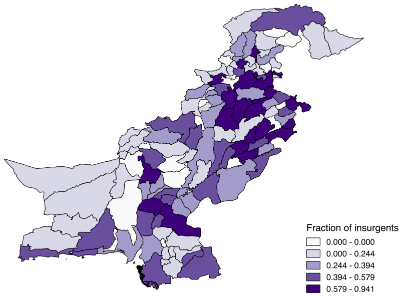
(d) Sindhis



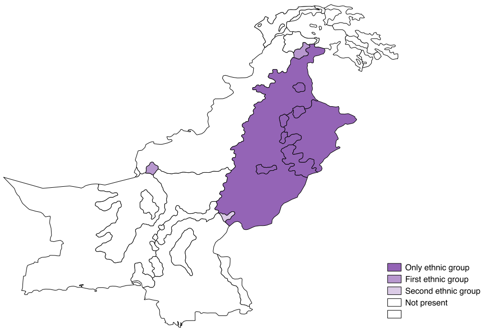
(e) Group 3



(f) Afghans



(g) Group 4



(h) Panjabis, Jhats, Awans

Table 3: Ethnic composition of groups shown in Figure 7c

	Group 1	Group 2	Group 3	Group 4
Baloch	0.62	0.25	0.00	0.12
	(0.09)	(0.09)	(0.10)	(0.10)
Sindhs	0.04	0.87	0.04	0.04
	(0.07)	(0.07)	(0.08)	(0.09)
Afghans	0.12	0.08	0.58	0.23
	(0.07)	(0.07)	(0.08)	(0.08)
Panjabis, Jhats, Awans	0.16	0.12	0.23	0.49
	(0.05)	(0.05)	(0.06)	(0.06)
Other	0.00	0.29	0.57	0.14
	(0.13)	(0.13)	(0.15)	(0.16)
N	115	115	115	115

Each column corresponds to a single regression without intercept.

The dependent variable is a dummy variable indicating whether a given district was clustered into the specified group number in the clustering shown in Figure 7c. Districts shown as white in the figure ("no data") are dropped: the remaining 115 districts are used in the regression.

The independent variables are a set of dummy variables, indicating whether the specified ethnicity was listed as the first ethnicity at the centroid of a given district.

Each row should sum to 1 because each coefficient in the table is a conditional mean giving the fraction of districts of the specified ethnicity that were clustered into the specified group, and the clustering in Figure 7c assigns each district to one group. Rows may not sum exactly to 1 because of rounding.

Standard errors in parentheses.

point there is no further recursion, with an estimate of $\hat{J} = 4$. The unified insurgent structure ($\hat{J} = 1$) that we recover for the Afghan case thus appears not to be present in Pakistan. This accords with qualitative analysis such as Dorronsoro [2009].

For completeness we now analyze the Pakistan data using the non-negative matrix factorization approach of Section 2.4. As with the Afghan data, results are generally in line with that obtained using the Q statistic approach. The left-hand column of Figure 8 shows a non-negative matrix factorization of the attack covariance matrix for Pakistan, using four factors. The result here is very close to that shown in Figure 7c. Furthermore, both of these figures show what appears to be a close relationship between the estimated group structure and the arrangement of ethnic groups in Pakistan. The relevant breakdown of these ethnic groups is shown in the right-hand column of Figure 8.

A qualitative comparison of the left and right columns of Figure 8 shows that there is one insurgent group present in Balochistan, another in the area populated by Sindhs, a third in the area populated by "Afghans" (i.e. Pashtuns), and a fourth in the Punjabi areas of Pakistan. The northernmost areas of Pakistan, with numerous smaller ethnicities, appear to be associated most closely with the "Afghans".[33] The major ethnic divisions of Pakistan can thus be successfully reproduced employing only the covariance matrix of insurgent attacks.

Tables 3 and 4 show the relationship between estimated insurgent groups and ethnic groups in a quantitative fashion. These tables are constructed to describe the distribution of ethnicities across the estimated insurgent groups. Each row corresponds to an ethnicity and sums to 100% (up to rounding error). The rows have been ordered so that the diagonal entries correspond to the qualitative relationship between insurgent groups and ethnicities just discussed. This is the same ordering of rows as is used in Figure 8.

We now consider an eigenratio type analysis of the Pakistan attack data. In the case of Afghanistan, analysis based on the Q statistic approach in Table 1 resulted in an estimate of $\hat{J} = 1$, but it was then necessary to use the results shown in Table 2 to show that the null hypothesis of $J = 0$ could be rejected. In contrast, with the Pakistan data, Table 1 gives $\hat{J} = 4$. This result would be extreme under a null hypothesis of $J = 0$, and thus it is not as important to seek alternate confirmation that there is indeed a group structure in the data.[34] This turns out to be fortunate, as the eigenratio type analysis shown in Table 5 is inconclusive in the case of the Pakistan data.[35]

[33] Adding a fifth group does not result in these "other" ethnicities being clustered into their own separate group: see Figure 7d. This may be because these areas consist of many small ethnic groups, and there is not a sufficient number of attacks for these smaller groups to be estimated correctly.

[34] In order for Table 1 to lead to an estimate of $\hat{J} = 4$ it must be that for each of two groups, three groups, and four groups, the improvement in geographic clustering is at least one standard deviation better than would be expected if there were no group structure. A result of $\hat{J} = 4$ is thus already very extreme under the null that $J = 0$.

[35] Very few entries are statistically significant at the 95% level, and those that are appear to be computational artifacts of some sort, giving very high estimates for $\hat{J}$. In Appendix G.1 we compare an eigenratio analysis to an analysis based on our Q statistic and show that the eigenratio approach is less likely to return the correct number of groups and sometimes returns very high numbers of groups. We join a number of other researchers in discarding a $\hat{J}$ estimate based on eigenratios in favour of other evidence: both Henzel and Rengel [2014] and Alquist and Coibion [2014] discard $\hat{J} = 1$ in favour of two factors, and Bleaney et al. [2012] discards $\hat{J} = 1$ or $\hat{J} = 3$ in favour of four or more factors, while Rezitis [2015] discards $\hat{J} = 2$ in favour of five factors.

Table 4: Ethnic composition of groups shown in Figure 8

	Group 1	Group 2	Group 3	Group 4
Baloch	0.58	0.07	0.12	0.23
	(0.05)	(0.04)	(0.06)	(0.06)
Sindhs	0.14	0.35	0.19	0.32
	(0.04)	(0.03)	(0.05)	(0.05)
Afghans	0.15	0.07	0.48	0.30
	(0.04)	(0.03)	(0.04)	(0.05)
Panjabis, Jhats, Awans	0.22	0.11	0.23	0.44
	(0.03)	(0.03)	(0.03)	(0.04)
Other	0.05	0.11	0.61	0.24
	(0.07)	(0.06)	(0.08)	(0.09)
N	115	115	115	115

Each column corresponds to a single regression without intercept, concerning group $j \in \{1, 2, 3, 4\}$.

The dependent variable is $\hat{\alpha}_{ij} / \sum_{j' \in \{1,2,3,4\}} \hat{\alpha}_{ij'}$. This is the fraction of organized insurgents present in a district that are from group j. This data is displayed in the left column of Figure 8, and it is available for the same 115 districts that were analyzed in Table 3.

The independent variables are a set of dummy variables, indicating whether the specified ethnicity was listed as the first ethnicity at the centroid of a given district.

Each row should sum to 1 (up to rounding) by the same argument as in Table 3: each coefficient in the table is a conditional mean for the ethnicity in question, every district is coded as one ethnicity, and the group shares must sum to one.

Standard errors in parentheses.

Table 5: Estimated number of groups via NNR, Pakistan

	Not by Month		By Month		
	I	II	III	IV	V
Pakistan (BFRS, May 2008 - Oct 2011, weighted)	1	1	1	1	1
	0.63	0.63	0.16	0.28	0.55
Pakistan (BFRS, May 2008 - Oct 2011, unweighted)	2	2	16	16	16
	0.73	0.68	0.00	0.01	0.03

Notes: same as Table 2, except with Pakistan data.

5 Discussion and Applications to GTD Data

This section provides a set of applications designed to display some of the additional capabilities of our methodology. We continue to focus on the insurgencies in Afghanistan and Pakistan; however, the applications we present translate to other contexts as well, both in the conflict literature and in some cases more broadly in political economy.

5.1 Tracing Insurgent Coalition Structures Over Time

Often qualitative analysis of conflict traces a static or slow-moving picture of alliances and coordination among different factions or insurgent groups. For example, in the Introduction we discussed how the "perpetrator" coding in the GTD uses a generic "Taliban" attribution for the majority of attacks over the 2004-2016 period. Shifts in internal organization and alliances, however, are common and can be frequent [Christia, 2012]. Our methodology displays promise in tracing such structures over time.

Consider the Afghan case, where observers have highlighted shifts in the Taliban organization over time after 2009. For instance, in late 2015 the Washington Post writes that:

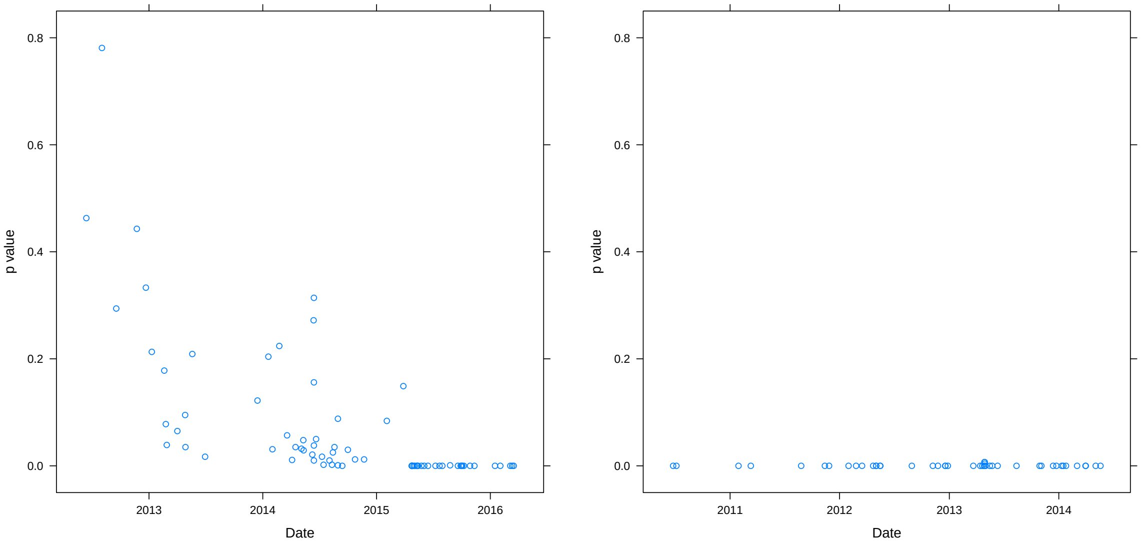
> "*the Taliban are no longer viewed as a monolithic entity capable of uniting Afghans under a religious identity. The discovery of Mullah Omar's death in 2013, the mythologized Robin Hood of Afghanistan, has ended the Taliban's identity as Afghanistan's unifying entity. Fractured movements tend to prolong conflicts. This is evident with contemporary conflicts around the globe from Yemen, Syria, and Libya.*"[36]

Several other articles maintain the first assertion and the final two sentences justify our interest in the question.[37]

[36] Shawn Snow 12/21/2015 Washington Post *"Why a fractured Taliban is endangering the U.S. mission in Afghanistan".*

[37] Sudarsan Raghavan 2/15/2015 Washington Post *"As the U.S. mission winds down, Afghan insurgency grows more complex".* The article reports that *"the Taliban is transforming into a patchwork of forces with often conflicting ideals and motivations, looking less like the ultra-religious movement it started out as in the mid-1990s. The fragmentation may suggest the movement is weakening, but it is forcing Afghanistan's government to confront an insurgency that is becoming increasingly diverse, scattered — and more lethal."*

Figure 9: Test for splitting into multiple groups



(a) Afghanistan

(b) Pakistan

Although the methodology presented in Section 2 assumes a constant group structure, we can apply our method to data with a changing group structure by repeatedly running an analysis using a moving window of data. Specifically, each dot in Figure 9a shows the p value corresponding to a moving window of 51 simultaneous attacks from the GTD attack data for Afghanistan, computed the same way as in Column I of Table 1, with the dots located on the horizontal axis at the date of the 26th attack.[38] Low p values indicate rejection of the null hypothesis of the presence of a single, unified group for our hierarchical split test. A single unitary actor cannot be rejected in the period before 2012 (consistently with our results using WITS for 2004-2009 in Section 4), but the situation progressively evolves towards fragmentation after 2012. The Taliban appear composed of separate factions after 2013-2014, information not available in the GTD group identity coding. The test traces a continuous progression, not a drastic structural breakpoint, starting around 2013. By early 2015 we can systematically reject at the 5% statistical confidence level the hypothesis that the Taliban is a unitary organization - the Taliban still coordinate but they do so in different

[38] The only GTD components of the attack data used are geolocation and exact date of the event. The GTD is valuable here because of the longer time coverage, which spans the full 2004-2016 period (at a loss of higher sparsity of incidents recorded). We do not employ in any part of the analysis here information on group identity in the "perpetrator group" coding of the GTD.

clusters. As further supporting evidence, one obtains qualitatively similar evidence of increased fragmentation of the Taliban when rerunning the exercise of Figure 9a employing only incidents explicitly labeled as "Taliban" in the GTD. That is, the pattern of organizational change appears internal, as it is clear even when performing the analysis within Taliban incidents only, according to the GTD attribution.[39]

As final validation of this method, Figure 9b shows that an equivalent analysis using GTD data for Pakistan always yields low p values over time. The multiple groups discussed in Section 4 using BFRS data are a constant feature of insurgency in this country and are confirmed in this application as well.[40]

5.2 Detection of New Insurgent Groups

The methods presented in this paper also allow for the early detection of emerging insurgent groups based on attack data alone. As proof of concept, consider here the case of the Sindhudesh Liberation Army (SDLA), a violent independence movement in the Pakistan province of Sindh.

In Appendix L we document how the GTD reports coordinated, large-scale activity of the SDLA in the region on February 25th, 2012. This identification was made only in 2012 notwithstanding explicit claims of responsibility by the SDLA itself in previous events in 2010 and 2011 (the GTD reports only 6 of them, while the BFRS shows hundreds of unattributed but coordinated incidents).[41] The GTD still attributes attacks by the SDLA to an *"unknown group"* in 2011. None of the GTD attacks before February 2012 are listed as part of multiple incidents.

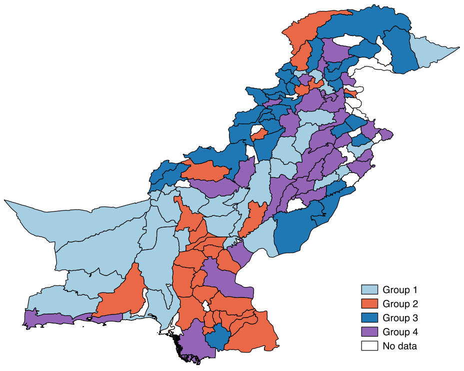
Estimates based on our methodology show an insurgent group using coordinated attacks operating in Sindh in April 2011. Figure 10 shows a clustering based on

[39] The figure appears virtually identical and it is available upon request. This also clarifies that none of these findings are due to the entry of ISIS in the Afghan arena. There are two additional arguments, besides the within-Taliban analysis, to doubt that ISIS may be related to these findings. The first argument is timing. The first recorded ISIS attack in the GTD is in 2015, while the p values start oscillating well before. The other argument is related to sheer magnitudes. Against a total of $5,420$ generic "Taliban" incidents recorded in GTD over the same period, ISIS is attributed only 132 incidents in total for Afghanistan between January 1st, 2010 and December 31st, 2016.

[40] The GTD contains relatively few simultaneous attacks for Pakistan, and thus we use a window of 26 attacks rather than the 51 used for Figure 9a. A narrower window will in general *increase* calculated p values, and we have verified that if a 51 attack window were used all p values for Pakistan would be less than 0.01.

[41] Claims of responsibility were documented in the public sources used by GTD, but presumably the GTD coders could not be certain of the veracity of these claims, which were in pamphlets left at the scene of the bombings.

Figure 10: Pakistan groups, 10 months before GTD identifies SDLA



BFRS data truncated to end in April 2011, and Appendix Table M.2 shows that our Q statistic method still indicates the presence of 4 groups. Figure 10 is almost identical to Figure 7c, which used the full sample ending in October 2011 (even this later date is still a full quarter before the SDLA is identified in the GTD). A researcher with access to microlevel data such as the BFRS and employing our methodology in real time could have been able, almost a year in advance, to detect a significant cluster of activity in Sindh. Employing only the BFRS incident description section would not have been a valid substitute for our method: only 9 incidents are coded to the SDLA in BFRS event descriptions across all the 2008-2011 sample: this, out of almost $1,300$ attacks listed in Sindh between 2010 and the first quarter of 2011 alone. We thus see that our methodology can detect changes in insurgency structures beyond what is reported in the best publicly available datasets.

5.3 Validation of Sources: The Case of the Haqqani Network

The WITS dataset provides a "group" coding for about half of all attacks: in almost all of these, the perpetrator is listed as undifferentiated "Taliban". Our results for Afghanistan in Section 4.1 agree with this "group" coding even though we did not use this coding in our analysis. This confirmation of the WITS coding was not a foregone conclusion because there is substantial controversy regarding the structure of the insurgency in Afghanistan. In fact, we have shown in Section 5.1 that we actually *disagree* with a largely similar "Taliban" coding when it is used by the GTD in 2010-

2016. Furthermore, *contra* WITS, the GTD codes the hard-line Islamist "Haqqani Network" as a distinct group, and there is thus disagreement even between databases.

Much discussion in the counterinsurgency literature has been dedicated to the Haqqani Network and whether it is part of the Taliban proper or is instead an independent entity with links to Pakistan and its security services. Appendix Figure M.6 reproduces Figure 1 from Jones [2008], a well-reputed study of insurgency in Afghanistan during the period covered by WITS. Maps similar to Appendix Figure M.6 are frequently digitized and employed for spatial conflict analysis, much like the Murdock [1959] ethnic maps are used in studies of ethnicity and conflict [König et al., 2017].

The Haqqani Network is given a distinct geographic territory in Appendix Figure M.6, and is coded in the Global Terrorism Database as being responsible for 77 attacks, including some very large simultaneous attacks. On the other hand, the Haqqani Network does not appear in the WITS group coding, and the son of the founder made the following remarks to BBC Pashto on October 3, 2011:

> *[Siraj Haqqani] pledged loyalty to [Taliban Leader] Mullah Omar, saying he "is our leader and we totally obey him." "In every operation we get the order, planning and financial resources from the [Taliban] Emirate's leadership and we act accordingly," Mr Haqqani said.*

Claims of responsibility occur frequently in the media in the aftermath of violent incidents or in the context of strategic manipulation of the conflict narrative.[42] Incentives for truthful communication, as opposed to over-claiming for psychological effect,

[42] For example, the US government has classified the Haqqani Network separately as a terrorist organization, and specifically blames it for certain attacks. Consider a simultaneous attack reported by Reuters on April 16, 2012:

> *The Taliban claimed sole responsibility for the attacks ... Afghan and U.S. officials have blamed the attacks on the al Qaeda-linked Haqqani network, based along the porous Afghan-Pakistan mountain border...*
>
> *'The attacks were very successful for us and were a remarkable achievement, dealing a psychological and political blow to foreigners and the government,' [Taliban spokesman] Mujahid said. '... the Haqqanis are part of the Taliban ... This is a baseless plot from the West, who wants to show that we are separate.'*

Our results agree with the Taliban position regarding this attack. In general, given a sufficiently large data set of explicit, but unverified, claims by different group leaders and corresponding attack data, our methods can provide a quantifiable and non-subjective metric of reliability, furthering the conflict literature focused on strategic communication.

are ambiguous for insurgents. To the best of our knowledge, the conflict literature lacks objective and replicable quantitative approaches designed to assess this type of messages: our methods help fill this gap.

Based on Table 1 and Figure 3, we prefer the WITS coding that includes Haqqani with the Taliban to the GTD coding of a separate Haqqani Network. In general, the contrast between the results of our method and other sources can provide insight regarding the degree to which quantitative evidence actually supports various qualitative reports, and our methodology can be employed as a tool for an objective assessment of claims made by insurgents and counterinsurgents.

6 Prevalence of Coordination

In Sections 4 and 5 we employed the presence of simultaneous attacks to study the organization of insurgency in Afghanistan and Pakistan. A remaining question is how prevalent these sort of attacks are and whether further insight can be gathered from their empirical properties.

Suppose that our data actually contained no planned simultaneous attacks at all. According to our model in Section 2, the fact that more than one attack occurred on a given day would then be due purely to random chance, and the number of attacks on any day in district i would follow a Binomial(ℓ_i, η) distribution. If we assume that the number of potential disorganized insurgents that could launch attacks is large, and the probability η of any one of them launching an attack is low, then this distribution should be approximately $x_{it} \sim$ Poisson $(\eta \ell_i)$. The total number of attacks observed on any day across all districts will then be Poisson $(\eta \sum_i \ell_i)$. Within our model, consider for example the simple case of two potential values of $\epsilon_{jt} \in \left\{\epsilon^h; \epsilon^l\right\}$ occurring with probability p_{ij} and $1 - p_{ij}$ respectively. Overdispersion in the presence of a single insurgent group j would then take the mixture form

$$x_{it} \sim p_{ij} Poisson\left(\epsilon^h \alpha_{i,j} + \eta \ell_i\right) + (1 - p_{ij}) Poisson\left(\epsilon^l \alpha_{i,j} + \eta \ell_i\right).$$

We can thus check for the existence of simultaneous attacks by looking for overdispersion in the observed pattern of attacks relative to a Poisson. According to the Cameron and Trivedi [1990] test, the distribution of attacks is overdispersed relative to a Poisson distribution with $p < 0.01$ both in Afghanistan and in Pakistan.[43] We can thus reject with high statistical confidence the absence of coordinated attacks in both countries.

We are further interested in assessing the quantitative importance of simultaneous organized attacks. The traditional definition of overdispersion refers to additional variance above and beyond what expected from a Poisson distribution, but this does not have an easy interpretation as a specific quantity of coordinated attacks. We will thus use a non-traditional definition of overdispersion, one that provides an estimate of how many simultaneous attacks there are in our dataset.

An overdispersed Poisson distribution will have more high realizations (days with a large number of attacks) compared to a Poisson distribution with the same mean. These additional high realizations correspond to the simultaneous attacks that are of interest.[44] Our measure will take the actually observed distribution of attacks across days and compare it to the theoretical Poisson distribution with the same mean.

Let $g(x) = (\bar{f}(x) - f(x))x$, where f is the theoretical probability of observing x attacks in a day, given a Poisson with the mean equal to the actual mean number of attacks per day. Let $\bar{f}$ be fraction of days when x attacks occurred in the empirical distribution. Thus, $g(x)$ is the excess number of attacks in the empirical distribution, considering only days where there were exactly x attacks. Let $G(x) = max_x \sum_{r=x}^{\infty} g(r)$, where the object being maximized is the excess number of attacks in the situation where there were x or more attacks in a day. We will define the number of excess attacks of interest as $G(x)$.

Using this definition, in our Afghanistan data 4% of all attacks are simultaneous attacks of interest, and in our Pakistan data, 9%. These figures point to the use of coordinated attacks in between one in twenty and one in ten episodes of violence for our samples. In addition, such attacks appear more deadly (even per individual incident) than regular attacks and are also more visible and traumatic for the civilian population (consider for example the 9/11, Mumbai, and Bataclan attacks). Our analysis in this paper has thus focused on a quantitatively relevant component of violence within these countries, comprising several hundred recorded incidents.

[43] Including month fixed effects gives a p-value of approximately 0.002.

[44] That is, if there is a group that has an ϵ that is constant across all days, then it will lead to no overdispersion and thus, according to our definition, no simultaneous attacks of interest. There are attacks that occur on the same day by random chance: it is just that the fraction of days where there are two attacks will be that of a poisson distribution, which is our baseline, and thus there are no *additional* attacks due to coordination within the group.

This evaluation of coordination relies on the Poisson approximation to our particular model of attacks. One might be concerned that the actual attack structure is, for some unknown reason, not an overdispersed Poisson distribution. To provide direct verification of the percentages reported above, we use attacks recorded in the Global Terrorism Database (GTD).

For our periods of interest, there are $1,692$ attacks recorded in Afghanistan and $2,610$ in Pakistan. For each attack, the GTD records whether there are other "related" attacks: these are generally attacks that occurred the same day but in a different location.[45] Importantly, this variable is not coded mechanically. Attacks are only coded as "related" if there actually appears to be evidence of intentional relationship between the attacks. For our period, 4% of the attacks in Afghanistan and 10% of the attacks in Pakistan have "related" attacks in the GTD, for 511 attacks in total. This is extremely close to our 4% and 9% estimates above based on overdispersion of WITS and BFRS, validating our econometric approach.[46]

One might further wonder whether this tight relationship between overdispersion and related attacks holds for countries other than Afghanistan and Pakistan. We consider all countries listed in the GTD, and calculate the percentage excess of attacks in each country using the GTD data. Figure N.7 shows a clear relationship between our measure of overdispersion and the fraction of attacks that are coded as related. Table N.8 considers regressions using our measure of overdispersion coded at the country-year level, which allows us to include country fixed effects. We see that in years in which there are more overdispersed attacks in a given country, that country is more likely to have a greater fraction of their attacks coded as related.

The GTD also proves extremely useful in performing a number of validity checks of our main approach. These include: (i) identification of the insurgent groups based on GTD perpetrator information; (ii) assessment of the insurgent structures we uncover for time periods beyond the ones considered in the main analysis; (iii) ruling out that coordinated attacks take place over time periods longer than the single day; (iv) evidence that a single group and not multiple coordinated groups is typically behind a set of simultaneous attacks; (v) evidence of the extent of credit claiming for

[45] The GTD codebook technically allows for "related" attacks that do not occur on the same day, but we confirm in Appendix N that here are no related attacks of this type in our sample.

[46] Although the percentages are almost the same, our WITS and BFRS figures are based more than three times as many attacks, and imply that we consider as coordinated hundreds of additional attacks relative to the GTD.

coordinated attacks; (vi) support for the assumption of a trade-off between military value opportunity and signaling value of attacks; (vii) panel data evidence of the relationship between group strength and extent of coordination. This additional analysis is available in Appendix N, where all details on specifications and interpretation are provided.

7 Conclusions

This paper focuses on the empirical analysis of insurgency, with applications to Afghanistan and Pakistan. Often the only type of information available about the level and geographic diffusion of insurgent activity comes from incident-level attack data. Recent advances in the analysis of the economics of conflict and post-war reconstruction have been possible thanks to this data, however limited it might be.[47]

Progress in understanding insurgency seems key in furthering knowledge of the determinants and consequences of political violence in developing countries. Although much of the analysis in this paper is necessarily context-dependent, it is informative nonetheless for regional stabilization and local development goals. From a methodological perspective, our contributions have a more general appeal and, as the availability of microlevel datasets expands within the conflict literature, they may show promise in other environments.

Some of the applications we have discussed may find a useful role in the study of crime, especially for the case of criminal organizations. Outside of research in political economy, the methods we propose based on conditional covariance structures across units may be applicable in the field of industrial organization. Examples may include the detection of collusion, price fixing, and of horizontal anticompetitive behavior across firms, within and across markets, or the estimation of unobserved networks amongst competitors.

REFERENCES

[1] Ahn, S.; Horenstein, A. (2013) “Eigenvalue Ratio Test for the Number of Factors. ”Econometrica. 81(3): 1203-1227.

[47] Berman, Shapiro, and Felter [2011] and Trebbi, Weese, Wright and Shaver [2017] are recent examples.

[2] Alquist, R.; Coibion, O. (2014) “Commodity-Price Comovement and Global Economic Activity ”. NBER Working Paper 20003. March 2014.

[3] Ashford, J.R. and R.G. Hunt (1973) “The Distribution of Doctor-Patient Contacts in the National Health Service” Journal of the Royal Statistical Society Series A 137 (3), 347-383.

[4] Barno, David (2006) Challenges in Fighting a Global Insurgency Parameters. Summer 2006. 15-29.

[5] Besley, T, M Reynal-Querol. (2014). The legacy of historical conflict: Evidence from Africa. American Political Science Review, 108(2), pp. 319-336.

[6] Bleaney, M.; Mizen, P.; Veleanu, V.2012). “Bond Spreads as Predictors of Economic Activity in Eight European Economies .” University of Nottingham, Centre for Finance, Credit and Macroeconomics (CFCM) Discussion Paper. December 2011.

[7] Benmelech, Efraim, Claude Berrebi, and Esteban F. Klor. (2012). “Economic Conditions and the Quality of Suicide Terrorism.” The Journal of Politics 74 (1): 113–128.

[8] Berman, Eli (2009). Radical, Religious and Violent: The New Economics of Terrorism. MIT Press.

[9] Berman, Eli, Joseph H. Felter, Jacob N. Shapiro, (2011) Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq. Journal of Political Economy Vol. 119, No. 4: 766-819

[10] Berman, Eli, Aila Matanock. (2015). “The Empiricists’ Insurgency.” Annual Review of Political Science, Vol 18: 443-464.

[11] Berman, Nicolas, Mathieu Couttenier, Dominic Rohner, and Mathias Thoenig. (2017). “This Mine Is Mine! How Minerals Fuel Conflicts in Africa.” American Economic Review, 107(6): 1564-1610.

[12] Besley, Tim, Torsten Persson. (2017). “The Joint Dynamics of Organizational Culture, Design, and Performance” mimeo LSE.

[13] Birgin, E.; Martinez, J.M.; Raydan, M. (2000). “Nonmonotone Spectral Projected Gradient Methods on Convex Sets.” SIAM J. Optim.. 10(4): 1196–1211.

[14] Blattman, Christopher and Edward Miguel (2010) “Civil War” Journal of Economic Literature 2010, 48:1, 3–57

[15] Brahimi, A. (2010). “The Taliban’s Evolving Ideology.” Working Paper. LSE Global Governance. WP 02/2010.

[16] Bruzzese, D., Vistocco, D. (2015). “DESPOTA: DEndrogram Slicing through a PemutatiOn Test Approach”. Journal of Classification. 32(2):285-304.

[17] Bueno de Mesquita, Ethan. (2013). “Rebel Tactics.” Journal of Political Economy 121 (2): 323–357

[18] Bueno de Mesquita, Ethan, and Eric S. Dickson. (2007). “The Propaganda of the Deed: Terrorism, Counterterrorism, and Mobilization.” American Journal of Political Science 51 (2): 364–381.

[19] Cameron, A.C. and Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. Journal of Econometrics, 46, 347-364.

[20] Christia, Fotini , Semple, Michael (2009). Flipping the Taliban- How to Win in Afghanistan. Foreign Affairs, 88, 34-45

[21] Christia, Fontini (2012). Alliance Formation in Civil Wars. Cambridge UP.

[22] Condra, Luke N., Jacob N. Shapiro, (2012) Who Takes the Blame? The Strategic Effects of Collateral Damage. American Journal of Political Science Vol. 56, No. 1: 167-187.

[23] Deloughery Kathleen (2013) Simultaneous Attacks by Terrorist Organisations. Perspectives on Terrorism, 7(6): 79-90.

[24] Ding, C., He, X., and Simon, H. (2005). On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. Proceedings of the Fifth SIAM International Conference on Data Mining, 606-610.

[25] Dorronsoro, Gilles (2009). The Taliban’s Winning Strategy in Afghanistan. Carnegie Endowment for International Peace Paper.

[26] Ellison, Glenn and Edward L. Glaeser (1997). Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. Journal of Political Economy 105(5): 889-927.

[27] Fearon, James (2008). Economic development, insurgency, and civil war in Institutions and Economic Performance, ed. Elhanan Helpman, Harvard University Press

[28] Ferson, W.; Kim, M. (2012). The factor structure of mutual fund flows. International Journal of Portfolio Analysis and Management, 1(2), 112-143.

[29] Ghobarah, Hazem Adam, Paul Huth and Bruce Russett. (2003) Civil Wars Kill and Maim People Long After the Shooting Stops. American Political Science Review 97(2):189-202.

[30] Giustozzi, Antonio (2009). The Pygmy who turned into a Giant: The Afghan Taliban in 2009. LSE mimeo.

[31] Guo-Fitoussi, L.; Darne, O.. (2014) “A Comparison of the Finite Sample Properties of Selection Rules of Factor Numbers in Large Datasets”. HAL Working Paper hal-00962247. March 2014.

[32] Gutierrez-Sanin, Francisco. (2008) Telling the Difference: Guerrillas and Paramilitaries in the Colombian War. Politics and Society 36(1):3-34.

[33] Henzel, S.; Rengel, M. (2014) Dimensions of Macroeconomic Uncertainty: A Common Factor Analysis. . SSRN Scholarly Paper 2507743.

[34] Hornik, K.; Feinerer, I.; Kober, M.; Buchta, C. (2012). Spherical k-Means Clustering. Journal of Statistical Software. 50(10).

[35] Horowitz, Michael C. and Potter, Philip (2014) Terrorist Intergroup Cooperation and the Consequences for Lethality. Journal of Conflict Resolution. 58(2): 199-225.

[36] Huang, K., Sidiropoulos, N. (2014) “Putting nonnegative matrix factorization to the test: a tutorial derivation of pertinent cramer-rao bounds and performance benchmarking.” IEEE Signal Processing Magazine. 31(3):76–86.

[37] Huang, K., Sidiropoulos, N., and Swami, A. (2014) Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition. IEEE Transactions on Signal Processing 62(1):211-224.

[38] Khuram Iqbal and Sara De Silva (2013) Terrorist lifecycles: a case study of Tehrik-e-Taliban Pakistan. Journal of Policing, Intelligence and Counter Terrorism. Vol. 8, No. 1, 72-86.

[39] Jenkins, Brian Michael (2014) The Dynamics of Syria’s Civil War, RAND Perspectives no. 115.

[40] Jones, Seth G. (2008). The Rise of Afghanistan’s Insurgency, International Security, 32(4), pp. 7-40.

[41] Karlis, D., Xekalaki, E. (2005) “Mixed Poisson Distributions”. International Statistical Review. 73(1):35-58.

[42] Kilcullen, David (2009) The accidental guerrilla: Fighting small wars in the midst of a big one. Oxford University Press

[43] König, Michael D., Dominic Rohner, Mathias Thoenig and Fabrizio Zilibotti, 2017 ”Networks in Conflict: Theory and Evidence from the Great War of Africa.” Econometrica.

[44] Krishna, K.; Narasimha, M. (1999). ”Clustering High Dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering”. Trans. Sys. Man Cyber. Part B. 29 (3): 433–439.

[45] Ludhianvi, M.R. (2015). Obedience to the Amir: An early text on the Afghan Taliban Movement. Trans. Y. Mitha and M. Semple. Berlin: First Draft Publishing.

[46] Luxburg, Ulrike von (2007) “A tutorial on spectral clustering” Statistics and Computing Volume 17, Issue 4, pp 395-416

[47] Marchenko, V.A.; Pastur, L.A. (1967). ”Distribution of Eigenvalues for Some Sets of Random Matrices”. Matematicheskii Sbornik. 72 (114): 507–536

[48] Mirza, H.; Storjohann, L. (2014). ”Making Weak Instrument Sets Stronger: Factor-Based Estimation of Inflation Dynamics and a Monetary Policy Rule.” Journal of Money, Credit and Banking. 46(4): 643-664.

[49] Munir, M. (2011). ”The Layha for the Mujahideen: an analysis of the code of conduct for the Taliban fighters under Islamic law ”. International Review of the Red Cross. Vol. 93 No. 881.

[50] Murdock, George (1959). Africa: Its Peoples and their Culture History. McGraw-Hill Inc.

[51] O’Neill, Bard (1990) Insurgency and Terrorism, Inside Modern Revolutionary Warfare, Dulles, VA.: Brassey’s Inc.

[52] Pak Institute for Peace Studies (2016) Pakistan Security Report 2015.

[53] Raleigh, C., Witmer, F., O’Loughlin, J. and Denemark, R.A., (2010). A review and assessment of spatial analysis and conflict: The geography of war. The international studies encyclopedia, 10, pp.6534-6553.

[54] Rezitis, A.N. (2015). “Empirical Analysis of Agricultural Commodity Prices, Crude Oil Prices and US Dollar Exchange Rates Using Panel Data Econometric Methods.” SSRN Scholarly Paper 2631534. July 2015.

[55] Schelling, Thomas C. (1960) The Strategy of Conflict. Cambridge: Harvard University Press.

[56] Ben Smith (2005) Afghanistan: Where Are We? Conflict Studies Research Center, Central Asia Series. Report 05/30. June 2005.

[57] Spence, M. (1973). “Job Market Signaling.” *The Quarterly Journal of Economics* 87(3): 355374.

[58] Thruelsen, Peter Dahl (2010) “The Taliban in southern Afghanistan: a localised insurgency with a local objective” Small Wars & Insurgencies, Volume 21, Issue 2, pp.259-276

[59] Trebbi, Francesco, Eric Weese, Austin Wright, Andrew Shaver. 2017. “Insurgent Learning” NBER WP 23475

[60] Tullock, Gordon (1974) The Social Dilemma, Blacksburg: Center for the Study of Public Choice, VPISU Press.

[61] United Nations (2013) Third report of the Analytical Support and Sanctions Monitoring Team, submitted pursuant to resolution 2082 (2012) concerning the Taliban and other associated individuals and entities constituting a threat to the peace, stability and security of Afghanistan. S/2013/656

[62] United Nations (2016) “Humanitarian Response Plan - Syrian Arab Republic” United Nations Office for the Coordination of Humanitarian Affairs.

[63] Weidmann, N.; Rod, J.K.; Cederman, L.E. (2010) “Representing ethnic groups in space: A new dataset.” Journal of Peace Research, 47(4), 491–499.

[64] Wigner, E.P. (1955) “Characteristic Vectors of Bordered Matrices With Infinite Dimensions”. The Annals of Mathematics, 62(3), 548–564.

Online Appendices – Not For Publication

A Insurgency Organization & Economic Recovery

This section briefly discusses case studies chosen to highlight the economic importance of understanding insurgent organization in conflict and post-conflict environments. We focus on three different episodes: Iraq, Syria, and Libya.

Insurgent groups owe their success to their deep ties with noncombatant populations. By impeding reconstruction efforts, they can fuel popular dissatisfaction with central authorities, thereby maintaining a steady flow of recruits and ensuring logistic assistance for their agents. Insurgencies thus have a particular incentive to delay aggregate economic recovery.

In Iraq, insurgents disrupted the electricity grid and seized control of oil resources. Henderson [2005] describes the loop that linked insecurity and economic stagnation:

> *Inability to provide security had a profound impact on Iraq's economic recovery. In turn, inability to provide recovery had a profound impact on Iraq's security. Reconstruction delays fed into Iraqi feelings of resentment and despair, which fueled insurgency and crime, thereby worsening the security climate.*

The connection of the study of insurgency with economic development comes from this tight link between insurgent strategies and the failure of reconstruction efforts. Understanding the exact nature of the Iraqi insurgency early on in the conflict could have proven crucial in breaking the vicious cycle that Henderson [2005] observes.[48]

Uncertainty about the organization of the insurgency in post-2003 Iraq took several forms. First, there was disagreement regarding the extent to which attacks represented an insurgency at all.[49] There was also confusion regarding its magnitude: as late as the fall of 2004, the U.S. military still attributed 80 percent of attacks to random and

[48] Henderson is critical of the strategy actually used: *"as violence worsened, the response of coalition officials in charge of reconstruction was not to find a way to fight it more effectively. Instead, their response was to withdraw into the heavily protected world of the Green Zone."*

[49] Eisenstadt and White [2005] write that *"In the summer of 2003, Secretary of Defense Donald Rumsfeld and General John Abizaid (head of U.S. Central Command) publicly disagreed about whether the violence in the Sunni Triangle was the final act of former regime "dead-enders" or an incipient insurgency against the emerging political order"*. There was a similar disagreement in 2005 between Vice President Richard Cheney and General Abizaid.

not to political violence. Finally, there was heated debate about the organization of the insurgency, once it was clear that one existed.[50] Further complexity in the Iraqi case stemmed from signs of evolution over time, as the New York Times reported: *"the insurgency was now organized regionally, and that evidence pointed to some planning across regional boundaries".*[51]

The difficulty, and the importance, of understanding the structure of insurgencies is not limited to Iraq. Consider recent Western efforts in Syria: *"Sixteen months into the uprising in Syria, the United States is struggling to develop a clear understanding of opposition forces inside the country, according to U.S. officials who said that intelligence gaps have impeded efforts to support the ouster of Syrian President Bashar al-Assad."*[52]

Beginning with a series of pro-democracy protests in 2011, the situation in Syria quickly escalated into a full-blown civil war that has cost $250,000$ lives and displaced almost 11 million Syrian citizens to the beginning of 2016. In the backdrop of a ethnically and religiously divided population, this conflict quickly displayed a high degree of complexity in the heterogeneity of parties involved [Smith, 2012], including the Syrian state army loyal to Bashar al-Assad, Sunni Syrian rebels, the Islamic State, Jabhat al-Nusra, Kurdish forces, and Hezbollah. Lack of understanding of the structure of the insurgency in Syria has been one of the strongest deterrents to military and humanitarian involvement of Western powers in this conflict [Jenkins, 2014] and slowed down relief efforts.

Western countries were willing to lend support and provide prompt international aid to moderate Sunni organizations, but the difficulty laid in identifying these rebels and their true organizational linkages. The impossibility of separating the secular moderates from the religious extremists among the Sunni opponents of the Alawite-led government resulted in international paralysis. This led to further economic and social deterioration, radicalization, and escalation of the conflict. Syria is now a nearly failed state, fought over by Assad loyalists, the Islamic State, and the al-Qaeda

[50] The New York Times quotes senior U.S. intelligence sources stating that "*It's not just one group of insurgents rallying under one cause. It's multiple groups with different causes loosely tied together by the threads of anti-U.S. sentiment, some sort of Iraqi nationalism, Muslim-Arab unity or greed*". The lack of familiarity with this type of enemy appeared evident: *"What makes it more difficult is that you're dealing with an insurgency without a single face".*

[51] http://www.nytimes.com/2004/10/22/international/middleeast/22insurgents.html?pagewanted=2&_r=0

[52] http://www.washingtonpost.com/world/national-security/in-syria-conflict-us-struggles-to-fill-intelligence-gaps/2012/07/23/gJQAW8DG5W_story.html

affiliated Nusra front. Numerous attempts at a political solution by the Arab League and the United Nations have failed.

Another relevant case is Libya post-Colonel Gaddafi. This event would require in itself a fully accurate discussion, but as above for Iraq and Syria, we try to provide a basic picture from the perspective of the analysis of multi-group conflicts. After 2011 and the violent overthrowing of the Gaddafi regime, Libya gradually descended into full-blown factional violence with Islamic State factions jockeying for control of oil rich areas together with two main armed groups: the Tobruk government (elected democratically but in a deeply unstable political environment) and the Muslim Brotherhood-supported General National Congress. To further complicate the picture, other ethnic-based groups, like the Touareg, have also laid claim to certain parts of the former Libyan state. Repeated failures to achieve stable Unity governments and substantial ambiguity in the set of alliances struck among the various groups have severely hindered the pacification response led by the United Nations in the region. While the United Nations and the European Union have been holding off decisive intervention, the east/west divide in the country has been increasingly exacerbating.

B Simultaneous Attacks: Theoretical Framework and Qualitative sources

An insurgent group typically operates from an asymmetric position and does not usually aim for military victory over its adversary [Kilcullen 2009]. Baloch separatists need only convince the Pakistani government to allow an independent Balochistan, not necessarily topple the government. A group that appears strong, however, will have greater negotiating power vis-à-vis its opponent. It will also have more success in recruitment and fundraising, as the noncombatant population is more likely to side with a strong group. Launching simultaneous attacks is a signal of strength, because such an attack requires coordination.

The basic idea for our theoretical framework is provided by Shapiro [2013]: insurgent groups face a trade-off between their degree of internal control and the safety of their members, because the mere act of communicating makes members more vulnerable to detection by government forces. Suppose that some particularly effective insurgent groups have managed to develop and maintain secure communication channels, while other groups are plagued by government moles and eavesdroppers. Insurgents benefit from the support of the civilian population for recruitment and fundraising, and civilians are more interested in supporting well-organized and effective groups than failing ones. In cases such as Afghanistan, insurgents also benefit from convincing foreign civilians of their strength, as these foreign civilians then pressure their governments to withdraw troops from an "unwinnable" conflict. Civilians do not know exactly how strong the insurgents are, and insurgents thus wish to somehow signal that their organization is strong and uncompromised, both to win local support and to force foreign withdrawal.

A simultaneous attack necessarily involves communication in order to coordinate the attack.[53] If a weak insurgent group is vulnerable to government surveillance when it attempts to communicate, while a strong group has successfully developed communication methods that escape detection, then a simultaneous attack is costlier for the weak group due to the exposure of its members. Simultaneous attacks thus fit into the standard Spence [1973] signaling framework: such an attack is a credible signal of strength because launching it is less costly for the strong group than the

[53] Shapiro and Siegel [2015] discuss how insurgent coordination is achieved through mobile phone communication and ICT.

weak group.

The qualitative literature supports the idea that simultaneous attacks have a signalling motivation. For example, Barno [2006] gives a specific example of a simultaneous attack on three border checkpoints where the media was deliberately alerted to the attack and publicity appears to have been the main objective. Deloughery [2013] provides a recent review of the literature and presents systematic evidence of the advantages of simultaneous attacks for terrorist organizations in terms of psychological warfare, media coverage and appeal in the recruitment of new fighters, incentives that operate within insurgencies as well.[54]

In reality insurgent groups launch a mix of simultaneous and individual stand-alone attacks. We posit that this is because there is a trade-off between the signalling value of attacking simultaneously in many districts versus the military value of attacking separately in each district at the most opportune moment for that district. In Appendix N we discuss this hypothesis further and support it with regression evidence. We also check implications of the signalling model just outlined above: for example, it appears that (both in Afghanistan and in a cross-country sample) insurgents are less likely to launch simultaneous attacks relative to stand-alone attacks in areas where they have a limited presence and are thus potentially more vulnerable.

Our main objective in the paper is to use the fact that insurgent groups do launch simultaneous attacks in order to identify the number of such groups and their geographic extent. We do not formalize the above signalling model of attacks –the framework is standard. Instead, in Section 2 we build an econometric model of simultaneous attacks based on the assumption that all groups launch such attacks to at least some degree.[55]

From a Western perspective, the 9/11 attacks in the United States are the most obvious example of the salience of such simultaneous violence, but the phenomenon is widespread. For example, in southern Thailand insurgent movements have adopted similar tactics: “On April 28, 2004 groups of militants gathered at mosques in Yala, Pattani, and Songkhla provinces before conducting simultaneous attacks on security checkpoints, police stations and army bases” [Fernandes, 2008]. The Indian Mu-

[54] According to Kilcullen [2009], “the insurgents treat propaganda as their main effort, coordinating physical attacks in support of a sophisticated propaganda campaign” (p. 58). See also Arce and Sandler [2007]. Additional references for the qualitative literature are also provided in Appendix B.

[55] In the model presented below there are disorganized individual insurgents who attack randomly, and thus even a particularly weak insurgent group would have an incentive to launch the occasional simultaneous attack, in order to distinguish themselves from these “lone wolf” actors.

jahideen, responsible for the 2008 Mumbai attacks, typically carry out simultaneous attacks [Subrahmanian et al., 2013]. Kurdish nationalists and the Tamil Tigers are known to have adopted simultaneous attacks as a strategy. In Africa, Boko Haram in northern Nigeria has carried out coordinated attacks on multiple targets such as churches, and Anderson [1974] describes coordinated attacks in Portuguese colonies. Simultaneous attacks and suicides have been a trademark of international jihadist organizations and of al-Qaeda in particular, making our approach well-suited to the Afghan insurgency case. Because the empirical covariance matrix of attacks is observed, these assumptions implying positive covariances driven by co-occurring incidents are readily verifiable and they are in fact supported by the data. See discussion at the end of Section 2.

C Decomposition of Covariance Matrix

Let $\gamma_{ii'} = \sum_j \alpha_{ij}\alpha_{i'j}$ denote the off-diagonal entry on row i and column i' of Γ_L. Let $\bar{\gamma}_{ii'}$ be the corresponding entry of the covariance matrix in the observed sample. Unfortunately, no empirical counterpart to Γ_L is observed, and thus one will have to be created by modifying the diagonal of the observed covariance matrix $\bar{\Gamma}$.

To create a $\hat{\Gamma}_L$ from $\bar{\Gamma}$, a diagonal matrix $\hat{\Gamma}_D$ will be subtracted from the latter to produce the former. An intuitive method for doing this is "trace minimization", discussed at least as early as Ledermann [1940]. First, note that $\bar{\Gamma}$ is a (sample) covariance matrix, and is thus positive semi-definite. $\hat{\Gamma}_L$ should also correspond to a covariance matrix, and thus should also be positive semi-definite. Consider the optimization problem

$$\begin{aligned} &\min_{\hat{\Gamma}_D} \mathrm{Tr}(\hat{\Gamma}_L) \\ &\text{s.t. } \hat{\Gamma}_L = \bar{\Gamma} - \hat{\Gamma}_D, \quad \hat{\Gamma}_D \text{ diagonal}, \\ &\qquad \hat{\Gamma}_D \succ 0, \hat{\Gamma}_L \succ 0 \end{aligned} \tag{10}$$

Here Tr() denotes the sum of diagonal entries of a matrix, and $\succ 0$ indicates positive semi-definiteness. The intuition for trace minimization is that the "extra" variance present in the diagonal entries of Γ has the form of a full rank matrix, and thus in order to recover a low rank matrix such as Γ_L, as much of this as possible needs to be removed.

Saunderson et al. [2012] show that the intuition of Ledermann and others was correct in general. Specifically, the positive semi-definite matrix Γ_L can be recovered given Γ so long as it is sufficiently "incoherent", and this property is satisfied by most low rank matrices. Details are provided in Appendix C.1.

If $N = 200$, the the semi-definite program corresponding to (10) involves $200 \times 199 = 39,800$ constraints: each off-diagonal entry $\bar{\gamma}_{ii'}$ in the positive semi-definite matrix $\bar{\Gamma}$ must be equal to the corresponding entry in $\hat{\Gamma}_L$. Problems of this size are feasible using modern semidefinite programming algorithms. We thus compute $\hat{\Gamma}_L$ using (10), and will use it as the basis for producing an estimate of insurgent group presence in the next two subsections.

C.1 Recoverability of Low-Rank Matrix

We are interested in the conditions under which the $\hat{\Gamma}_L$ resulting from (10) will be a consistent estimator for Γ_L. It is clear that there are some matrices Γ_L for which the proposed method will be inconsistent:

Example 1. *Suppose that there are three districts, and two groups. Group memberships are $\alpha_{\cdot 1} = (1, 0, \delta)$ and $\alpha_{\cdot 2} = (0, 1, \delta)$, and thus*

$$\Gamma_L = \begin{bmatrix} 1 & 0 & \delta \\ 0 & 1 & \delta \\ \delta & \delta & 2\delta^2 \\ & & \end{bmatrix}$$

for some small value δ. Suppose that there are disorganized insurgents such that $\Gamma_D = I_3$. The minimum trace heuristic of (10), will then give an estimate

$$\hat{\Gamma}_L = \begin{bmatrix} \delta & 0 & \delta \\ 0 & \delta & \delta \\ \delta & \delta & 2\delta \\ & & \end{bmatrix}$$

which has lower trace than the true Γ_L so long as δ is small.

It is thus important to provide conditions for the matrix Γ_L such that the proposed method gives a consistent estimator. Saunderson et al. [2012] give such a characterization. First, Saunderson et al. [2012] define a subspace $\mathcal{U}$ as realizable if, for any Γ_L having column space $\mathcal{U}$, and any Γ_D, the minimum trace factorization algorithm of (10) applied to $\Gamma = \Gamma_D + \Gamma_L$ returns $\hat{\Gamma}_L = \Gamma_L$. Next, they define the "coherence" $\mu(\mathcal{U})$ of a subspace $\mathcal{U}$ of $\mathbb{R}^n$ as

$$\mu(\mathcal{U}) = \max_{i \in \{1,2,\ldots n\}} ||P_{\mathcal{U}} e_i|| \tag{11}$$

where e_i are the standard basis vectors, and $P_{\mathcal{U}}$ is the orthogonal projection matrix onto $\mathcal{U}$. They then provide the following sufficient condition:

Theorem 2 (Saunderson et al. 2012)**.** *If $\mathcal{U}$ is a subspace of $\mathbb{R}^n$ and $\mu(\mathcal{U}) < 1/2$, then $\mathcal{U}$ is realizable.*

From an intuitive perspective, this restriction on coherence is equivalent to nothing in the column space of Γ_L being too close to the standard basis vectors. In the context of estimating insurgent groups, the standard basis vectors represent groups that are only present in one district. It makes sense that groups of this sort will result in the procedure in (10) being inconsistent: a group that is only present in one district is indistinguishable from disorganized insurgents, as they both only appear in the diagonal entries of the covariance matrix.

Saunderson et al. [2012] also provide a further result, regarding the "realizability of random subspaces". They argue that "most" subspaces of dimension less than $n/2$ are realizable. The intuition here appears to be that a random subspace of low dimension is unlikely to include anything close to a standard basis vector. In general, then, if the number of groups is small relative to the number of districts, the heuristic given in (10) will provide a consistent estimator for the group structure. Cases where the estimator will not be consistent are those where one of the groups is overwhelmingly located in a single district.

D Spectral Clustering Estimator

Spectral clustering is based on the "graph Laplacian" matrix

$$L = D - \Gamma_L \tag{12}$$

where D is a diagonal matrix with entries equal to the row sums of Γ_L. The graph Laplacian thus has off-diagonal entries equal to the negative of those of the adjacency matrix, and diagonal entries such that all rows and columns sum to zero. The graph Laplacian L has a rank of $N - J$, and thus has J zero eigenvalues.[56] Spectral clustering focusses on the number of zero eigenvalues for the associated graph Laplacian matrix L, whereas the method used in the main text produces an estimate $\hat{J}$ of the number of insurgent groups by examining (in a very broad sense) the rank of Γ_L.

If Γ_L were known, the number of organized groups could be calculated immediately, and it would equal both the rank of Γ_L and the number of zero eigenvalues of L. However, the data available gives the sample covariances $\bar{\gamma}_{ii'}$ rather than the true $\gamma_{ii'}$, and thus a noisy $\hat{\Gamma}_L$ must be used instead of the true Γ_L. The simplest option for actually implementing a spectral clustering approach is to use a modification of Shi and Malik [2000]: use $\bar{\Gamma}$ to construct $\bar{L}$, and then count the "zero" eigenvalues of $\bar{L}$.

In a finite sample, however, these eigenvalues calculated from $\bar{L}$ are subject to finite sample variation. In particular, random variation will result in positive $\bar{\gamma}_{ii'}$ entries in some cases where the true $\gamma_{ii'}$ is zero, and negative $\bar{\gamma}_{ii'}$ entries in some cases where the true $\gamma_{ii'}$ is positive. This random variation will tend to increase the rank of the $\bar{L}$ relative to L. This problem is particularly severe for districts i for which there are few attacks: the data provides little information on the group structure in these districts, and if one object of interest is J, the total number of groups, the inclusion of these particularly noisy districts could result in a substantial amount of additional noise in the estimate $\hat{J}$.

[56] The number of zero eigenvalues of the graph Laplacian matrix corresponds to the number of connected components of the weighted undirected graph described by the adjacency matrix Γ_L. This is J, the number of blocks of Γ_L.

The intuition for this result is relatively straightforward. Each Γ_L^j block has rank one. The corresponding block of the diagonal matrix D has full rank. Setting the entries in this diagonal matrix so that rows and columns of the graph Laplacian L sum to zero ensures that the rows (and columns) of L corresponding to each Γ_L^j block are linearly dependent. The Γ_L^j block that was subtracted, however, is only rank one, and thus the null space of the resulting block of L must be rank one. This is true for every block in L, and thus the null space of L has dimension J. This will also be the number of zero eigenvalues of L.

A similar problem affects the approach presented in the main text, which is based on using the the largest eigenvalues (or other components) of $\hat{\Gamma}_L$. Finite sample variation will also affect these eigenvalues. The question thus arises whether it is better to use $\hat{\Gamma}_L$ directly, or instead use the corresponding graph Laplacian matrix L. Direct use of $\hat{\Gamma}_L$ requires confidence that the trace minimization algorithm in (10) will work well in finite samples, while use of L avoids this issue because the diagonal entries in question are subtracted away and thus are irrelevant. On the other hand, using L requires labelling some eigenvalues as "zero" eigenvalues, despite the fact that due to random noise all eigenvalues will probably be non-zero.[57] A particular concern here is that the eigenvalues in question are the smallest out of N eigenvalues. Monte Carlo exercises (available upon request) suggest that the approach based on using $\hat{\Gamma}_L$ directly has better finite sample performance. We thus use this approach in our analysis, as described in the main text. Below, we briefly discuss how the alternative approach (based on the smallest eigenvalues of the graph Laplacian) might be applied.

A heuristic method is available based on "eigengaps" similar to those used by Ng, Jordan, and Weiss [2002]. Sort the eigenvalues λ of L in increasing order, such that λ_1 is the smallest and λ_N the largest.[58] The difference $\lambda_{k+1} - \lambda_k$ is defined the kth eigengap. Ng, Jordan, and Weiss [2002] argue that a large eigengap indicates that perturbation of the eigenvectors of L would not change the clusters produced by spectral clustering. Luxburg [2007] thus suggests that the right choice for $\hat{J}$ is a number such that λ_k is "small" for $k \leq \hat{J}$, and the $\hat{J}$th eigengap is large.[59] The intuition here is that if there truly are $\hat{J}$ eigenvalues that are zero, then these appear to be non-zero in the finite sample only due to random variation. In contrast, the $\hat{J} + 1$th and larger eigenvalues would be strictly positive even if the true L were used. An examination of the $\hat{J}$th eigengap thus provides a heuristic test of whether the choice of $\hat{J}$ was reliable, or whether small changes due to random variation might

[57] Eigenvalues that would be zero asymptotically will not be zero in a finite sample, because some of the entries that are zero in Γ_L will be positive in the calculated $\hat{\Gamma}_L$. When using a covariance matrix that includes this finite sample variation, it is thus necessary to account for the fact that eigenvalues that are zero in the population may not be zero in the sample.

[58] A first step to dealing with the problem of finite sample is to exclude districts with very few attacks from estimation: for the analysis of the Afghan data, we used data only for those districts in which there were 3 or more attacks (other cutoffs yielded similar results). This approach does not fully solve the underlying issue, however. For simplicity the notation here assumes that no districts are excluded on this basis and thus there are still N districts, and N eigenvalues.

[59] The underlying difficulty here is determining what exactly constitutes a "zero" eigenvalue, when there is finite sample variation. The presence of a large eigengap would thus provide some confirmation that an appropriate definition of "zero" has been chosen.

result in a different number of zero eigenvalues.

Using this approach, the estimated $\hat{J}$ corresponds to an eigenvalue such that λ_k is "small" for all $k \leq \hat{J}$. The presence of high eigengaps for very high values of k is not relevant for the eigengap procedure, so long as $J_{\max}$ is lower than these values. Luxburg [2007] suggests that the cutoff between "small" and "large" should not be larger than the minimum degree in the graph. This is trivially met by $\hat{J} = 1$, but would be violated by any much larger estimate. Although the "eigengap" approach is intended to be heuristic rather than formal, it is possible to compare the first eigengap to simulated data where there is no group structure. Compared to data where the attacks in each district have been reassigned to a random date, the first eigengap in the actual Afghanistan attack data is larger, and this difference is statistically significant at the 95% level.

More formal tests could also be constructed. Each off-diagonal $\bar{\gamma}_{ii'}$ entry will converges to $\gamma_{ii'}$ as the number of time periods grows, and the $\bar{\Gamma}_L$ matrix will converge to Γ_L. Thus, $\bar{L}$ will converge to L. Asymptotically, the correct number of the sample eigenvalues of $\bar{L}$ will approach zero. Thus, from a theoretical perspective, a test statistic similar to that given in Yao, Zheng, and Bai [2015] could be used to determine the number of zero eigenvalues. This test statistic appears to have originated from Anderson [1963], and a simplified version appears to be appropriate in this case: the eigenvalues that are converging to zero are doing so at a $\sqrt{T}$ rate, and thus for the K smallest eigenvalues, the test statistic $\sqrt{T}\sum_{k=1}^{K}\lambda_k$ or $T\sum_{k=1}^{K}\lambda_k^2$ could be used.[60]

Unfortunately, the asymptotic distribution of these test statistics is not clear, and it is also not obvious that a subsampling bootstrap approach would yield the correct distribution either. Simulations suggest that there are certain cases where the correct number of groups will only be obtained with high probability when a very large number of time periods are observed. Specifically, consider the case where α_{ij} is positive but very close to zero for some i and j. That is, there are members of group j in district i, but there are very few of them. In this case $\gamma_{ii'}$ will be very close to zero for all the other i' that contain members of group j. It is thus difficult to distinguish between i containing its own separate group, and i being a part of group j. This suggests that a formal test following this approach might be difficult to implement.

[60] The asymptotic argument is made with a fixed number of districts, N, and a growing number of time periods, T.

E Covariance Matrix with differing values of σ^2

In the main text we assume that σ is constant for all districts, and we then normalize it to $\sigma^2 = 1$. Now suppose instead that some districts are easier to coordinate than others. Continue to assume that $\text{Var}(\epsilon_j) = 1$ for all groups j, but suppose that the signal to group j in district i is $\tilde{\epsilon}_{ij} = \tilde{\sigma}_i \epsilon_j$, where $\tilde{\sigma}_i$ is a district specific indicator of how much coordination will be occurring in this district. In this case we will have

$$\Gamma_L = \left[\begin{array}{ccccc} \tilde{\sigma}_1\tilde{\sigma}_1 \sum_j \alpha_{1j}\alpha_{1j} & \tilde{\sigma}_1\tilde{\sigma}_2 \sum_j \alpha_{1j}\alpha_{2j} & & & \\ \tilde{\sigma}_2\tilde{\sigma}_1 \sum_j \alpha_{2j}\alpha_{1j} & \tilde{\sigma}_2\tilde{\sigma}_2 \sum_j \alpha_{2j}\alpha_{2j} & & & \\ ... & & \tilde{\sigma}_i\tilde{\sigma}_i \sum_j \alpha_{ij}\alpha_{ij} & & \\ \tilde{\sigma}_1\tilde{\sigma}_i \sum_j \alpha_{ij}\alpha_{1j} & & & ... & \tilde{\sigma}_i\tilde{\sigma}_{i'} \sum_j \alpha_{ij}\alpha_{i'j} \\ ... & & & & \end{array}\right] \tag{13}$$

The transformation to a correlation matrix in this case will be

$$\Gamma_L^{\text{cor}} = D(\tilde{\sigma}_\cdot^2 \sum_j \alpha_{\cdot j}\alpha_{\cdot j})^{-1/2} \Gamma_L D(\tilde{\sigma}_\cdot^2 \sum_j \alpha_{\cdot j}\alpha_{\cdot j})^{-1/2},$$

where $D()$ indicates a diagonal matrix with the specified entries on the diagonal. The resulting Γ_L does not contain any $\tilde{\sigma}$ terms, and is thus identical to the Γ_L used in the main text. We thus see that district specific differences in coordination do not affect the analysis.

Now consider the case where σ differs across groups instead of across districts. That is, $\text{Var}(\epsilon_j) = \sigma_j$. In the case where groups do not overlap there is only one group per district, and thus the situation is identical to the above where $\tilde{\sigma}$ varied by district. In the case where groups do overlap, however, the transformation to Γ_L^{cor} would no longer eliminate the σ terms. Thus, if we assume that $\sigma^2 = 1$ for all groups when this is not in fact the case, our estimator for $\{\alpha_{ij}\}$ will be inconsistent. To see what will happen here, let $\tilde{\alpha}_{ij} = \sigma_j \alpha_{ij}$. The covariance matrix will have the form

$$\Gamma_L = \sigma^2 \left[\begin{array}{ccccc} \sum_j \tilde{\alpha}_{1j}\tilde{\alpha}_{1j} & \sum_j \tilde{\alpha}_{1j}\tilde{\alpha}_{2j} & & & \\ \sum_j \tilde{\alpha}_{2j}\tilde{\alpha}_{1j} & \sum_j \tilde{\alpha}_{2j}\tilde{\alpha}_{2j} & & & \\ ... & & \sum_j \tilde{\alpha}_{ij}\tilde{\alpha}_{ij} & & \\ \sum_j \tilde{\alpha}_{ij}\tilde{\alpha}_{1j} & & & ... & \sum_j \tilde{\alpha}_{ij}\tilde{\alpha}_{i'j} \\ ... & & & & \end{array}\right] \tag{14}$$

which is exactly the same as 1, except with $\tilde{\alpha}_{ij}$ replacing α_{ij}. Thus, our estimates $\hat{\alpha}_{ij}$ will be consistent for $\tilde{\alpha}_{ij}$. This would affect the estimates shown in Figures 6 and 8. If there is a group with low σ_j that thus launches almost no simultaneous attacks, this group would show up only in very light colours in these maps. This would not necessarily present a problem, since it would still be obvious where in the country such a group was operating. The only issue that would arise is that specific districts where there was overlap with other groups would seem to be dominated by those other groups, when the reality is that those other groups are simply engaging in more coordinated attacks.

If in reality σ differs based on pairs of districts, and so is actually $\sigma_{ii'}$, then the situation becomes more difficult. In the extreme case, insurgents in each district would coordinate with those in all adjacent districts but never with those that are further away. In this case, there is no plausible clustering of districts into groups, because each district exhibits the same similarity with all of its neighbours. The idea of clustering is that the underlying structure can be simplified into cluster memberships. In the extreme case this is ineffective, and thus our model is inappropriate.

A less extreme version of this would be that there is a group structure, but insurgents in the same group are more likely to coordinate with districts that are geographically close to them rather than districts that are further away. In this case, clustering the data could return meaningful results. The clustering algorithm would have to be carefully selected, however, to not incorrectly split a group just because there was some internal variation regarding which districts were coordinating with which other districts.

For example, suppose that districts are evenly spaced along a one dimensional line, and within the same insurgent group there will only be coordination between districts that are within a distance d of each other. In this case the covariance matrix does not consist of blocks as in Equation 2. Instead replacing each block will be a band, where the entries outside of the band are 0 because these district pairs, while in the same group, are too far away to coordinate. We thus have what we might call a diagonal matrix instead of a block diagonal matrix.

This situation would not be handled correctly by the approach we use in this paper, because we would incorrectly split a group based on the fact that it has this internal structure. It seems as though some sort of improved method should be able to cluster districts correctly here, because there is no correlation between districts in different groups but some positive correlation between at least some districts in the same group, and there are enough of these positive correlations to connect the entire group. One method that could potentially resolve this problem would be to use variant of correlation clustering [Bansal, Blum & Shuchi 2004]. We leave this for future work.

F Clustering Details and Estimate for α

As a first step, the correlation matrix $\hat{\Gamma}_L^{\text{cor}}$ is readily obtained by imposing diagonal elements equal to 1 and appropriately rescaling rows and columns of the covariance matrix $\hat{\Gamma}_L$ by the square root of the corresponding diagonal entry of $\hat{\Gamma}_L$.

For many k-means algorithms, however, a distance matrix rather than a correlation matrix is needed. Such a distance matrix can easily be constructed using cosine distances: $1-\gamma_{ii'}^{\text{cor}}$ is the cosine distance between i and i', where $\gamma_{ii'}^{\text{cor}}$ is the off-diagonal entry of Γ_L^{cor} corresponding to districts i and i'.[61] The cosine distance between two districts with the same group present will be zero asymptotically, while it will be one when the districts have different groups present.

For the particular data that we will be considering, a weighted clustering approach appears to be called for because a district with very low α_{ij} for the group j that is present will have very noisy off-diagonal entries. We do not explore optimal weights, instead using ad-hoc weights corresponding to the square root of the diagonal entries of $\hat{\Gamma}_L$. Krishna and Narasimha [1999] provide a weighted k-means algorithm, based on genetic optimization: we use the Hornik, Feinerer, Kober, and Buchta [2012] implementation of this algorithm. Using unweighted clustering instead does not change any of the results discussed below substantially.

Suppose that each organized group that is present has members in a large number of districts, and that no single district has a particularly large α_{ij}. Let I_j be the set of districts that have members of organized group j. Then, since an assumption of the model was that the organized groups do not overlap, an estimate of α_{ij} for $i \in I_j$ can be produced via the following approximation, using $\bar{\Gamma}^j$, the relevant block of the original $\bar{\Gamma}$.[62]

Specifically, note that a sum across the off-diagonal entries of a row of $\bar{\Gamma}$ corresponding to district i is $\sum_{i' \neq i} \alpha_{ij}\alpha_{i'j}$. If there are a large number of districts with

[61] The construction of a distance matrix is trivial because any correlation matrix is also an interpoint angle matrix, and these angles can be used directly to construct a cosine distance matrix.

[62] A potential alternative approach to the one presented here would be to use the diagonal entries of $\hat{\Gamma}_L$ to produce estimates of $\{\alpha_{ij}\}$. However, this matrix is itself the output of a semi-definite program based on $\bar{\Gamma}$. The approach presented below has the advantage of using the off-diagonal entries of $\bar{\Gamma}$ directly.

members of j, then it is reasonable to use the approximation

$$\begin{aligned}\sum_{i' \neq i} \alpha_{ij}\alpha_{i'j} &\simeq \sum_{i'} \alpha_{ij}\alpha_{i'j} \\ &= \alpha_{ij} \sum_{i'} \alpha_{i'j} \\ &= \alpha_{ij} a_j\end{aligned} \tag{15}$$

where $a_j = \sum_{i'} \alpha_{i'j}$ is the same for any choice of district i within I_j. The row sums of the off-diagonal entries of each block of $\bar{\Gamma}^j$ thus give the relative prevalence of organized group members in each district in I_j.[63]

[63] While it would be possible to use non-linear programming or other techniques to develop an estimator with more desirable properties, the approximate estimator has at least two advantages. First, the estimator has an intuitive interpretation: $\bar{\Gamma}$ is a covariance matrix, and the sum across the off-diagonal entries of a row of $\bar{\Gamma}$ thus gives an indication (in a heuristic sense) of how closely linked attacks in a given district are with attacks in other districts. Second, if in the data a given district i experiences only a small number of attacks, then the off-diagonal entries $\bar{\gamma}_{ii'}$ will be relatively small for that district, and thus i will not introduce substantial noise into estimates $\hat{\alpha}_{i'j}$ for other districts i'. Developing an unbiased estimator that also possesses such properties appears to be a non-trivial undertaking.

G Eigenratio type estimators: Simulations

To better understand the finite sample properties of eigenratio type estimators, we conduct a series of simulations. For simplicity, we do not use a model with discrete attacks, as presented in Section 2, but instead use a more standard model with normally distributed random variables. Let there be $J = 4$ groups, $N = 100$ districts, and $T = 2000$ days. Let there be exactly one group in each district, with $\alpha_{i1} \sim \text{Uniform}(0,1)$ i.i.d. for $i \in \{1, ..., 25\}$, and no other group present in those districts. In the same fashion, only Group 2 is present in districts 26-50, only Group 3 in districts 51-75, and only Group 4 in 76-100.

Our simplified model of attacks is that in each period t for each group j, an i.i.d. draw $\epsilon_{tj} \sim N(0, \sigma^2)$ is made. The number of attacks is then given by

$$x_{it} = \sum_j \alpha_{ij}\epsilon_{tj} + u_{it} \tag{16}$$

where $u_{it} \sim N(0,1)$, i.i.d.

We then consider eigenvalues associated with the (N by N) covariance matrix of attacks. We perform 100000 simulations for each of $\sigma^2 = 1$, $\sigma^2 = 0.1$, $\sigma^2 = 0.05$, and $\sigma^2 = 0$, generating a total of 400000 simulated sample covariance matrices.[64]

Figures G.1 - G.3 graphically display the results of these simulations. Figure G.1 shows the eigenvalues of the covariance matrix. We see that the group structure is immediately apparent at $\sigma^2 = 1$, still clear at $\sigma^2 = 0.1$, but somewhat unclear at $\sigma^2 = 0.05$. There is no group structure with $\sigma^2 = 0$, and thus Figure G.1d shows the distribution of eigenvalues under $J = 0$.

Figure G.2 shows eigenratios, with the leftmost eigenratio being the ratio between the largest (i.e. leftmost) and second-largest eigenvalues, and so forth. Here, on average the largest eigenratio clearly corresponds to $J = 4$ when σ^2 is large, but this is no longer the case with $\sigma^2 = 0.05$. Figure G.2d shows that the distribution of eigenvalues when $J = 0$ leads to a somewhat peculiar distribution of eigenratios: the first few and last few eigenratios are much larger than the others. Figure G.2d thus illustrates why it is important to have some maximum number number of possible

[64] Note that in the main text, the choice of $\sigma^2 = 1$ is a normalization, because the $\{\alpha_{ij}\}$ are unknown, and a decrease in the choice of σ^2 would simply result in higher $\hat{\alpha}$ estimates. In contrast, in the simulations in this appendix, the distribution of the $\{\alpha_{ij}\}$ are given, and thus choosing a different value σ^2 changes the signal to noise ratio for the attack covariance matrix.

groups, $J_{\max}$. The eigenratios associated with the very smallest eigenvalues (towards the right hand side of Figure G.1d) become quite large. With $N = 100$, and no $J_{\max}$, choosing $\hat{J}$ based on the largest of all the eigenratios would lead to many $\hat{J}$ estimates of 99 groups. However, as noted in Ahn and Horenstein [2013], any intermediate choice of $J_{\max}$ is unlikely to affect the results.

Figure G.3 shows the distribution of estimates $\hat{J}$ with $J_{\max} = 50$. Figures G.3a and G.3b show that the eigenratio approach works very well when the signal to noise ratio in the covariance matrix is relatively high. Figure G.3c, however, shows that with a noisier covariance matrix, the estimated values for $\hat{J}$ tend to be too low. Figure G.3d shows the distribution of estimates of $\hat{J}$ when there is no group structure.

In both of Figures G.3c and G.3d, $\hat{J} = 1$ is the modal estimate. Figure G.3d shows that the median estimated $\hat{J}$ is below the true value $J = 4$ (the mean is above, but this is less apparent from the figure). However, Figure G.3d shows the case with no group structure at all, and thus would not change regardless of the true value of J. The bias of the estimator thus cannot be signed: this is a natural result of J and $\hat{J}$ both being integers bounded between 0 and 50. Bias correction appears to be non-trivial.

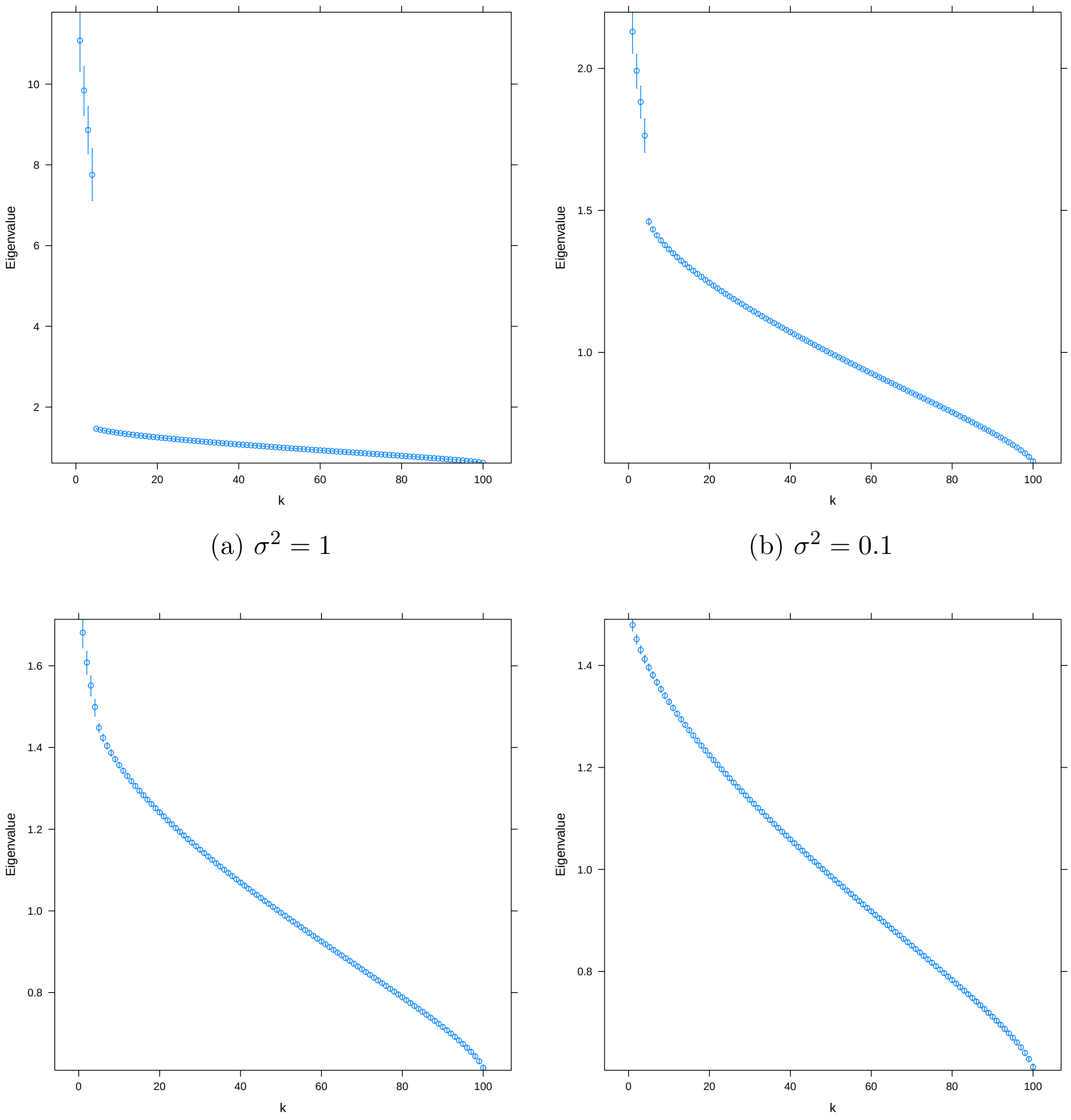
Figure G.3c provides a possible explanation for why estimates of $\hat{J} = 1$ appear so frequently in Table 5. The finite sample properties of eigenratio type estimators are such that there is a tendency to estimate low values of $\hat{J}$ in cases where the covariance matrix is noisy. This is due to the distribution of eigenvalues resulting from the noise, as shown in Figure G.1d. The evidence provided in Table 5 should thus mainly be taken as an indication that the null hypothesis of no group structure should be rejected. Figure G.3c shows how estimates $\hat{J} = 1$ occur frequently when there is actually a group structure with $J > 1$.[65]

G.1 Comparison with Hierarchical Splits

To compare our eigenratio type estimator with the estimator based on hierarchical splits, we need to simulate data with discrete attacks, as the permutation test used require integer numbers of attacks to permute. Let the number of attacks by group j in district i at time t be drawn from a Poisson(λ_{ijt}) distribution, where $\lambda_{ijt} = 0$ with probability 0.9, and $\lambda_{ijt} = \alpha_{ij}$ with probability 0.1 (this is equivalent to ϵ_{it} having

[65] In the empirical literature, "low" estimates for the number of factors (compared to other methods) are obtained by Choi et al. [2014] and Wu et al. [2011]. Figures G.3b and G.3c appear in line with results reported (using actual data) in the supplement to Baurle [2013].

Appendix Figure G.1: Eigenvalues



(a) $\sigma^2 = 1$

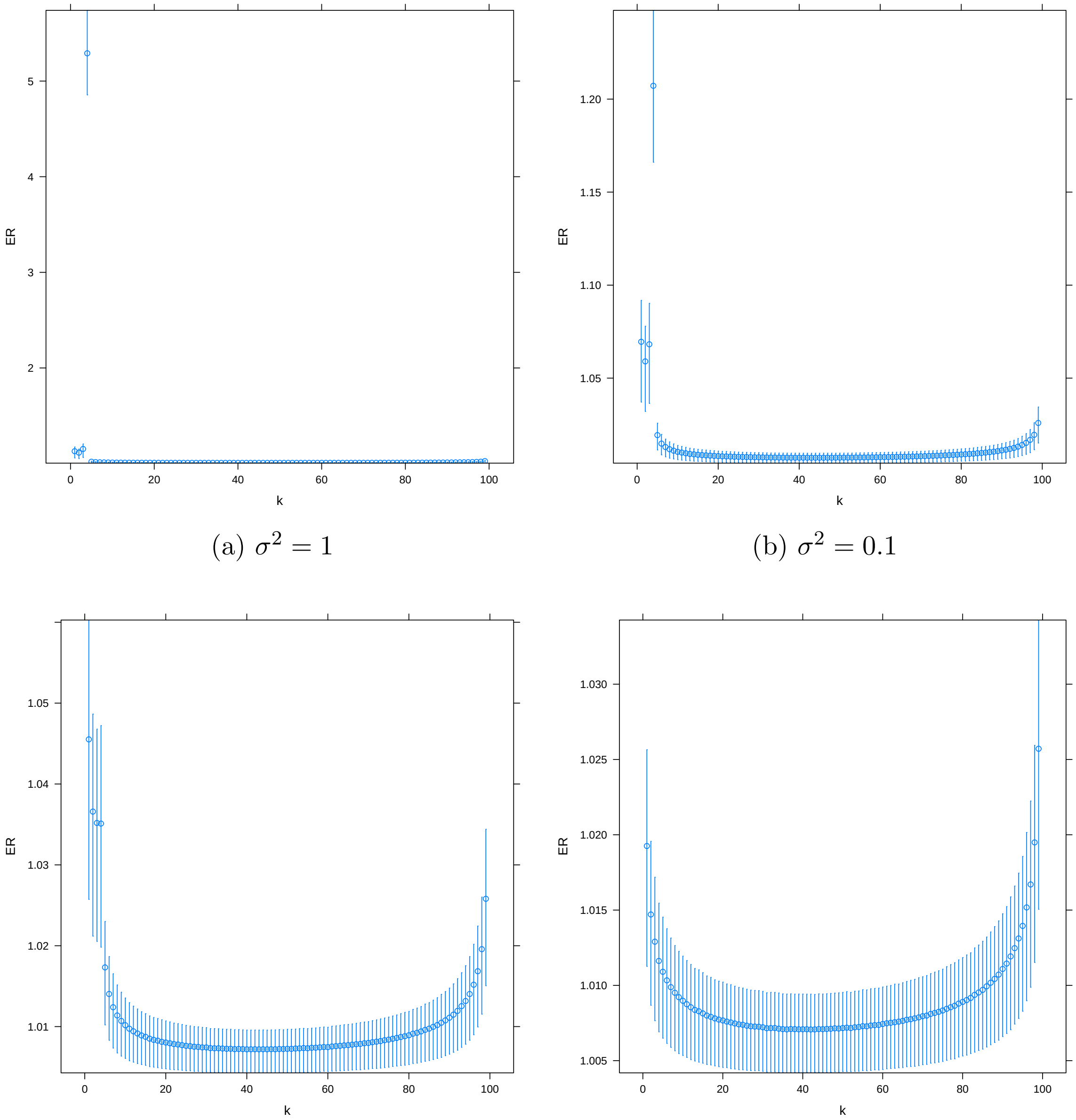
(b) $\sigma^2 = 0.1$

(c) $\sigma^2 = 0.05$

(d) $\sigma^2 = 0$

Points indicate means over 100000 simulations. Bars show interquartile range.

Appendix Figure G.2: Eigenratios



(a) $\sigma^2 = 1$

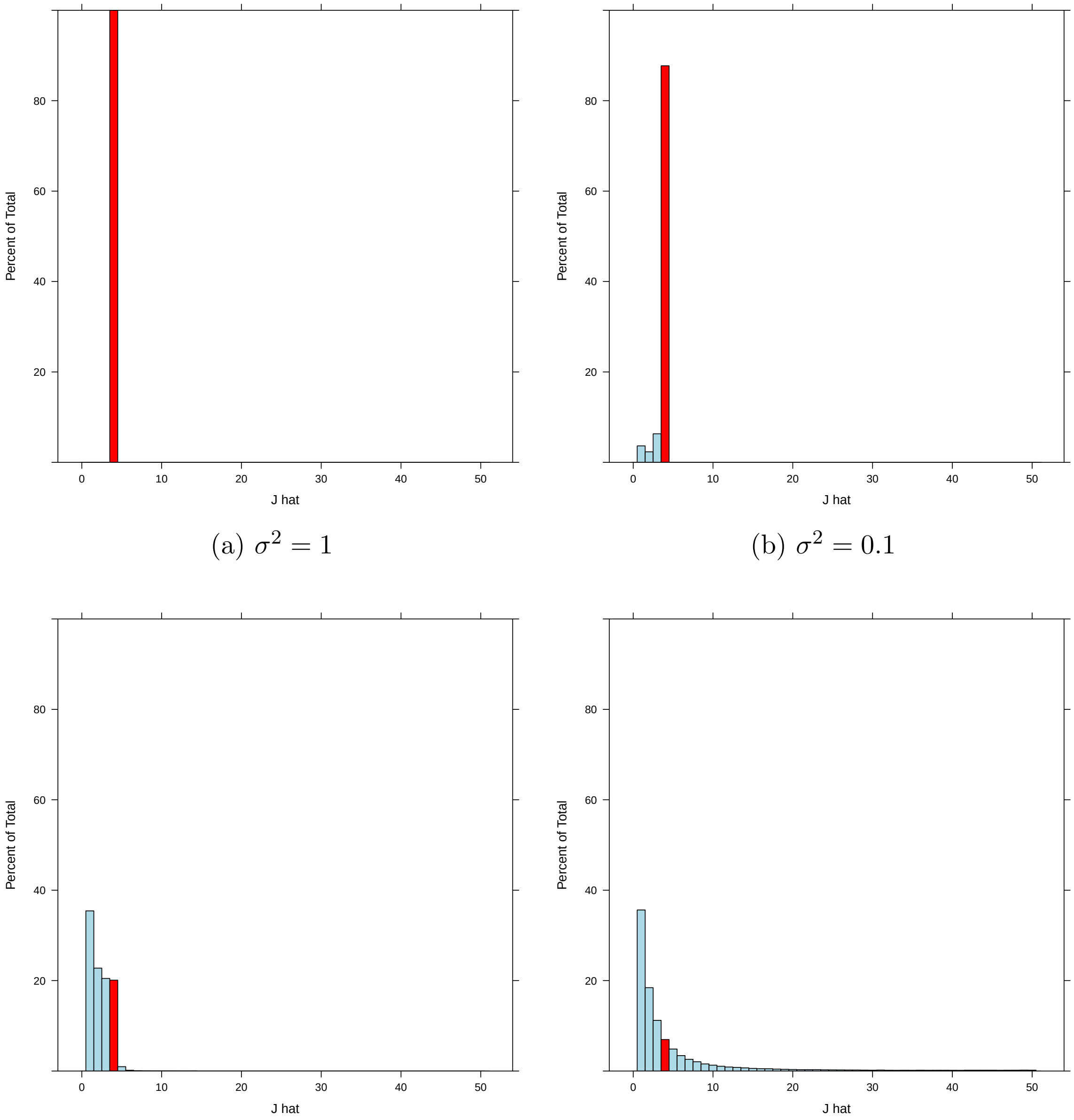
(b) $\sigma^2 = 0.1$

(c) $\sigma^2 = 0.05$

(d) $\sigma^2 = 0$

Points indicate means over 100000 simulations. Bars show interquartile range.

Appendix Figure G.3: Estimated number of groups ($\hat{J}$)



(a) $\sigma^2 = 1$

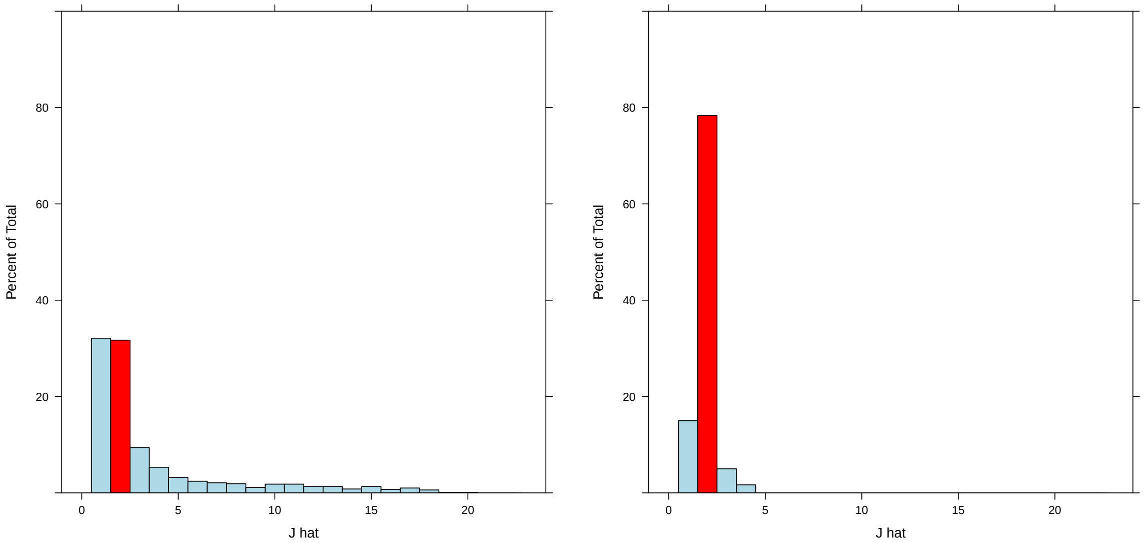
(b) $\sigma^2 = 0.1$

(c) $\sigma^2 = 0.05$

(d) $\sigma^2 = 0$

Histograms of estimated number of groups, over 100000 simulations. True value $J = 4$ shown in red.

Appendix Figure G.4: Estimated number of groups ($\hat{J}$)



(a) Eigenratio

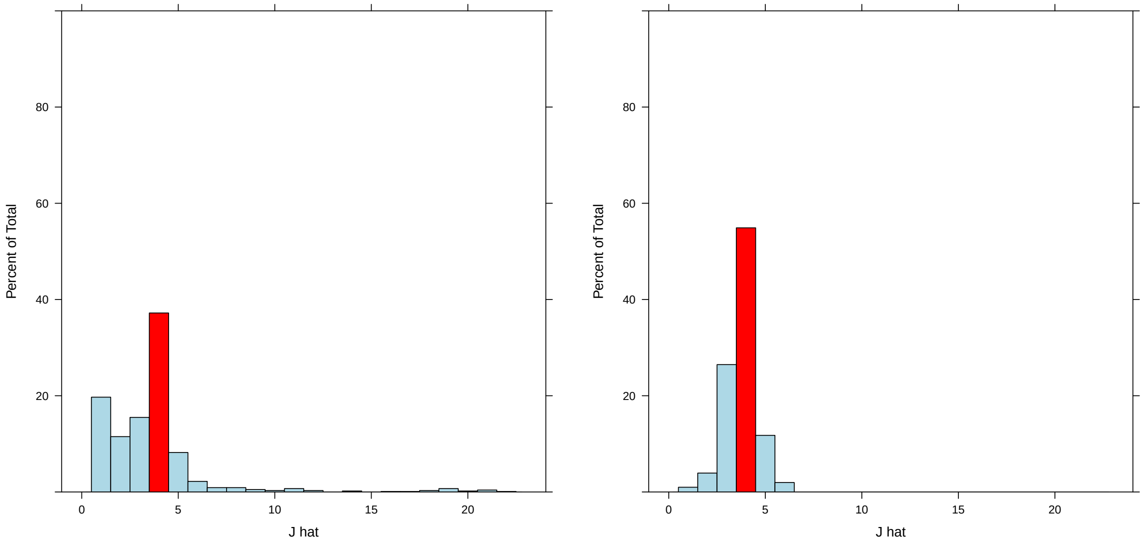
(b) Hierarchical Splits

Histograms of estimated number of groups, over 100 simulations. True value $J = 2$ shown in red.

a bernoulli distribution with probabilities 0.9 and 0.1). We consider $J = 2$ with the non-zero α_{ij} entries drawn from a Uniform$(0, 0.25)$ distribution, as well as $J = 4$ with the non-zero α_{ij} entries drawn from a Uniform$(0, 0.5)$ distribution.

Results are shown in Figures G.4 and G.5. In both cases, the method based on hierarchical splits substantially outperforms that based on eigenratios. A particular advantage of the hierarchical splits is that there are no estimates with very large values of $\hat{J}$, whereas with the eigenratio type approach a small number of simulations yield extremely large values for $\hat{J}$. The hierarchical split based method is also less likely to stop at $\hat{J} = 1$, and thus estimates in both tails appear to be less likely with this method.

Appendix Figure G.5: Estimated number of groups ($\hat{J}$)



(a) Eigenratio

(b) Hierarchical Splits

Histograms of estimated number of groups, over 100 simulations. True value $J = 4$ shown in red.

H NNMF Consistency

Conditions under which $\hat{\Gamma}_L$ will converge to Γ_L have been discussed in Appendix C.1. We now consider conditions under which a non-negative matrix factorization of Γ_L will recover the $\{\alpha_{ij}\}$ group structure. It is clear that the index numbering of the groups cannot be recovered, because Γ_L is invariant to relabelling of groups. The index numbering of groups is irrelevant throughout our analysis, however, and thus we are only concerned with whether the group structure can be recovered up to a reindexing.

Huang, Sidiropoulos, and Swami [2014] discuss uniqueness of symmetric non-negative factorizations at some length. They conclude that while there are no obvious necessary conditions to check for uniqueness, simulations reveal that multiplicity of solutions does not appear to be a problem unless the correct factorization is extremely dense: factorizations with 80% non-zero entries are still reconstructed successfully. The Γ_L matrices considered in this paper would generally be expected to have a relatively sparse factorization.

I Reference Distributions

We consider three different "reference distributions". First, suppose that the structural model presented in Section 2 is correct. In this case, the distribution of the number of attacks by disorganized militants in district i is the same for all periods, with expected value $\eta \ell_i$. Thus, under the null hypothesis that there is no group structure, the observed attack data is weakly exchangeable: within a given district, permuting the time indices does not change the joint distribution of the attacks.[66] The total number of such permutations is huge, and thus rather than perform calculations using the entire set we consider only a random subset of these permutations. By construction, the permuted data exhibits no group structure: all the off-diagonal entries of the sample covariance matrix will be zero asymptotically. To construct the desired reference distribution, we treat each of these permutations as if it were the observed data.

Now, suppose instead that the structural model assumed is not exactly correct, and there is some cross-time variation in the expected number of attacks by disorganized militants within a district. Specifically, suppose that the probability that a disorganized militant launches an attack is not a constant η, but rather varies across months. The expected number of attacks on a given day in month m is then $\eta_{im} \ell_i$, and will differ by month. In this case, the observed attack data is still weakly exchangeable, but only within a given district *and* a given month. We can thus still construct a reference distribution, provided that observations are permuted only within each month for each district. In this case, the covariance matrices may not have all off-diagonal entries zero asymptotically: it could be that η_{im} and $\eta_{i'm}$ are positively correlated, for example.

Finally, suppose that the expected number of attacks by disorganized militants varies at the daily level, rather than the monthly level. The general case, with $\eta_{it} \ell_i$ attacks expected in district i at time t, is so general that it does not appear to allow for any permutations. However, suppose that the number of expected attacks is instead $\eta_t \ell_i$, where η_t now does not differ across districts.[67] This might be the case,

[66] The intuition here can be provided by an example. Suppose there are three periods. If there is no group structure, then the probability of observing $\{x_1, x_2, x_3\}$ in a given district must be equal to the probability of observing $\{x_1, x_3, x_2\}$, because the number of attacks is i.i.d. across time within a given district.

[67] This gives the disorganized militants the same structure an additional organized group. The test against the null hypothesis in this case is thus related to whether there is an organized group present

for example, if there were particular days that, for whatever reason, generated large amounts of random violence. In this case, observations are "approximately" weakly exchangeable via the following sort of permutation, inspired by Good [2002]. Find a pair of districts i and i', and a pair of times t and t', such that the following two conditions hold: there were the same number of attacks x in district i at time t and in district i' at time t', and there were the same number of attacks x' in district i at time t' and in district i' at time t. Permute the data by swapping x and x' in these four entries.[68] These permutations are attractive from an intuitive perspective, as they retain not only the same number of total attacks in each district, but also the same number of total attacks on each day. In the Afghan data, there are relatively few attacks on any given day and thus an enormous number of possible permutations of this sort.

I.1 Additional reference distribution

The purpose of generating permutations is to compute distributions of test statistics, and one of the most obvious test statistics is the fraction of covariance explained by the group structure. Covariance matrices are positive semi-definite, and thus have a spatial interpretation as points in Euclidean space that can be used in order to consider the "between sum of squares" and "within sum of squares" produced by any given clustering. With the permutations just proposed, however, the contribution of different districts to the total sum of squares will generally be different between different permutations, and thus some permutations may be more amenable to clustering than others. In addition, the permutations may in general be more amenable to clustering than the actually observed data, which complicates the interpretation of

that is active in some districts but not others. Under the null hypothesis, the off-diagonal entries of the sample covariance matrix should be directly proportional to the total number of attacks in the districts in question.

[68] To see why this weak exchangeability holds "approximately", note that the distribution of attacks is binomial. Approximate the binomial with a Poisson distribution with expectation $\eta_t \ell_i$. Then for observations of the type just described

$$\begin{aligned}\Pr(x|\eta_t\ell_i)\Pr(x'|\eta_{t'}\ell_i)\Pr(x'|\eta_t\ell_{i'})\Pr(x|\eta_{t'}\ell_{i'}) &= \frac{(\eta_t\ell_i)^x}{x!}e^{-\eta_t\ell_i}\frac{(\eta_{t'}\ell_i)^{x'}}{x'!}e^{-\eta_{t'}\ell_i}\frac{(\eta_t\ell_{i'})^{x'}}{x'!}e^{-\eta_t\ell_{i'}}\frac{(\eta_{t'}\ell_{i'})^x}{x!}e^{-\eta_{t'}\ell_{i'}} \\ &= \Pr(x'|\eta_t\ell_i)\Pr(x|\eta_{t'}\ell_i)\Pr(x|\eta_t\ell_{i'})\Pr(x'|\eta_{t'}\ell_{i'})\end{aligned}$$

by rearranging terms. The canonical reference for multivariate permutations appears to be Pesarin [2001], although this specific type of permutation is not described. Good [2005] provides an accessible introduction to permutation tests.

the permutation test. A way to avoid this would be to use only those permutations where each district makes the same contribution to the total sum of squares as in the actually observed data. While this would be an improvement, the correlation matrix Γ^{cor} is what is block diagonal, and thus is the most appropriate object to analyze using a sum of squares decomposition. To keep the contribution of each district to the total sum of squares the same when considering this correlation matrix, we can add an additional requirement that the diagonal entries of the covariance matrix remain the same as those in the actually observed data. This ensures that the transformation to the correlation matrix will involve division by the same quantities as in the actual data, and thus the contribution of each district to the total sum of squares in the correlation matrix will remain the same in the permutation as in the actual data. The permutations that satisfy these additional criteria are a subset of the "swap" permutations discussed above; however, there does not appear to be a way to generate a permutation of the desired type by randomly choosing swaps. It is possible, however, to create valid permutations through the use of an integer program. Let the variables for this program be binary variables x_{ti}^{r}, which is equal to one if there were r attacks on day t in district i, and equal to zero otherwise. A valid permutation will satisfy the constraints

$$\sum_{r} x_{ti}^{r} = 1 \qquad \forall t, i \tag{17}$$

$$\sum_{t=1}^{T} x_{ti}^{r} = \sum_{t=1}^{T} x_{ti}^{r,\text{actual}}, \qquad \forall i, r \tag{18}$$

$$\sum_{i=1}^{N} \sum_{r} r x_{ti}^{r} = \sum_{i=1}^{N} \sum_{r} r x_{ti}^{r,\text{actual}} \qquad \forall t \tag{19}$$

$$\sum_{t=1}^{T} \left(\sum_{r} r x_{ti}^{r}\right)\left(\sum_{r} \sum_{i=1}^{N} r x_{ti}^{r,\text{actual}}\right) = \sum_{t=1}^{T} \left(\sum_{r} r x_{ti}^{r,\text{actual}}\right)\left(\sum_{r} \sum_{i=1}^{N} r x_{ti}^{r,\text{actual}}\right) \qquad \forall i \tag{20}$$

where $x_{ti}^{r,\text{actual}}$ is a constant corresponding to the actually observed data. The first constraint simply ensures that there is a number of attacks on each day in each district. The second constraint ensures that distribution of attacks within each district is the same as in the actually observed data: this also ensures that the diagonal entries of the covariance matrix are the same as in the actually observed data. The third constraint ensures that the number of attacks on each day is the same as in the actually observed data.[69] The fourth constraint ensures that the sum of each row (and column) of the covariance matrix is the same as in the actually observed data.

A solution to this binary integer program always exists, because the actually observed data will always satisfy the constraints. To randomly generate a solution to the program, we choose a random objective function, and stop at the first integer solution obtained. Running the program repeatedly generates a random sample of permutations with the desired characteristics.

Table I.1 performs the same analysis as 1, except using the above reference distribution instead of using auxiliary geographic information.

[69] This is slightly weaker than the "swap" permutations described above, which preserve the distribution of attacks within each day. There does not appear to be a need for this stronger constraint, and so we relax it here.

Appendix Table I.1: Hierarchical model without geographic information

		Afghanistan	Pakistan
		I	II
Split at (1)?	Randomly shuffled data (mean)	234.06	101.68
	Std. dev.	0.15	0.11
	Actual data	234.02	101.01
	p-value	0.40	0.00
Split at (2)?	Randomly shuffled data (mean)		47.02
	Std. dev.		0.09
	Actual data		46.78
	p-value		0.01
Split at (3)?	Randomly shuffled data (mean)		49.32
	Std. dev.		0.08
	Actual data		49.17
	p-value		0.04
Split at (4)?	Randomly shuffled data (mean)		17.01
	Std. dev.		0.03
	Actual data		17.01
	p-value		0.48
Split at (5)?	Randomly shuffled data (mean)		24.92
	Std. dev.		0.06
	Actual data		24.82
	p-value		0.08
Split at (6)?	Randomly shuffled data (mean)		20.08
	Std. dev.		0.06
	Actual data		19.98
	p-value		0.07
Split at (7)?	Randomly shuffled data (mean)		21.52
	Std. dev.		0.06
	Actual data		21.45
	p-value		0.14

J Estimation using monthly covariance matrices

Suppose that attack probabilities are relatively small. Then the number of attacks by unorganized militants can be approximated using a Poisson($\zeta_{im}\eta\ell_i$) distribution instead of using the actual Binomial($\zeta_{im}\eta, \ell_i$) distribution. Similarly, the distribution of attacks by members of an organized group can be approximated with Poisson($\zeta_{im}\epsilon_{tj}\alpha_{ij}$) in place of Binomial($\zeta_{im}\epsilon_{tj}, \alpha_{ij}$).

Now, suppose that there are a total of x_{im} attacks in district i. Conditional on there being a total of x_{im} attacks, the distribution of these attacks across days is given by a Multinomial(x_{im}, p_i) distribution, where p_i is a probability vector with elements of the form

$$p_{it} = \frac{\eta\ell_i + \sum_j \epsilon_{tj}\alpha_{ij}}{\sum_{t'}(\eta\ell_i + \sum_j \epsilon_{t'j}\alpha_{ij})}$$

If in some other district i' there were $x_{i'm}$ attacks, then the covariance of daily attacks has the useful form

$$\begin{aligned}\text{Cov}(x_{im\cdot}, x_{i'm\cdot}) &= x_{im}x_{i'm}\sum_t p_{it}p_{i't} - \frac{x_{im}}{T}\cdot\frac{x_{i'm}}{T} \\ &= x_{im}x_{i'm}(\sum_t p_{it}p_{i't} - \frac{1}{T}\cdot\frac{1}{T}) \\ \frac{\text{Cov}(x_{im\cdot}, x_{i'm\cdot})}{x_{im}x_{i'm}} &= \text{SCov}(p_{it}, p_{i't})\end{aligned}$$

where SCov($p_{it}, p_{i't}$) gives the sample covariance for a given draw of ϵ. The first line of the above holds because each attack decision is independent given both the total number of attacks and the realization of ϵ. If the ϵ are constructed such that $\sum_{t'}\epsilon_{t'j} = 1$, then the denominator in the expression above for p_{it} will simplify such that

$$\text{SCov}(p_{it}, p_{i't}) = \frac{\sum_j \alpha_{ij}\alpha_{i'j}\sigma_j^2}{(T\eta\ell_i + \sum_j \alpha_{ij})(T\eta\ell_{i'} + \sum_j \alpha_{i'j})}$$

The $T\eta\ell_i + \sum_j \alpha_{ij}$ term can be taken to be the "average" number of attacks, which implies that $\tilde{\alpha}_{ij} = \frac{\alpha_{ij}}{T\eta\ell_i + \sum_j \alpha_{ij}}$ is the fraction of attacks in district i that group j will be responsible for. Then

$$\mathrm{Cov}(p_{it}, p_{i't}) = \sum_j \tilde{\alpha}_{ij} \tilde{\alpha}_{i'j} \sigma_j^2$$

Here $\tilde{\alpha}$ and σ^2 are not separately identified. If the normalization $\sigma_j^2 = 1$ is used, then the estimated $\tilde{\alpha}$ describe relative degrees to which groups are more or less responsible for attacks, across districts.

K Coding of attack vs. defence

A possible concern with the attack data we use is that, while classified as insurgent attacks, these incidents are actually in response to government actions. Thus, any correlation we discover between districts would not be indicative of the structure of insurgent groups, but rather the organization of the counter-insurgency.

There are two situations that are of particular concern. First, there is the danger that a police attack on an insurgent stronghold might be included in our data as an attack simply because the insurgents shoot back. Second, even if our data only includes incidents initiated by the insurgents in a tactical sense, these incidents may be initiated by the government in a strategic sense. For example, suppose that a mountainous area is known to be insurgent controlled, and the government wants to change this. It might send several patrols deep into the mountains. Insurgents that happen to be present in the area might then attack these patrols as targets of opportunity. These attacks could then show up in our dataset as simultaneous attacks, but this would be evidence of coordination by the government, rather than by the insurgents.

The easiest dataset to use to consider these issues is the Global Terrorism Database (GTD), which has much more detailed coding of events than either WITS or BFRS. The GTD has a smaller number of incidents overall, which is why we do not use it as our main datasource, but as shown in Section , this dataset gives effectively the same results, albeit with a smaller number of districts. Thus, if we can show that the above problems do not occur with the GTD, this suggests that they are not responsible for the results we report in the paper.

The GTD only includes incidents where non-state actors are the attackers. Thus, it specifically excludes incidents such as police raids. This can be seen in the dataset because a small number of incidents (about 0.1%) are coded as doubtful because the attack could been by a state actor. In a few of these, the additional notes explicitly give as the reason that the police may have fired first. The other 99.9% of attacks are not believed to be initiated by government forces, and thus simultaneous government attacks are not contaminating the data.

The second possibility is that a strategic decision by the government leads naturally to simultaneous attacks by the insurgents without any insurgent planning can also be checked using notes that accompany the GTD entries. Attacks on government forces could occur when these forces are on patrol, or they could occur when the government forces are stationary. If the forces are on patrol, it could be that they have entered an insurgent held area, and it is obvious that if many patrols simultaneously enter then they will be simultaneously attacked. On the other hand, if the forces are stationary, then there is no particular reason for the insurgents to naturally attack these forces simultaneously, unless there is coordination on the part of the insurgents. A police checkpoint, for example, could be attacked today, but could equally well be attacked tomorrow, and thus, beyond random chance, the simultaneous attacks that do occur would be due to insurgent coordination.

The question thus becomes whether insurgents strike mainly when government forces are on patrol, or when they appear to be stationary. In the GTD data, there are a total of 124 sets of simultaneous attacks listed for Afghanistan. In the summary description of these attacks, “patrol” occurs in descriptions in 4 sets of attacks, “checkpoint” appears in descriptions in 25 sets of attacks, and “post” or “checkpost” appears in descriptions in 31 sets of attacks. A qualitative examination of the descriptions suggests that many of the remaining attacks are aimed at targets that would best be described as “stationary” (e.g. police chiefs, embassies, towns). It thus appears that insurgents mainly attack government forces when they are stationary. This strongly suggests that government strategic decisions do not determine the precise day when the insurgents will attack, and thus the observed simultaneity really is due to insurgent coordination, rather than being a mechanical product of the strategy of government forces.

L Sindhudesh Liberation Army

The first recorded attack in the GTD under the SDLA banner is recorded on November 2nd, 2010 when incident (ID number 201011020003) states:

> *"11/02/2010 : On Tuesday, in Hyderabad, Sindh, Pakistan, a portion of rail track was damaged when unidentified militants detonated an improvised explosive device, wounding four people. Another bomb was found and defused be security forces at the scene. A two-page pamphlet issued by Sindhu Desh Liberation Army (SDLA) 'chief commander' Darya Khan was found on the spot. The pamphlet enlisted 19 points, mentioning issues of Sindh and targeting what it called Punjabi imperialism."* [70]

When reading the entries for the GTD simultaneous incidents of Karachi (201102110009) and Hyderabad (201102110005) on February 2, 2011, which are explicitly listed as *not being part of multiple incidents*, the GTD appears uninformed by these events. Consider (201102110005) notwithstanding explicit claiming by the SDLA (possibly discarded as not credible):

> *"02/11/2011: On Friday morning, in Bengali Colony of Hussainabad in Hyderabad, Sindh, Pakistan, unidentified militants blew up railway tracks, causing no casualties but damaging the tracks. A few pamphlets were found at the blast sites carrying the name of an unknown group, Sindhu Desh Liberation Army (SDLA)."*

Subsequently, the GTD identifies two attacks in Sindh in the month of November 2011. None of these attacks is again classified as part of a multiple incidents event (i.e. coordinated attacks). It is only on February 25th, 2012, about a year after our methodology singles out SDLA activity in Sindh, that coordination of SDLA is finally

[70] And the next day GTD records incident 201011030021 *"11/03/2010: On Wednesday, near Nawabshah, Sindh, Pakistan, unknown assailants detonated an improvised explosive device on the Karachi-Lahore railroad. The blast damaged an eight-inch long portion of up-track and caused rail traffic to be suspended for over three hours. A bomb disposal squad later discovered a second bomb and successfully defused it. No casualties were reported. Sindhu Desh Liberation Army (SDLA), has claimed the responsibility for the blasts. The organization's purported chief commander, Darya Khan, has threatened that it would continue to carry out such attacks in future in order to get their "right to liberation" recognized by the United Nations."*

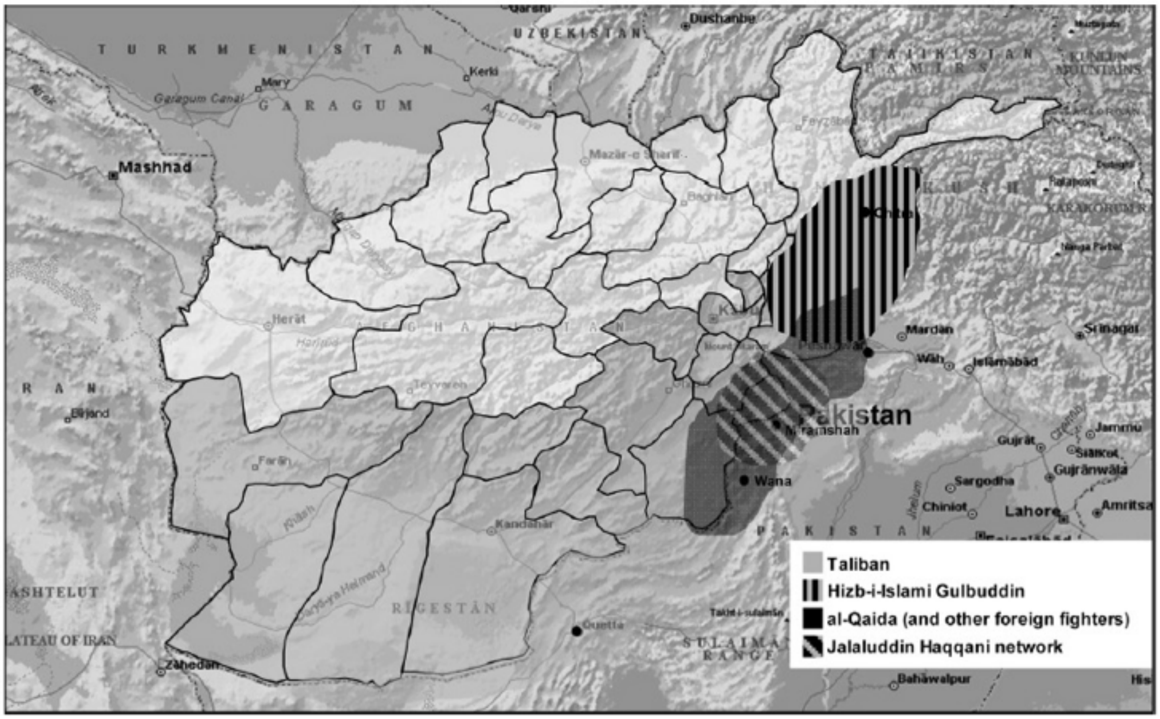
detected in the GTD, with multiple entries (12 entries, listed explicitly as being part of multiple incident).[71]

The BFRS data mentions in its comment section the SDLA only on 4 of the 41 attacks taking place on the month of November 2011 in Sindh. Of these 41 attacks, we can observe that only 10 are isolated incidents, while 31 attacks occur in bundles of 2 in a day (5 multiple incidents) or 3 attacks in a day (7 multiple incidents). By this time, our methodology is picking up SDLA coordinated activity since early 2011, information clearly missed both in the GTD and in the BFRS.

[71] Incident 201202250003 states *“02/25/2012: Explosives planted along railway tracks detonated in Jamshoro district, Sindh province, Pakistan. The tracks were damaged, but there were no human casualties. This was one of 12 explosive devices planted on railroad tracks in Sindh province on February 25, 2012. Sindhu Desh Liberation Army (SDLA) claimed responsibility, stating that people were fighting nationally and internationally for Baloch independence.”’* In addition to the one above, incident GTD ID’s are all those listed 201202250012-201202250022. On May 2nd 2012 the SDLA followed suit with 21 coordinated bomb attacks on the same day on banks around Sindh province. In the month of May 2012, SLDA activity caused 9 deceased and 30 wounded victims.

M Additional Figures & Tables for Section 5

Figure 1. The Afghan Insurgent Front



Appendix Figure M.6: from Jones [2008]

N Additional Analysis: Global Terrorism Database

Check 1. Our results indicate that the group structure we estimate for Pakistan corresponds to ethnic homelands. We might thus be concerned that in fact our method is not picking up individual insurgent groups, but rather some broader aspect of coordination within the same ethnic group. The GTD includes some information on the identities of attackers, and we can use this to cross-check our estimated group structure.[72] We examine this data for simultaneous attacks in Pakistan during the period that we study. In Balochistan, the GTD lists 38 attacks. Of these, 32% are ascribed to the Baloch Republican Army, and the remainder are listed as unknown. In the Federally Administered Tribal Areas and North-West Frontier Province, there were 167 attacks. Of these, 31% were ascribed to the (Pakistani) Taliban, 4% to Lashkar-e-Islam (which later joined the Taliban), and the remainder were unknown. Thus, in these cases, our results match what evidence is available: our method finds one group in Balochistan and one more in the area near the Afghan border. The GTD records very few attacks in Punjab, and most of these are Taliban attacks in the part of Punjab nearest to the Afghan border. A comparison for Punjab is thus unfortunately not available.

In Sindh, the GTD reports 24 attacks, but 20 of these involve an unknown group. In the next year, however, there are 54 attacks reported, with 61% of these ascribed to the Sindhu Desh Liberation Army. As discussed in Section 5.2, our method appears to pick up an organized group operating across Sindh almost a year before this would have been visible by examining the group identification in the best available datasets. Overall, we see that the GTD reports a single group corresponding to our estimated groups for Balochistan, Sindh, and the area near the Afghan border.

Check 2. As an additional verification of our model, we can consider whether our estimated group structure in Pakistan can predict the geographic structure of attacks in a later period. BFRS and WITS data is not available for more recent years, so we use data from the GTD for this analysis. We use data from the Nov. 2011 - Dec. 2016 period, and run a clustering of this data into four groups.[73] The resulting

[72] Neither the BFRS nor WITS record the group identity of the assailants in a systematic way.

[73] Another possibility would have been to examine data from a point *earlier* than our main period. One of the data requirements for our method to be effective, however, is that there must be simultaneous attacks in the districts that we wish to cluster. Although Pakistan has a long history of terrorism, much of this violence is concentrated in major cities. For example, in 1995 the Global Terrorism Database lists 666 attacks in Pakistan: of those, 614 of them occur in Karachi. The BFRS

Appendix Table M.2: Estimation of $\hat{J}$ based on hierarchical splits

		Pakistan (to Apr '11)
Split at (1)?	Randomly shuffled data (mean)	138.49
	Std. dev.	7.82
	Actual data	159.00
	p-value	0.01
Split at (2)?	Randomly shuffled data (mean)	45.82
	Std. dev.	4.69
	Actual data	58.00
	p-value	0.01
Split at (3)?	Randomly shuffled data (mean)	32.36
	Std. dev.	4.09
	Actual data	50.00
	p-value	0.00
Split at (4)?	Randomly shuffled data (mean)	11.00
	Std. dev.	2.24
	Actual data	13.00
	p-value	0.24
Split at (5)?	Randomly shuffled data (mean)	16.69
	Std. dev.	2.76
	Actual data	21.00
	p-value	0.08
Split at (6)?	Randomly shuffled data (mean)	13.12
	Std. dev.	2.45
	Actual data	15.00
	p-value	0.27
Split at (7)?	Randomly shuffled data (mean)	11.60
	Std. dev.	2.42
	Actual data	13.00
	p-value	0.34

A test statistic Q is computed as described in Section 2.3, based on a within-month covariance matrix as described in Section 2.5. Figure 5 shows the order of the potential splits. Data used is the Pakistan BFRS dataset for May 2008 - April 2011. This is 6 months less data than is used in Table 1, which uses data through to October 2011.

group structure is shown in Figure N.10. There are obvious similarities here to the clustering on the original data shown in Figure 7. To quantify these similarities, we run regressions predicting the new group membership using the original group membership: these are shown in Table N.9. In both figures, we see a group that matches the Sindh ethnic homeland, another in Balochistan, and a third in the area near the Afghan border. The GTD data includes fewer attacks than the BFRS data, and has very few incidents in Punjab. We thus do not see any group corresponding to the Punjabi ethnicity, which is main difference between Figures 7 and N.10. Overall, however, the data shows a high degree of persistence in the structure of simultaneous attacks, which suggests that the methods we describe can be used to predict patterns of insurgent coordination in the future.

Check 3. In our estimation strategy, we calculate a covariance matrix based on daily attack data. One might be concerned that in fact we are discarding useful information by considering only coordination within a single day. For example, perhaps one of the ways an insurgency coordinates is to arrange sequential attacks over consecutive days. We can use the GTD to verify that this does not appear to be the case.

The GTD allows for the component attacks of a multiple attack to occur on different days. 96% of all multiple attacks, however, take place only on one day. In addition, most of the attacks that are spread across multiple days actually take place on two consecutive days, and in some cases the notes for the attacks indicate that the attack took place during a single night, with some components occurring before midnight and others after midnight. We thus see that almost all multiple attacks are indeed same-day simultaneous attacks, rather than spread out across time.[74]

Check 4. A concern is that the "groups" that result from our method do not

data similarly has 79% of all attacks occurring in Karachi. It is thus unsurprising that attempting to cluster other districts does not yield meaningful results. For comparison, only 9% of attacks occur in Karachi in 2009, and this is the most attacks in any district during that year. Because of this feature of the earlier data, it is unfortunately not possible to track changes in the group structure in Pakistan across time.

[74] One of the major advantages that counterinsurgency forces have is that they are generally more numerous and better equipped than the insurgency that they are fighting. The insurgents, on the other hand, have the advantage of surprise, in terms of both timing and location of attacks. If an insurgent group were to advertise that they would attack a week later, the government would be able to place their forces on high alert, change their deployment strategy, cancel leave, and so forth. The insurgents thus face a higher cost in terms of casualties if they attack with advance warning. Horn [2013] cites a Taliban commander describing how simultaneous attacks prevent a concentration of security forces.

match what a qualitative researcher would consider a group to be. For example, they might be too narrow, classifying as different groups what are in reality simply different branches of the same organization. Alternatively, the groups we estimate might be too broad, lumping together different insurgent organizations that merely cooperate occasionally on campaigns. As our definition of a group is based on same-day simultaneous attacks, we can address this concern by examining how these attacks are attributed to insurgent groups by qualitative analysts.

The GTD is again useful here, because it reports group identities where available. Groups in the GTD are defined based on perpetrator information, where "*the perpetrator attributions recorded for each attack reflect what is reported in open-source media accounts, which does not necessarily indicate a legal finding of culpability*" and teams of researchers at START (http://www.start.umd.edu/gtd/using-gtd/) are responsible for the verification and consistency of the entries.

There are 6718 sets of multiple attacks in this database, with an average of 3.4 attacks in each set. Identities of the groups responsible are recorded for at least one attack in 67% of these sets. Only 170 sets of attacks (2.5% of the total) have multiple different groups recorded as being responsible for component attacks within the same set of attacks. Of these, the majority are cases where it is unclear whether the attacks were actually coordinated, and one of the groups is listed as unknown (the notes for these attacks often report this uncertainty). There are only 50 cases (0.7% of the total) where there are actually two distinct group names listed, and about half of these are cases where an identified group is clearly responsible for one of the attacks, but it is unclear whether it also committed the other one, and thus the second attack is listed with a more general group description (e.g. Revolutionary United Front vs. Rebels). There are only a few dozen cases where two different identified groups engage in a simultaneous attack. This happens, for example, in Colombia (ELN and FARC) and Chile (FPMR and MIR). It is thus true that sometimes multiple different groups will engage in simultaneous attacks, but these incidents comprise only a fraction of a percent of all simultaneous attacks.

We thus see that our simultaneous-attack based definition of a group is not too wide compared to the definition used by qualitative sources, because the GTD shows very little coordination of attacks between groups as they define them. A remaining danger is that our definition is too narrow, in that different cells in a group that has a cohesive objective may choose not to coordinate for some reason, and thus we detect too many groups using our method. However, we only detect one group in Afghanistan, and 4 in Pakistan. In Pakistan, separatists in Sindh and Balochistan have their own independent objectives, which are clearly not in alignment with Punjabi interests and also differ from those of the Taliban. It thus seems unlikely that we have detected too many groups in Pakistan, although there does not appear to be a more formal way of testing this using the data sources that we currently have available.

Check 5. Our model suggests that part of the value of the simultaneous attack relies on citizens knowing which group launched the attack, because the attack serves as a signal of this groups strength. It is thus more important that a simultaneous attack actually be attributed to a group, relative to a non-simultaneous attack. In particular, we should expect that groups will claim credit for these attacks at rates that are higher than for non-simultaneous attacks. Table N.5 shows that this appears to indeed be the case, even after controlling for variables that describe the total size and damage that the attacks cause.

Check 6. Another implication of our model is that types of attacks where decentralization is particularly important should be less likely to be simultaneous. For example, consider the difference between bomb attacks against a railroad, versus the assassination of senior government officials. The railroad is close to equally vulnerable every day, although there may be slight variations in the effect of a bombing due to differences in traffic. On the other hand, a given senior government official may be vulnerable to assassination only on certain days, and the probability of an attempt succeeding could vary greatly depending on when the attempt is made. Thus, there would be substantial costs to attempting to coordinate two assassinations: even if the coordinator had perfect information regarding when the targets were vulnerable, the time selected for the attack would still be a compromise that would not have either target at its most vulnerable. We should thus expect that assassinations are much less likely to be part of a simultaneous attack than bombings. Table N.4 shows that this appears to indeed be the case.

Check 7. Our theory of simultaneous attacks sketched in Appendix B is based on a cost-benefit trade-off of launching a simultaneous attack. In the case where a terrorist group is very weak and disorganized, it may be too difficult to attempt a simultaneous attack. Is there evidence that weaker groups are less likely to launch simultaneous attacks? The use of country-year data to answer this question is potentially problematic but provides a valuable starting point.

Appendix Figure N.7: Overdispersion and "Related" Attacks

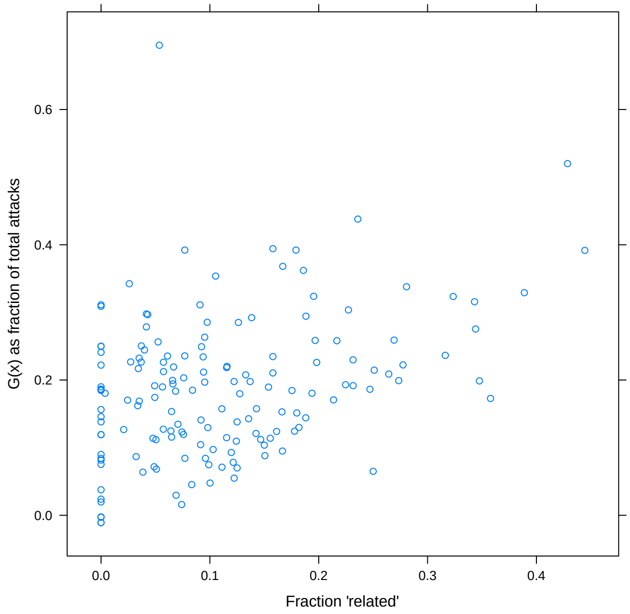
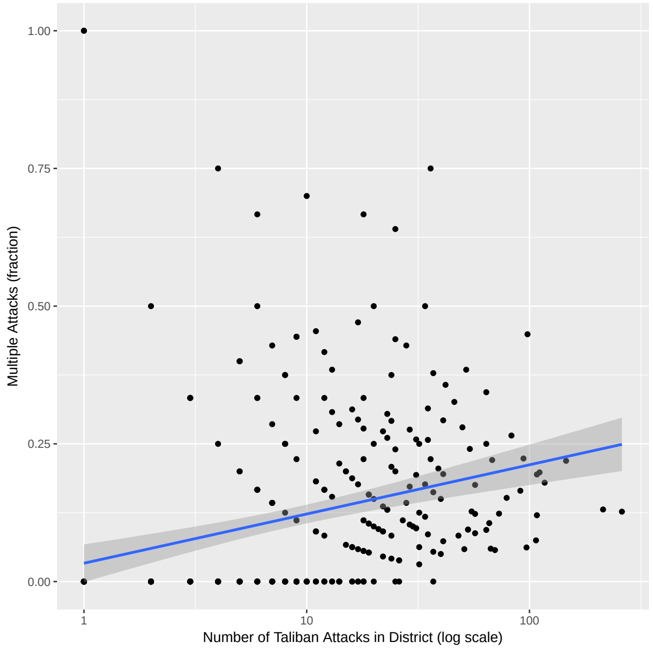


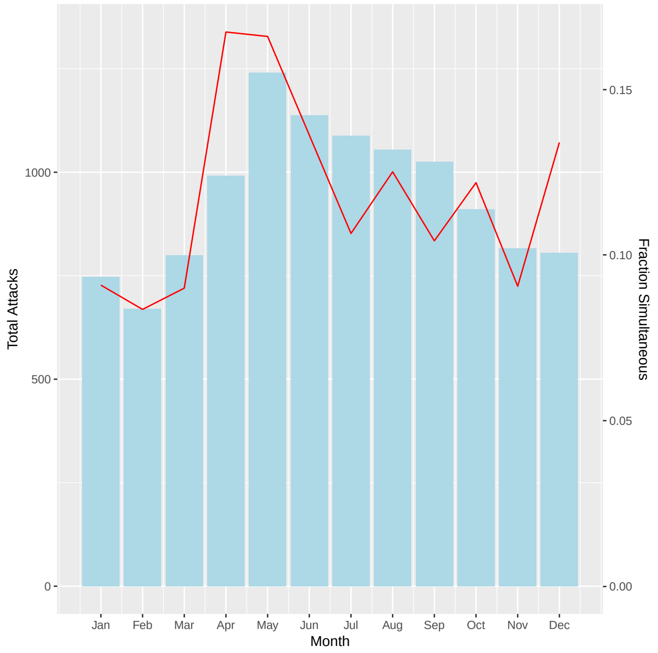
Table N.7 shows that a greater fraction of simultaneous attacks of interest is associated with a higher number of total attacks, even after controlling for country and year fixed effects. However, there is an obvious confounding effect here, because a group that is so weak that it can set off only a single bomb will not be able to launch any simultaneous attacks. One way to attempt to deal with this is by using lagged simultaneous attacks as an instrument for the fraction of simultaneous attacks this year: Columns III and IV of Table N.7 show that results do not change when this approach is used. So, at the very least, evidence from these conditional correlations seems not to counter our intuition.

We further address this point by considering districts of Afghanistan. The advantage here is that even if there is only one attack in a district, it can still be a simultaneous attack because it is coordinated with an attack in another district. Our hypothesis is that districts where the Taliban are weak are districts where it would be very costly for them to coordinate, and thus are districts where they will not engage in simultaneous attacks. Figure N.8 and Table N.3 show that this indeed appears to be the case.

Appendix Figure N.8: Fraction of Taliban Multiple Attacks



Appendix Figure N.9: Seasonality in Multiple Attacks in Afghanistan



Appendix Table N.3: Fraction of Taliban Multiple Attacks

	OLS	OLS	Logistic	Logistic
	I	II	III	IV
(Intercept)	0.235	0.254	-1.253^{*}	-4.630^{***}
	(0.169)	(0.242)	(0.756)	(1.420)
log(Num Attacks)	0.037^{***}	0.039^{***}	0.308^{***}	0.477^{***}
	(0.009)	(0.010)	(0.051)	(0.077)
log(Population)	-0.022^{*}	-0.022	-0.247^{***}	-0.073
	(0.013)	(0.016)	(0.060)	(0.102)
log(Area)	-0.002	-0.002	-0.088^{**}	-0.217^{***}
	(0.008)	(0.012)	(0.037)	(0.063)
Night Lights 92, 00, 12		Yes		Yes
Province FE		Yes		Yes
Observations	341	341	341	341

Note: *p<0.1; **p<0.05; ***p<0.01
Observations are districts in Afghanistan. Dependent variable is the fraction of Taliban attacks that are multiple attacks. Attacks by unidentified attackers and non-Taliban attackers are omitted.
Data source: Global Terrorism Database, 1998-2016

Appendix Table N.4: Is the attack part of multiple attacks?

	OLS	OLS	Logit	Logit
	(1)	(2)	(3)	(4)
Armed Assault	0.118^{***}	-0.016	-2.011^{***}	-3.514^{***}
	(0.002)	(0.010)	(0.021)	(0.120)
Assassination	0.032^{***}	-0.036^{***}	-3.397^{***}	-4.352^{***}
	(0.005)	(0.012)	(0.077)	(0.201)
Bombing/Explosion	0.172^{***}	0.070^{***}	-1.568^{***}	-2.584^{***}
	(0.002)	(0.009)	(0.012)	(0.101)
Facility/Infrastructure	0.288^{***}	0.042^{***}	-0.903^{***}	-2.904^{***}
	(0.005)	(0.015)	(0.034)	(0.145)
Hijacking	0.109^{***}	0.027	-2.100^{***}	-3.035^{***}
	(0.024)	(0.039)	(0.216)	(0.418)
Hostage-Barricade	0.141^{***}	-0.043	-1.808^{***}	-3.609^{***}
	(0.024)	(0.034)	(0.197)	(0.372)
Hostage-Kidnapping	0.100^{***}	-0.048^{***}	-2.200^{***}	-3.692^{***}
	(0.005)	(0.015)	(0.043)	(0.171)
Unarmed Assault	0.173^{***}	0.005	-1.562^{***}	-3.211^{***}
	(0.016)	(0.027)	(0.121)	(0.295)
Unknown	0.183^{***}	0.032	-1.498^{***}	-2.995^{***}
	(0.007)	(0.022)	(0.048)	(0.212)
log(Num Perpetrators)		0.050^{***}		0.435^{***}
		(0.002)		(0.023)
log(Num Killed + 1)		-0.001		-0.006
		(0.004)		(0.041)
log(Num Wounded + 1)		0.015^{***}		0.141^{***}
		(0.003)		(0.031)
Country FE		Yes		Yes
Observations	89,338	14,156	89,338	14,156

Note: *p<0.1; **p<0.05; ***p<0.01

Observations are individual terrorist attacks in all countries. Dependent variable is binary: whether or the attack is part of a set of simultaneous (same day) attacks. Attack types are an exhaustive set of dummy variables.

Appendix Table N.5: Was the attack claimed by a terrorist group?

	OLS	OLS	OLS	Logit	Logit	Logit
	I	II	III	IV	V	VI
(Intercept)	0.131^{***}	0.109^{***}		-1.892^{***}	-1.940^{***}	
	(0.001)	(0.006)		(0.011)	(0.038)	
Multiple Attack	0.105^{***}	0.112^{***}	0.068^{***}	0.720^{***}	0.555^{***}	0.531^{***}
	(0.003)	(0.013)	(0.012)	(0.023)	(0.076)	(0.102)
log(Total Num Perpetrators)		0.004	-0.012^{***}		0.020	-0.111^{***}
		(0.003)	(0.003)		(0.017)	(0.025)
log(Total Num Killed + 1)		0.054^{***}	0.023^{***}		0.298^{***}	0.169^{***}
		(0.005)	(0.004)		(0.029)	(0.043)
log(Total Num Wounded + 1)		0.033^{***}	0.022^{***}		0.174^{***}	0.184^{***}
		(0.004)	(0.003)		(0.023)	(0.032)
Country FE			Yes			
Terrorist Group FE			Yes			Yes
Observations	87,901	13,441	13,441	87,901	13,441	13,441

Note: *p<0.1; **p<0.05; ***p<0.01

Observations are individual terrorist attacks in all countries. Dependent variable is binary: whether or not a terrorist group claimed responsibility for the attack. In the case of multiple attacks, “Total Num” refers to the total number of perpetrators (etc.) in all of the attacks combined.
Column VI omits country fixed effects due to convergence issues (very few terrorist groups span multiple countries).

Appendix Table N.6: Dependent variable is Herf. fragmentation of terrorist groups

	I	II	III	IV
(Intercept)	0.53*	0.06		
	(0.01)	(0.08)		
Overdispersion	-0.28^*	-0.32^*	-0.28^*	-0.28^*
	(0.04)	(0.03)	(0.03)	(0.03)
Max Possible Overdispersion		0.50*	0.51*	0.45*
		(0.05)	(0.06)	(0.06)
FKMS Controls	No	Yes	Yes	Yes
Country FE	No	No	Yes	Yes
Year FE	No	No	No	Yes
N	1144	1143	1143	1143
R^2	0.05	0.22	0.85	0.86

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Observations are an unbalanced panel in country and year. Dependent variable is the Herfindahl fragmentation of terrorist attacks by terrorist group within a given country-year. The range of the dependent variable depends on the number of terrorist attacks that occurred: for example, with only one terrorist attack, the only possible fragmentation is 0, while with two terrorist attacks the possible levels are 0 and 0.5. The control variable "Max Possible Overdispersion" is the maximum possible fragmentation given the number of attacks that occurred. A more sophisticated adjustment appears not to exist: see Gotelli and Chao [2013] for discussion.
"FKMS Controls" are the covariates used in Table 1 of Freytag et al. [2011].[75]

Appendix Table N.7: Relationship between number of attacks and overdispersion

	OLS	OLS	IV	IV
	1	2	3	4
(Intercept)	1.517^{***}		0.543^{*}	
	(0.041)		(0.297)	
Overdispersion	2.681^{***}	1.567^{***}	9.447^{***}	10.719^{**}
	(0.198)	(0.115)	(1.574)	(4.413)
FKMS Controls	No	Yes	No	Yes
Country FE	No	Yes	No	Yes
Year FE	No	Yes	No	Yes
Observations	1,940	1,939	1,568	1,568

Note: *p<0.1; **p<0.05; ***p<0.01

Observations are an unbalanced panel in country and year. Dependent variable is the log number of terrorist attacks in a given country-year. “Overdispersion” is $G(x)$, as defined in the text. “FKMS Controls” are the covariates used in Table 1 of Freytag et al. [2011]. Columns 3 and 4 use the previous year’s overdispersion as an instrument for current overdispersion.

Appendix Table N.8: “Related” attacks and overdispersion

	I	II	III	IV
(Intercept)	0.01	0.02		
	(0.00)	(0.02)		
Overdispersion	0.34^{*}	0.34^{*}	0.34^{*}	0.35^{*}
	(0.03)	(0.03)	(0.03)	(0.03)
FKMS Controls	No	Yes	Yes	Yes
Country FE	No	No	Yes	Yes
Year FE	No	No	No	Yes
N	2006	2005	2005	2005

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

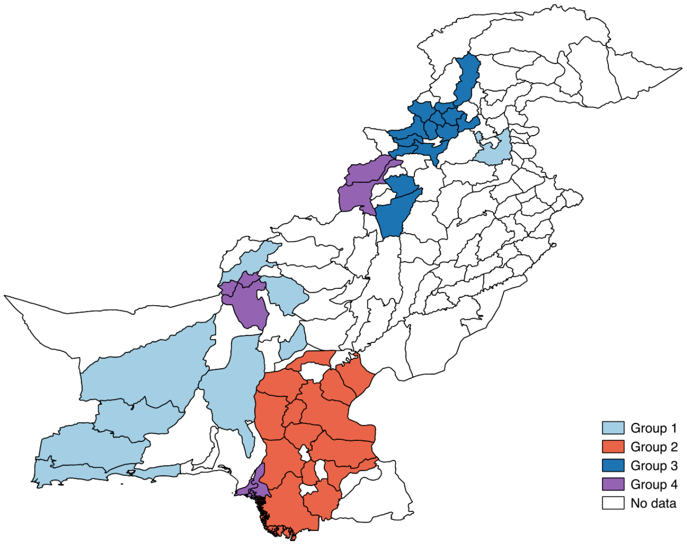
Observations are an unbalanced panel in country and year. Dependent variable is the fraction of terrorist attacks in a given country-year that had “related” attacks. “Overdispersion” is $G(x)$, as defined in the text. “FKMS Controls” are the covariates used in Table 1 of Freytag et al. [2011].

Appendix Table N.9: Pakistan Comparison using post- Nov 2011 GTDB data

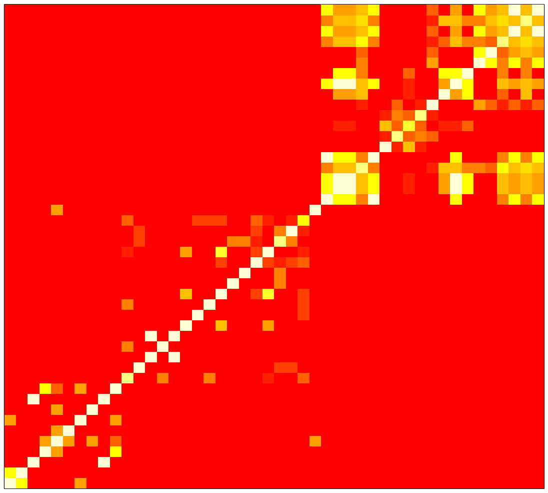
	Group 1	Group 2	Group 3	Group 4
Group 1 (mostly Baloch)	0.63	0.00	0.13	0.25
	(0.13)	(0.12)	(0.13)	(0.12)
Group 2 (mostly Sindhs)	0.07	0.80	0.07	0.07
	(0.09)	(0.08)	(0.10)	(0.09)
Group 3 (mostly Afghans)	0.08	0.08	0.77	0.08
	(0.10)	(0.09)	(0.11)	(0.09)
Group 4 (mostly Panjabis)	0.33	0.17	0.33	0.17
	(0.15)	(0.14)	(0.15)	(0.14)
N	42	42	42	42

Each column corresponds to a single regression without intercept.
The dependent variable is a dummy variable indicating whether a given district was clustered into the specified group number in the clustering shown in Figure N.10.
The independent variables are a set of dummy variables, indicating whether a given district was clustered into the specified group number in the clustering shown in Figure 7c. Districts shown as white ("no data") in either Figure 7c or N.10 are dropped: the remaining 42 districts are used in the regression.
Each row should sum to 1 because each coefficient in the table is a conditional mean giving the fraction of districts of the specified ethnicity that were clustered into the specified group, and the clustering in Figure N.10 assigns each district to one group. Rows may not sum exactly to 1 because of rounding.
Standard errors in parentheses.

Appendix Figure N.10: Pakistan Groups with post- Nov 2011 GTDB Data



Appendix Figure N.11: Covariance Matrix for post- Nov 2011 GTDB Data



Cells of cross-district covariance matrix, coloured from low covariance (red) to high covariance (white). Ordering of rows and columns is the default order for GIS maps of Pakistan, which places districts in the same province together. Three groups are clearly visible. The GTDB data contains very few attacks in Punjab: no group corresponding to Punjab is visible. This data is clustered to produce Figure N.10.

REFERENCES

[1] Anderson, Carl A. (1974) “Portuguese Africa: A Brief History of United Nations Involvement” Denver Journal of International Law & Policy 133

[2] Anderson, T.W. (1963) “Asymptotic Theory for Principal Component Analysis” Annals of Mathematical Statistics 122-148.

[3] Baurle, G. (2013) “Structural Dynamic Factor Analysis Using Prior Information From Macroeconomic Theory ” Journal of Business & Economic Statistics. 31(2):136-150.

[4] Choi, W.G.; Kang, T.; Kim, G.Y.; Lee, B. (2014) “Global Liquidity Transmission to Emerging Market Economies, and Their Policy Responses” . SSRN Scholarly Paper 2580627. December 2014.

[5] Eisenstadt, Michael and Jeffrey White (2005) “Assessing Iraq’s Sunni Arab Insurgency” The Washington Institute for Near East Policy Policy Focus No.50.

[6] Fernandes, Clinton (2008) Hot Spot: Asia and Oceania. ABC-CLIO

[7] Freytag, Andres, Jens J. Kruer, Daniel Meierrieks, Friedrich Schneider. (2011). The origins of terrorism: Cross-country estimates of socio-economic determinants of terrorism. European Journal of Political Economy 27 S516

[8] Good, Phillip. (2002) Extensions of the Concept of Exchangeability and their Applications. Journal of Modern Applied Statistical Methods. 1(2) 243-247.

[9] Good, Phillip. (2005) Permutation, Parametric, and Bootstrap Tests of Hypotheses. New York: Springer.

[10] Gotelli Nicholas J., and Chao Anne (2013) Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data. In: Levin S.A. (ed.) Encyclopedia of Biodiversity, second edition, Volume 5, pp. 195-211. Waltham, MA: Academic Press..

[11] Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning. New York: Springer.

[12] Henderson, Anne. (2005) The Coalition Provisional Authority's Experience: with Economic Reconstruction in Iraq: Lessons Identified. USIP Special Report No. 138. http://www.usip.org/files/resources/sr138.pdf

[13] Ledermann, W. (1940). "On a Problem concerning Matrices with Variable Diagonal Elements." Proceedings of the Royal Society of Edinburgh. 60(1). 1–17.

[14] Leites, Nathan and Charles Wolf. (1970). Rebellion and Authority. Chicago, IL: Markham.

[15] Ng, A. Y., Jordan, M., &Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), Advances in neural information processing systems, 14. Cambridge, MA: MIT Press.

[16] Pesarin, Fortunato (2001) Multivariate Permutation Tests, New York: Wiley.

[17] Saunderson, J.; Chandrasekaran, V.; Parrilo, P.; Willsky, A. (2012). "Diagonal and Low-Rank Matrix Decompositions, Correlation Matrices, and Ellipsoid Fitting." SIAM Journal on Matrix Analysis and Applications. 33(4): 1395–1416.

[18] Shi, J. and Malik, J. (2000). "Normalized cuts and image segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888 – 905.

[19] Smith, B. (2012). "Syria: No End in Sight?" House of Commons Library. Research Paper 12/48.

[20] Subrahmanian, V. S., Aaron Mannes, Animesh Roul, R. K. Raghavan (2013) Indian Mujahideen: Computational Analysis and Public Policy, Springer.

[21] Tibshirani, R., Walther, G., and Hastie, T. (2001) "Estimating the number of clusters in a data set via the gap statistic". J. R. Statist. Soc. B, 63, Part 2, 411-423.

[22] Wu, Y.; Moon, H.R.; Deng, Y. (2011) "Factor Analysis on US Housing Price Indexes." USC Lusk Center Working Paper.

[23] Yao, J., Zheng, S., and Bai, Z. (2015) Large Sample Covariance Matrices and High-Dimensional Data Analysis, Cambridge University Press..