Computational Methods to Discover Sets of Patterns of Behaviors that Precede Political Events of Interest*

Kurt Rohloff
BBN Technologies
10 Moulton St.
Cambridge, MA 02138
krohloff@bbn.com

Victor Asal
University at Albany, SUNY
135 Western Ave.
Albany NY 12203
vasal@email.albany.edu

Abstract

In this paper we present an approach to identify sets of patterns of behaviors which precede political events of interest (EoIs) such as the the onset of regime change, insurgency, ethnic violence, etc.. We define a pattern to be an identified set of values of sampled, quantized factor data which occurs before at least two instances of an EoI and only before the occurrences of EoIs. Not all EoIs instances exhibit the same patterns preceding their occurrence, but we hypothesize that there exist sets of patterns which, taken together, precede all EoIs of the same type. A set of patterns which taken together precede all EoIs of the same type are called a "cover". We describe a computationally efficient cover discovery operation based on a randomized greedy algorithm which grows patterns simultaneously with the cover. This cover discovery algorithm was implemented in the Java programming language. Although the optimal cover discovery problem is NP-complete, our algorithm runs in polynomial time and returns nontrivial results.

Introduction

A major challenge in computational social science is the problem of identifying sets of symptomatic precursors to political Events of Interest (EoIs) such as rebellion, insurgency, civil war, etc. which can describe all occurrences of those EoIs. We hypothesize that EoI causal mechanisms have the property of equifinality - a country may or may not follow multiple patterns leading to EoI occurrence simultaneously, and there may be multiple means to a same end. For example, a country such as India may contain multiple types of rebellion or may exhibit the antecedents for several rebellions simultaneously.

Unfortunately, and as may be expected, due to the equifinality hypothesis there may be multiple patterns which precede the onset of EoI occurrence. Hence not all occurrences of an EoI can be described by a single pattern. There has been previous work that explores the concept of patterns preceding occurrences of EoIs ((Rohloff and Asal 2008)), but this previous work focused on finding single patterns that describe countries' behaviors before chosen sets of EoI occurrences. This paper addresses the major challenge of finding the smallest (or at least a small) set of patterns which "cover" all observed occurrences of an EoI. The computational challenge associated with this problem is to identify which EoI occurrences have common preceding patterns and select a small subset of those patterns so that all EoI advents in a set of training data are covered by at least one pattern.

The major contributions of this paper are two-fold - 1) a formalization of the cover discovery problem for patterns preceding EoI occurrences and 2) a scalable, real-world approach to solve this problem. We present a greedy covering approach to identify sets of patterns, called covers, which describe the dynamic conditions immediately preceding sets of EoI occurrences. With knowledge of patterns preceding all previously observed occurrences of EoIs we would not only be able to better understand the factors that commonly drive (or at least precede) political instability, but we would have an approach to forecasting the occurrence of political instability.

Our patterns are formalized through the output of the "backwards chaining" methodology to discover quantitative changes in the properties of a country which precede political instability events (Rohloff and Asal 2008). Our approach to pattern discovery is based on the supposition that the phenomena which cause (or at least are related to) the occurrences of some non-trivial subset of EoIs exhibit similar symptomatic behaviors across these multiple EoI occurrences. For example, for countries with rebellions driven by the desire for freedom by internal ethnic groups commonly exhibit increasing ethnic tension and violence before the occurrence of ethnic rebellions. Our methodology to discover patterns is served by the use of regularly sampled factor data. A sampled factor is an quantifiable measurement that may be taken (or sampled) from a country at discrete, regular points in time. Example factors potentially include GDP, the rates of occurrence of various words in the national press, the average caloric intake, etc... The sampling period of factor data may be over any regular period (such as yearly, monthly, weekly) as long as the sampling period and measurement methodology is constant. Naturally, some factors may be easier to measure accurately than others.

To identify patterns of changes in factors preceding political instability events we used the backwards chaining

*Supported by AFRL/HECS contract FA8650-07-C-7748

methodology. The backwards chaining methodology permits us to identify which factors change identically over a fixed number of time steps in the time period leading up to the occurrences of events of interest in selected countries. The identification of how specific factors change over time leading to the occurrence of an EoI defines a pattern in the context of our backwards chaining methodology.

Unfortunately, the size of the set of all possible patterns for subsets of EoI occurrences is exponential in the number of EoI occurrences in the worst case. Therefore we cannot use a brute-force approach to discovering all possible patterns that could be used to form a cover. We present a greedy algorithm which incrementally grows patterns that cover a set of EoI occurrences. We implemented this algorithm in the Java programming language.

This document is organized as follows. The section immediately following discusses the theoretical basis for our analysis. We then formalize the covering discovery problem and then review the relevant concepts the backwards chaining approach to pattern definition described in (Rohloff and Asal 2008). We then describe our greedy covering algorithm which runs in polynomial time. We also discuss how our solution methods are scalable. We close the paper with a discussion of our results.

A Theoretical Basis for the Pattern Cover Concept

We start our investigation with the belief that the outbreak of violence will be characterized by equifinality, "many alternative causal paths to the same outcome," (George and Bennett 2005) in other words there is not one cause of EoI outbreaks. There may be a set of factors that make those outbreaks more or less likely but we believe there are a set of potential causal pathways that are likely to lead to outbreak of non-state actor violence towards the state. We are interested in exploring the combinational power of various factors as they lead to EoIs. To do this we build on efforts to use Boolean analysis (Ragin 2000) to understand political activity while allowing for multiple causal mechanisms (Chan 2003). Given that this approach is usually pursued to examine a small number of cases, the focus is on those factors that are thought to be likely to be explanatory rather than also collect data on the general environment (the dogs that dont bark - and the cats that never would) as well as those occasions where there is an outbreak of an EoI (for example see: (Chan 2003)).

The ability to apply a computational approach to this analytical effort allows us to do several things that otherwise not be possible. First we are not limited to a small number of cases nor are we limited to a small number of variables but we are able to look across a wide range of variables to see which variables build patterns that lead to the outbreak of an EoI including focusing on might be unlikely but possible combinations of variables. Second, this computational approach allows us to take into account the way variables combine over time to produce outcomes. Thus a major innovation in our efforts is the ability to meld a component of temporality from a process tracing perspective with the Boolean perspective that is usually used across a wide selection of cases. From a policy perspective our efforts allow us to identify combinations of changing patterns of behaviors within particular larger sociopolitical contexts that are likely, based on past experience to lead to EoI outbreaks at a temporal point distant from the actual outbreak.

For the backwards chaining process we identify which factors change in the same manner for several time steps leading up to all occurrences of a particular EoI. The combination of all of the factors which all change in the same manner for a fixed number of time steps leading up to all occurrences of a particular EoI define a pattern. By identifying which specific factors which in combination exhibit symptomatic behavior leading to EoI occurrences, we are closer to our ongoing goal of obtaining early warning derivative results informed by factor combinatorics in a semi-automated manner. Our hypothesis is that by identifying such a pattern, one can use this pattern to detect conditions which are precursors to the occurrences of EoIs and hence have a forecasting capability for EoI occurrence. It is important to note that a country may or may not follow multiple patterns leading to EoI occurrence simultaneously, and there may be multiple means to the same end. For example, a country such as India may contain multiple types of rebellion or may exhibit the antecedents for several rebellions simultaneously.

It is important to underline that our efforts here differ fundamentally from the usual computational approaches taken to the study of social processes. Most computational analysis embraces a quantitative statistical methodology that examines if a host of variables each separately or (occasionally) in interaction impact significantly a dependent variable of interest. The relative nature of this impact can then be expressed in probabilities and caveated by confidence intervals. The approach has the advantage of controlling for a variety of variables and for having an established metric for the impact of each variable. Researchers who use qualitative literature have identified several key problems with this approach though. For our purposes it is important to note two of these critiques. First the statistical approach focuses on probabilities rather than patterns. That is that the possibility that it is a wide range of factors forming a specific constellation of factors that account for the advent of a particular EOI is missing from this approach (Ragin 2000). In addition what is missing is an accounting of this constellation might combine over time to produce certain results that are temporally dependent and contingent (George and Bennett 2005). We take this temporal path dependency seriously. Finally our approach differs because our approach takes equifinality, the argument that different combinations of factors can result in the same EOI, seriously (George and Bennett 2005). Instead of using a probabilistic approach, we look to identify those temporal pattern combinations that explain the outbreak of an EOI and we look to identify all the patterns that do so in a coherent fashion. Our covering efforts allow us to do this in a novel way examining a wide range of data at the same time that goes far beyond what similarly minded qualitative approaches have been able to do so far.

Our analytical effort is being driven by an approach that does not privilege any particular social science theoreti-

cal bias related to violent conflict (for example the greed - grievance argument). Instead we have endeavored to draw from various theoretical models we have been able to find and to operationalize theories in an analytically useful way. Our efforts are directed at creating a synthesis that draws from each of major theoretical models that tries to explain substate political violence. We drew explicitly from the greed (Collier et al. 2003), grievance (Gurr 2000) , resource mobilization (McAdam, McCarthy, and Zald 1996), political opportunity structure, (external and demographic) pressures (Tarrow 2001), culture/values (Goldstone 2001) and leadership literatures (Herman and Herman 1989). In this effort we are looking to identify factors that help explain rare events - the outbreak of different kinds of political violence. In general we view EoI emergence, as a product of interactions between causal factors at different levels of a social environment. Some aspects of that social environment are more changeable and stochastic while other aspects are relatively rigid and predictable. These variables combine in different patterns to produce future behavior-in our case patterns of violence.

In our current analysis we focus on the variables that change leading up to the EoIs but we should point out that the backward chaining approach we are using allows us to also identify the ongoing unchanging factors that go into creating a state space where the changes in particular behavioral and policy factors move a country towards experiencing an EoI. The larger state space of countries primed for an EoI are defined by elements like general political instability married to anocratic (regimes between democracy and autocracy) political structures or low levels of militarization. What we find is that when these types of conditions exist the stage is set for an EoI. The stage though is not what sparks the EoI.

Across the different EoIs we are finding that a general decline in good rhetoric and behavior and a rise in bad rhetoric and behaviors give rise to the outbreak of EoIs given certain structurally negative state spaces as noted above. Certain variables repeat across EoIs as leading behavioral indicators. Specifically:

- The general count of good behavior goes down.
- Efforts at public diplomacy go down.
- Protest behavior tends to go down.

The first two changes are fairly intuitive - good behavior is an encoding of how often good expressions appeared in the popular press for a given country over a 1-month time span. The changes latter may be seen as counterintuitive in that we can think of protests as a step forward on the continuum of contentious behavior (McAdam, Tarrow, and Tilly 2001). On the other hand if we envision contention as a choice between several kinds of contention (each one exacting a cost) then the withdrawal from a non-violent contention may be a sign that opposition groups are repositioning their resources for more violent approaches and governments are expending resources to drive this behavior down - perhaps unwittingly pushing opponents into activities that are much more dangerous. Overall, our approach allows us to model different patterns of change within a broader unchanging state space that leads to violence

The Pattern Covering Problem

In this section we formalize the cover discovery problem.

For every class of EoI (such riot, rebellion, coups, etc...) there are a set of instances of those EoIs. More formally, $E = \{e_1, e_2, \ldots\}$ where e_i is an instance of an EoI and E is the set of instances of a class of EoI. The set E could include all instances of coups over a geographic area and region in time.

A pattern P describes the dynamic sets of behaviors which precede some subset of EoI instances. Formally, each pattern describes behavior preceding multiple EoI occurrences: $\{e_1^{P_1}, e_2^{P_1}, \cdots\} = E_{P_1} \subseteq E$. We call a pattern *sufficient* if it discriminates EoI instances. That is, a pattern is sufficient for a geographic area and region in time if an EoI occurs after every time a country follows a pattern. For any set of EoI instances $\{e_{i_1}, e_{i_2}, \ldots\}$ there may or may not be a pattern that describes those sets of EoIs.

With this motivation we define a cover to be a set of patterns such that each EoI instance is described by at least one pattern. More formally, a cover C is a set of patterns $C = \{P_1, P_2, \cdots\}$ such that

$$E = \cup_{P_i \in C} E_{P_i}. \quad (1)$$

We therefore define the cover discovery problem such that when given a set of EoI occurrences $E = \{e_1, e_2, \ldots\}$, find a set of patterns $C = \{P_1, P_2, \cdots\}$ such that $E = \cup_{P_i \in C} E_{P_i}$. Ideally the cardinality of the cover ($|C|$) should be as small as possible.

Unfortunately the brute-force approach to cover discovery is computationally intractable in the worst case. For a set of patterns $\{P_1, P_2, \ldots\}$, the problem of selecting the smallest set of patterns to form a cover C is equivalent to the Set-Covering problem which is well known to be NP-complete (Garey and Johnson 1979). Notice also that the set of all subsets of E is exponential in the cardinality of E so it is infeasible to compute all sets of patterns that can be used to define a cover. Therefore we cannot directly use standard randomized set-covering algorithms that could otherwise be used to discover minimal pattern covers with high probability (Cormen, Leiserson, and Rivest 1990; Motwani and Raghavan 1995).

Fortunately, using the backwards chaining methodology (outlined immediately below), when given a set of events $\{e_{i_1}, e_{i_2}, \ldots\}$ we can quickly test for and find a pattern P such that $E_P = \{e_{i_1}, e_{i_2}, \ldots\}$ if such a pattern exists. We can take advantage of this pattern discovery approach to efficiently find a small pattern cover with high probability using a randomized greedy covering algorithm described below.

The Backwards Chaining Methodology

In this section we present our methodology for automatically discovering patterns based on the backwards chaining methodology process. We implemented this process in the Java programming language to automatically obtain data

from a data server, search over the data in an automated manner to identify key factor changes that precede selected EoIs to identify patterns. A version of this pattern discovery process is discussed in more detail in (Rohloff and Asal 2008).

To discover patterns using the backwards chaining methodology we developed algorithms and wrote software to identify factors that change "identically" over a fixed number of sample times in the time period leading up to the occurrences of user-selected EoI advents. We define an equivalence relationship for the factor values based on quantization levels of those factors that was implemented in our factor identification tool. We use that equivalence relationship to determine when changes in factors are similar enough to be called "identical".

For the backwards chaining methodology, we define a pattern for the advent of an EoI to be:

1. A set of factors, and
2. A description on how each of those factors change quantitatively for a number of time steps before the advent of an Event of Interest in at least two distinct instances.

We require that patterns *discriminate*. That is, the patterns should not match the behavior in any country when and where an EoI does not occur. The set of factors which define a pattern may include a factor that represents previous occurrences of the advent of the EoI itself. By the nature of the factor data used in our pattern discovery process, we can incorporate discrete event occurrences in our patterns by representing the occurrences of these events as a binary factor. In this manner we can incorporate catastrophic events in our patterns such as the ongoing global financial crisis or a natural disaster of very large proportions like the tsunami in 2006 that may not be directly related to a political event of interest (EoIs) but have a significant impact on EoIs.

An example of a hypothetical identification of two factors that change identically in the time preceding the occurrence of an EoI is seen in Figure 1. This figure shows the values of two factors (quality of government and level of corruption) for two countries for several quarters preceding the occurrence of the EoI rebellion. The trajectory of one country is shown using a black line and the trajectory of a second country is shown using a light blue line. In this example, the values of the Quality of Government and Corruption are nearly the same for up to three quarters before the occurrence of Rebellion.

Although the graphical example in Figure 1 identifies a pattern in two different countries for two factors, our patterns could be derived from the behavior of any number of countries using any number of factors. Most importantly, our pattern discovery process search over the set of available factors to identify the specific factors that change identically in the time-steps leading up to EoI occurrences.

In the context of our backwards chaining methodology, the set of factors that define a pattern represents the specific aspects of the condition of a country at particular moments in time. Based on discovered patterns of changes in factors leading to the advent of EoIs, we can generate early-warning forecasts of EoIs if early portions of the patterns are observed in real-time for a specific country.

In general, we search for patterns that lead to EoIs driven by (or at least related to preceding) government policy and immediate antecedent behavior. Examples of this preceding (and possibly driving) behavior include shifts in government policy and economic performance. These antecedent conditions can create agitation and spark violence amongst the country's population when expectations are let down or there is a spike in repression. It is important to note that we are searching for and forecasting on factors acting in combination and over time which cause the advent of events of interest. Contextual information is non-trivial: some factors in a pattern state space might not change over time, but they set an important context for the country's state evolution.

An overview of our process of identifying factors for a pattern are as follows: (We describe these steps in detail below.)

1. Identify EoI occurrences for which patterns should be identified.
2. Quantize Factor Data.
3. Determine which factors are identical for all instances for a user-specified number of times steps before EoI occurrence.

This back-chaining process is discussed in more depth in (Rohloff and Asal 2008). However, to present our cover discovery algorithm in the immediately following section we represent the result of our backwards-chaining pattern discovery algorithm as $\tilde{P}(E_i, F)$ for the factor data F and a set of EoIs $E_i = \{e_{i_1}, e_{i_2}, \ldots\}$. In general $\tilde{P} : 2^E \times F \rightarrow \{\emptyset, P_1, P_2, \ldots\}$ where $\tilde{P}(E_i, F) = \emptyset$ if a run of the backwards chaining algorithm cannot find a pattern to describe

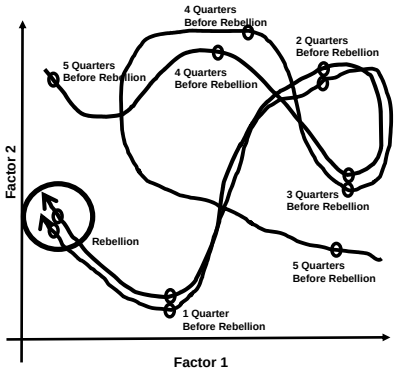

Figure 1: The identification of factors leading to an EoI.

the behavior preceding the EoIs in E_i. Note that although $\tilde{P}(E_i, F)$ is a random variable, $\tilde{P}(E_i, F) = \emptyset$ if and only if there is no pattern of behavior to describe the changes in factor data preceding the occurrences in E_i.

Pattern Covering

We use our pattern discovery process to identify covers. The basis of our cover discovery operation is that we incrementally attempt to identify patterns that describe a specific portion of the EoI occurrences in our training data which are not adequately described by other discovered patterns. Our basis of this cover discovery is a randomized process that repeatedly selects randomized subsets of EoIs and discovers a minimized pattern for that subset of EoIs. This pattern discovery process terminates when all EoIs are either covered by at least one pattern or a pattern has attempted to have been constructed for every EoI not covered and every other EoI. An algorithm for our pattern discovery process is below in Algorithm 1.

Data: A set of EoIs E and factor data F.
Result: A cover $C = \{P_1, P_2, \ldots\}$ such that $E = \cup_{P_i \in C} E_{P_i}$.
$C := \emptyset$;
$P := \emptyset$;
$toCover := E$;
$covered := \emptyset$;
while *toCover is non-empty* **do**
  Remove randomly selected e from $toCover$;
  $E' := \{e\}$;
  $toTest := toCover \cup covered$;
  while $toTest \neq \emptyset$ **do**
    Remove randomly selected e from $toTest$;
    $E' := E' \cup \{e\}$;
    Compute $P := \tilde{P}(E', F)$;
    if $P == \emptyset$ **then**
      Remove e from E';
    else
      if $e \in toCover$ **then**
        move e from $toCover$ to $covered$;
      end
    end
  end
  if $P \neq \emptyset$ **then**
    $C := C \cup P$
  end
end

Algorithm 1: Randomized Greedy Cover Discovery

In Algorithm 1, patterns are iteratively grown in a randomized manner such that every grown pattern covers at least one previously uncovered EoI. The algorithm terminates when all EoIs are covered. To discover the patterns in the cover, the algorithm iteratively and randomly grows a subset of EoIs and a corresponding pattern from a "seed" EoI that has not been previously covered by a pattern in the cover. As this subset of EoIs and the pattern is grown, the algorithm iteratively and randomly attempts to add additional EoIs to the pattern subset. If an EoI can be added, then it is kept in the subset. If the EoI cannot be added, then it is removed from the subset.

The variable $toCover$ tracks the set of EoIs that have not been successfully added to a pattern. Note that if there exists a cover, Algorithm 1 will find one. The only way that an EoI is not covered by a pattern in a cover would be if the EoI were one of the "seed" EoIs such that no other EoIs could be used with it to construct a valid pattern P. This happens if $\forall e' \in E, \tilde{P}(\{e, e'\}, F) = \emptyset$. If this is the case then there can be no cover for all EoIs in E.

Note that Algorithm 1 runs in polynomial time with respect to $\|E\|$. This is discussed in more detail in the section immediately following. Although Algorithm 1 will not always find the *optimal* cover, we have found in practice that it generally finds a cover with an acceptably low cardinality. Indeed, due to its similarity to greedy randomized algorithms for the set covering problem, one can show that Algorithm 1 has a good performance ratio for the closeness of the cardinality of its covering solution to the optimal covering solution.

Scalability

An advantage of our cover discovery process is that it scales well to situations with very large numbers of factors and very large numbers of EoIs. The keystone of this scalability for cover discovery is the scalability of the back-chaining process for pattern discovery. As discussed in (Rohloff and Asal 2008), the computation time of the our backwards chaining pattern discovery process grows linearly in the number of factors and linearly in the number of countries used to define a pattern. This is because the back-chaining process tests factor for inclusion in the pattern independently of one another and are tested once for each country at a time. To use assymptotic big-O notation, $O(t(\tilde{P}(E, F))) \subseteq O(|E| poly(|F|))$

Algorithm 1 calls the pattern discover back-chaining process a polynomial number of times with respect to the number of EoIs to be covered to leverage the scalability of the pattern-discovery process for a scalable cover discovery process. The outer loop of Algorithm 1 iterates at most $|E|$ times and the inner loop of Algorithm 1 similarly iterates at most $|E|$ times. Consequently the run time of Algorithm 1 is in $O(|E|^2) \times O(t(\tilde{P}(E, F))) \subseteq O(|E|^3 poly(|F|))$. Hence Algorithm 1 is scalable.

Conclusion

In this paper we presented the notion of a pattern cover to describe patterns of possible behavior preceding the onset of EoIs. We formalized the problem of cover discovery based and presented an computationally efficient randomized algorithm that discovers covers. If a pattern exists, then our algorithm will find a cover to solve the cover discovery problem.

References

Chan, S. 2003. Explaining war termination: A boolean analysis of causes. *Journal of Peace Research* 40(1):49–66.

Collier, P.; Elliott, L.; Hegre, H.; Hoeffler, A.; Reynal-Querrol, M.; and Sambanis, N. 2003. *Breaking the Conflict Trap: Civil War and Development Policy*. Oxford: World Bank and Oxford University Press.

Cormen, T. T.; Leiserson, C. E.; and Rivest, R. L. 1990. *Introduction to algorithms*. Cambridge, MA, USA: MIT Press.

Garey, M., and Johnson, D. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

George, A. L., and Bennett, A. 2005. *Case Studies and Theory Development in the Social Sciences*. Boston: The MIT Press.

Goldstone, J. A. 2001. Toward a fourth generation of revolutionary theory. *Annual Review of Political Science* 4(1):139–187.

Gurr, T. R. 2000. *People vs. States*. Washington D.C.: United States Institute of Peace.

Herman, M. G., and Herman, C. F. 1989. Who makes foreign policy decisions and how: An empirical inquiry. *International Studies Quarterly* 33(4):361–387.

McAdam, D.; McCarthy, J. D.; and Zald, M. N. 1996. *Comparative perspectives on social movements : political opportunities, mobilizing structures, and cultural framings*. Cambridge, New York: Cambridge University Press.

McAdam, D.; Tarrow, S.; and Tilly, C. 2001. *Dynamics of Contention*. Cambridge: Cambridge University Press.

Motwani, R., and Raghavan, P. 1995. *Randomized algorithms*. New York, NY, USA: Cambridge University Press.

Ragin, C. C. 2000. *Fuzzy-set social science*. University of Chicago Press.

Rohloff, K., and Asal, V. 2008. The identification of sequential patterns preceding the occurrence of political events of interest. In *Proceedings of the 2nd International Conference onComputational Cultural Dynamics (ICCCD)*.

Tarrow, S. 2001. Transnational politics: Contention and institutions in international politics. *Annual Review of Political Science* 4(1):1–20.