Advance Access publication September 3, 2010 *Political Analysis* (2010) 18:403–425
doi:10.1093/pan/mpq024

(Mis)Using Dyadic Data to Analyze Multilateral Events

Paul Poast
Department of Political Science, University of Michigan, 5700 Haven Hall, Ann Arbor, MI 48109
e-mail: poastpd@umich.edu

Dyadic (state-pair) data is completely inappropriate for analyzing multilateral events (such as large alliances and major wars). Scholars, particularly in international relations, often divide the actors in a multilateral event into a series of dyadic relations. Though this practice can dramatically increase the size of data sets, using dyadic data to analyze what are, in reality, *k*-adic events leads to model misspecification and, inevitably, statistical bias. In short, one cannot recover a *k*-adic data generating process using *dy*-adic data. In this paper, I accomplish three tasks. First, I use Monte Carlo simulations to confirm that analyzing *k*-adic events with dyadic data produces substantial bias. Second, I show that choice-based sampling, as popularized by King and Zeng (2001a, Explaining rare events in international relations. *International Organization* 55:693–715, and 2001b, Logistic regression in rare events data. *Political Analysis* 9:137–63), can be used to create feasibly sized *k*-adic data sets. Finally, I use the study of alliance formation by Gibler and Wolford (2006, Alliances, then democracy: An examination of the relationship between regime type and alliance formation. *The Journal of Conflict Resolution* 50:1–25) to illustrate how to apply this choice-based sampling solution and explain how to code independent variables in a *k*-adic context.

1 Introduction

Consider a simple counterfactual: would Belgium and Turkey be alliance members if not for the presence of the United States in creating and supporting the North Atlantic Treaty Organization? This seems unlikely given political differences, relatively small military capabilities, and the large geographic distance between the two countries. Similarly, should one portray the European theater of World War II as a bilateral war between Germany and Greece? Probably not, given this was a minor campaign in a much larger conflict. Finally, for comparative politics scholars, would the French Greens and French Communists have formed an electoral coalition on their own, without the Socialist party being involved? This seems unlikely. In short, empirical scholars cannot treat the Belgium-Turkey country pair (or dyad) as having formed a bilateral

Author's note: I thank William Roberts Clark, Justine Ersaey, Benjamin Fordham, Robert Franzese, Douglas Gibler, John Jackson, Gary King, Walter Mebane, Robert Mickey, Eric Neumayer, Allan Stam, Johannes Urpelainen, three anonymous referees, and the editors for helpful comments. I also thank Douglas Gibler for sharing his data. A replication and supplemental materials packet, STATA and R software for converting dyadic observations into *k*-adic observations, and code for creating independent variables in *k*-adic data are available on the *Political Analysis* Web site. One can also type ''findit kadcreate'' into the command line of STATA.

alliance, Greece and Germany as having fought an isolated bilateral war, and the French Greens and Communists as having formed a bilateral electoral coalition. This is because, simply put, disregarding these actors' relations with outside actors will cause inferential error.

However, this is exactly the common practice of many empirical scholars, particularly (but not solely) in international relations.[1] When analyzing multilateral events, scholars divide the actors involved into a series of *dyadic* relations (i.e., a U.S.-France-U.K. event is converted into three events: U.S.-France, U.S.-U.K., and France-U.K).[2] This subset of observations is then added to a set of purely dyadic observations. Though this practice can dramatically increase the size of data sets, using *dy*-adic data to analyze what are *k*-adic events leads to model misspecification and, inevitably, statistical bias.

Dyadic data are known to violate the independence assumption underpinning many statistical estimation techniques. Specifically, dyadic data commit four major violations of the independence assumption. First, the observations in the dyad-year are temporally correlated (e.g., the Russia-Germany 1938 dyad and the Russia-Germany 1939 dyad; Beck, Katz, and Tucker 1998). Second, the dyads typically share unexplained heterogeneity (Beck and Katz 2001; Green, Kim, and Yoon 2001; King 2001). Third, the dyads have monadic similarity (e.g., the presence of the United States in the U.S.-France and U.S.-Brazil dyads; Hoff and Ward 2004; Hoff 2005; Ward, Siverson, and Cao 2007). Fourth, Signorino (1999) highlights the failure of scholars to adequately capture the strategic interaction between nations that is implied by dyadic data.

Though accounting for such nonindependence is critical for drawing proper inferences, I am not presenting an alternative procedure for modeling such spatial, strategic, temporal, or monadic interdependencies in the data. Such features of the data will still be present and must still be modeled. Instead, this paper highlights a prior, conceptual issue arising in the context of multilateral decision-making processes—namely, if the data are formed by interactions among $k > 2$ actors, then a dyadic format will not reflect this process regardless of how one models other interdependencies. For example, suppose one accounts for strategic interdependence using an estimator based upon logit quantal response equilibria. In this case, the probability of each outcome is derived by multiplying the probabilities of the actions that lead to the outcome. If the outcome is the result of actions taken by $k = 4$ actors, then considering only the actions of $k = 2$ actors (i.e., using dyadic data) will fail to capture the true probability of an outcome.[3]

This particular limitation of dyadic data is not unknown to scholars. Croco and Teo (2005), using a series of case studies, highlight the inferential bias introduced by splitting multilateral events into dyadic observations. Gibler, Rider, and Hutchison (2005), citing Weede (1980), discuss how Wallace (1976, 1979) overstates the ability of arms races to escalate into wars because he disaggregates one event of arms race–induced escalation into several events,

[1]Examples outside international relations include studies of preelectoral coalitions in comparative politics (see Golder 2006). Interestingly enough, studies on governing coalitions in comparative politics do go beyond the dyad to consider all combinations of governing parties (see Martin and Stevenson 2001 and Franklin and Mackie 1984).

[2]Bremer (1992) was highly influential in making the dyad the most prominent unit of analysis in IR, though the practice of disaggregating multilateral events into dyadic observations is too widespread to summarize. One need only to pickup a statistical study in international relations over the past few decades to find an example (assuming the study did not focus solely on bilateral relations). Prominent examples that apply dyadic data to multilateral wars include Bremer (1992), Russett and Oneal (1997), Peceny, Beer, and Sanchez-Terry (2002), and Reiter and Stam (2002, 2003). Studies that have applied dyadic data to the creation of multilateral trade agreements include Mansfield, Milner and Rosendorff (2002) and Mansfield and Reinhardt (2003). Lai and Reiter (2000) and Leeds et al. (2002) apply dyadic data to the creation of multilateral alliances. See Bennett and Stam (2000) for an excellent discussion of the promise and pitfalls of estimating dyadic data in international relations.

[3]See Gent (2007) and Findley and Teo (2006) for examples of modeling strategic interactions in multilateral events.

thereby inflating the number of positive cases. Signorino (1999) also identifies this problem, pointing out that dividing a k-nation event into a series of dyadic observations of size $k(k-1)/2$ greatly expands the size of the data set but does so without adding new information and by introducing bias. However, none of these studies, nor any previous study to which the author is aware, has sought either to identify the size of the bias introduced by evaluating k-adic events with dyadic data or offer a suggestion for how one should alternatively structure the data. Instead, scholars continue to divide multilateral events into a series of dyadic observations.

This paper has two goals. The first is to illustrate the bias produced when analyzing k-adic processes with dyadic data. I show, using a Monte Carlo simulation under the simplest of conditions (a cross-sectional data set in which each grouping of countries has an independent ability to form an alliance and the decision-making process is nonstrategic) that one cannot recover a k-adic data generating process (DGP) using *dy*-adic data. One must instead evaluate the DGP using k-adic data. In other words, one must use a data set containing all combinations of actors (i.e., actors A, B, and C can form four multiactor combinations: AB, AC, BC, and ABC). Of course, if the number of potential actors is even moderately large (perhaps $N = 100$), this can quickly produce a data set with observations numbering in the millions, billions, or more.[4]

Thus, the second goal of the paper is to illustrate how choice-based sampling, an approach recommended by King and Zeng (2001a, 2001b) for analyzing "rare events" data, enables one to create and analyze k-adic data sets of manageable size. Specifically, one can use a data set consisting of all k-ads in which a binary dependent variable is coded 1 (indicating, e.g., if members of the k-ad formed an alliance treaty or began a conflict) and a random sample of various sized k-ads in which the binary dependent variable is coded 0. This solution is not without costs. For instance, some measures, such as distance, are most easily understood in a dyadic context. However, intuitive tractability and data collection simplicity do not justify continued reliance upon flawed inferences.

This paper is organized as follows. First, using Monte Carlo simulations, Section 2 illustrates the bias introduced when dyadic data are used to evaluate k-adic events. Though I place this simulation (and the subsequent simulations) in the context of alliance formation, this is intended simply to give the simulation a point of reference (the main statistical points they raise could be illustrated just as easily with randomly constructed covariates devoid of any substantive motivation). Conflict onset, the formation of international trade agreements, governing party coalitions, as well as numerous other subjects could be used to contextualize the simulation. Second, Section 3 uses Monte Carlo simulations to illustrate how choice-based sampling can generate a feasibly sized data set that, when estimated, produces substantially less bias. Section 4 uses the study of alliance formation by Gibler and Wolford (2006) to illustrate how one may apply choice-based sampling to the construction of k-adic data. Section 5 offers a discussion of why some alternative methods, particularly spatial interdependence models and evolving network models, are not viable substitutes for using k-adic data to evaluate multilateral events. Section 6 concludes.

2 Illustrating the Problem with Monte Carlo Simulations

In this section, I use Monte Carlo simulations to illustrate how dyadic data cannot capture the process that produces data formed by interactions among $k > 2$ actors. Again, this is an

[4]This is a problem that the governing coalition papers of Martin and Stevenson (2001) and Franklin and Mackie (1984) did not have to address as the number of potential actors in their studies were relatively small (the largest being $N = 10$ for a few countries where up to 10 parties existed at the time of a government formation).

issue of how one conceptualizes the construction of the observations and, hence, is separate from concerns of modeling spatial, temporal, strategic, or monadic interdependencies among the observations. For the sake of illustration, I place this simulation (and the subsequent simulations) in the context of alliance formation. This is intended only to give the simulation a substantive point of reference.

2.1 *Motivating the Simulation*

According to Morrow (1991), pairings of states with highly asymmetric relative physical capabilities are natural alliance partners. In essence, alliances serve as a type of ''protection racket'' where a small state gives foreign policy autonomy to a larger state (in the form of policy concessions or the granting of territorial access to the large state's military forces) in exchange for the large state's promise to defend it in a time of crisis. One can easily extend to multilateral agreements the Morrow story of asymmetry influencing alliance formation. The game theoretic work on *N*-player prisoner dilemmas (see Bianco and Bates 1990) and its extension to international cooperation (see, most recently, Stone, Slantchev, and Tamar 2008) view the presence of a large state as the key factor in creating multilateral agreements. This is because the large state can impose punishments on states that fail to meet contribution requirements.

This suggests that the *capability ratio* of a grouping of states is a (if not ''the'') major factor in determining if the states will form an alliance. Though the exact influence of the largest state's capabilities relative to the entire group's capabilities is not known, we do know that, in theory, the larger this ratio, the more likely is a multilateral agreement. For the sake of simplicity and to avoid the issue of the improper use of control variables that is rampant throughout the empirical international relations literature,[5] I will assume that the true DGP for alliance formation can be specified as:

$$\begin{aligned}&\Pr(\text{Alliance between states A through K})\\&\quad=\Phi\left(\text{cons}+\beta\frac{\max(\text{capability A, capability B},\ldots,\text{capability K})}{(\text{capability A}+\text{capability B}+\ldots+\text{capability K})}+\mu\right),\end{aligned} \tag{1}$$

where ''cons'' is a constant term, μ is a random element capturing the unknown and/or unobserved determinants of alliance formation, and Φ is a function taking on values strictly between zero and one ($0 \leqslant \Phi \leqslant 1$). β is the true parameter specifying the relationship between the value of a latent, unobserved dependent variable that determines the probability of alliance formation, y^*, and the capability ratio of states A through K.

2.2 *Describing the Trilateral Alliance Simulation*

I consider a scenario in which states can only form the most basic of multilateral alliances, *trilateral* alliances. I construct the simulation according to the following steps:

Step 1: I create a data set of 100 observations, where each observation represents a single country. I then assign a country code (ccode) value to each country.

Step 2: I randomly assign military ''capabilities'' to each of these countries. Capabilities range from 0 to 100. These capabilities are stored in the variable *cap*.

[5]For works that detail the improper use of control variables in international relations, see Kadera and Mitchell (2005), Ray (2005), Achen (2005), Clarke (2005), and Starr (2005).

Step 3: I reorganize the 100 countries into all possible three-country groupings. Since order is not important, these 100 countries produce 161,700 three-country combinations of states (or triads). The *triadic* data set includes the following variables: *triadid*, *mem*1, *mem*2, *mem*3, *cap*1, *cap*2, and *cap*3. The variable *triadid* is simply a code identifying triad i (with $i \in \{1, 161{,}700\}$). The variable *mem*1 is the ccode number of the first member state in triad i, *mem*2 is the ccode number of the second state in triad i, and *mem*3 is the ccode number of the third state in triad i. The variables *cap*1, *cap*2, and *cap*3 capture the capabilities of *mem*1, *mem*2, and *mem*3, respectively.

Step 4: I compute the "capabilities ratio" of each triad. Specifically, this is captured by the variable "*capratio*" that is calculated as:

$$\text{capratio} = \frac{\max(cap1, cap2, cap3)}{(cap1 + cap2 + cap3)}, \tag{2}$$

Step 5: I write the DGP of trilateral alliance formation as:

$$xb = \text{cons} + \beta\,\text{capratio} + \mu, \tag{3}$$

where xb represents the underlying latent variables that determine alliance formation. I set cons $= -4$ and $\beta = .25$. The variable μ is a logistically distributed random error term.

Step 6: *ALLY*, the dependent variable, is a dichotomous variable equal to 1 if a triad forms an alliance, zero otherwise. To generate realizations of this dependent variable, I code *ALLY* = 1 if $xb > 0$, zero otherwise. Table 1 reports the values of these variables for the first 10 observations. One should notice that *ALLY* = 0 for each of these 10 observations.

Step 7: I now convert this triadic data set into dyadic data. Thus, if a triad contains states A, B, and C, this step divides this triad into dyad A with B, dyad B with C, and dyad A with C. If *ALLY* = 1 for triad A, B, and C, then this means *ALLY* = 1 for dyad A with B, *ALLY* = 1 for dyad B with C, and *ALLY* = 1 for dyad A with C. Next, I use the capabilities scores of each dyad member to compute that dyad's capability ratio.

Step 8: I take this dyadic data set and attempt to estimate β, the parameter characterizing the relationship between *capratio* and *ALLY*. Since the errors are drawn from a logistic distribution, I use logit estimation. The goal is to see if the logit estimate of β, $\hat{\beta}$, is an unbiased estimate of the true β (which is equal to .25).

These eight steps create one realization of my data set. Of course, this realization is determined by a single random draw of u from a logistic distribution. Because the dependent variable is computed using an error term drawn from a probability distribution, I repeat the creation of the dependent variable via a Monte Carlo simulation. In Monte Carlo simulations, random numbers are drawn so as to model a process. The goal is to determine how random variation (or lack of knowledge or error) affects the sensitivity and reliability of the parameters characterizing the process. In this particular simulation, I wish to know how randomness impacts my ability to estimate the impact of relative capabilities on the formation of alliances. The essence of Monte Carlo simulations is to iterate the process numerous times and then obtain an average value from these iterations.

Step 9: I repeat 500 times steps 6 through 8. This produces 500 values of $\hat{\beta}$. After each iteration, I also keep the estimated SE around $\hat{\beta}$ (giving me 500 values of the SE of $\hat{\beta}$).

Table 1 Sample of complete triadic data set

Dyad	*mem1*	*mem2*	*mem3*	*cap1*	*cap2*	*cap3*	*capratio*	*ALLY*
102	1	2	3	70.01	62.24	26.23	0.44	0
102	1	2	4	70.01	62.24	16.30	0.47	0
102	1	2	5	70.01	62.24	51.48	0.38	0
102	1	2	6	70.01	62.24	85.39	0.39	0
102	1	2	7	70.01	62.24	35.41	0.42	0
102	1	2	8	70.01	62.24	24.88	0.45	0
102	1	2	9	70.01	62.24	24.29	0.45	0
102	1	2	10	70.01	62.24	34.11	0.42	0
102	1	2	11	70.01	62.24	66.35	0.35	0
102	1	2	12	70.01	62.24	5.38	0.51	0

2.3 *Results from the Trilateral Alliance Simulation*

I use the stored values of $\hat{\beta}$ and the estimated SEs to compute three common criteria for evaluating estimator performance in Monte Carlo simulations: Bias, Root Mean Squared Error, and Overconfidence. Bias is the difference between the average value of the coefficient estimate and the true coefficient value. Root Mean Squared Error is calculated in three steps: (1) computing the squared difference between each iteration's coefficient estimate and the true coefficient value; (2) summing up these values and dividing the total by the number of iterations; and (3) taking the square root of this average value. Overconfidence is the standard deviation (SD) of the coefficient estimates divided by the average reported coefficient SE. This is a measure of SE accuracy. For all three measures, the smaller the value, the more accurate the estimator.

Column 1 of Table 2 shows how estimating the triadic data set with triadic data produces, as one would expect, relatively unbiased coefficient estimates. However, column 2 shows quite convincingly that the $\hat{\beta}$ produced using the dyadic data set does not accurately estimate β. These results suggest that the existing approach of dividing multilateral alliances into a series of dyadic observations produces biased estimates of the true parameter. This is unsurprising as one should not reasonably expect a dyadic measure of capability asymmetry to be equivalent to the triadic measure of capability asymmetry employed in this simulation's actual DGP. Unfortunately, this is exactly the technique employed by scholars of international relations (to divide multilateral events, in this case the formation of alliances, into a series of bilateral observations).

Table 2 Trilateral alliance simulation results (true $\beta_1 = 0.25$)

	(1) Triadic DGP estimated with triadic data	*(2) Triadic DGP estimated with dyadic data*	*(3) Triadic DGP estimated with choice-based sample*
Average $\hat{\beta}_1$	0.251	0.46	0.251
Bias	0.001	0.21	0.001
Root Mean Squared Error	0.13	0.66	0.16
Overconfidence	0.17	1.09	0.24

2.4 *Indentifying the Source of the Bias*

What is the exact cause of the bias in the multilateral simulation? Setting aside problems of nonindependence of observations (which, as already highlighted, is present in dyadic data as well), dividing a k-adic event into dyadic observations leads to a classic case of measurement error in $\mathbf{X}$, the vector of independent variable values. Recall that the capability ratio in the true DGP is

$$\text{capratio} = \frac{\max(cap1, cap2, cap3)}{(cap1 + cap2 + cap3)}, \tag{4}$$

whereas the independent variable in the estimated model is the dyadic capability ratio

$$\text{capratio} = \frac{\max(cap1, cap2)}{(cap1 + cap2)}. \tag{5}$$

Suppose that the third member of the alliance is never the largest member. Therefore, the dyadic data will always have the correct numerator value (either *cap*1 or *cap*2). However, even in this ideal scenario, the estimated independent variable is systematically higher because the denominator is missing one term, *cap*3. As a result, the observation $\mathbf{X}_i$ for $i = 1, \ldots, n$ is actually $\mathbf{X}_i = \mathbf{W}_i + \mathbf{U}_i$, where $U_1, \ldots, U_n$ are uniformly distributed (because a uniform distribution was used to generate the capability scores of each state) with $E[\mathbf{U}] \geqslant 0$.[6] Thus, for either the probit or logit model, we obtain (ignoring the constant term)

$$\begin{aligned}
\Pr(Y_{T_i} = 1) &= \Pr\left(Y^*_{T_i} \geqslant 0\right) \\
&= \Pr(\beta \mathbf{X}_i + \epsilon_i \geqslant 0) \\
&= \Pr(\beta(\mathbf{W}_i + \mathbf{U}_i) + \epsilon_i \geqslant 0) \\
&= \Pr(\epsilon_i \geqslant -\beta(\mathbf{W}_i + \mathbf{U}_i)) \\
&= 1 - F(-\beta(\mathbf{W}_i + \mathbf{U}_i)) \\
&= F(\beta(\mathbf{W}_i + \mathbf{U}_i))
\end{aligned},$$

where the last line is possible if F is symmetric (which is the case for both the logistic and normal distributions). Of course, the parameter will be estimated via maximum likelihood, where each outcome of Y_{T_i} follows a Bernoulli density function, $f\left(Y_{T_i}\right) = p_i^{Y_{T_i}}(1-p_i)^{1-Y_{T_i}}$. In other words, each Y_{T_i} takes on either a value of 0 or 1 with probability $f(0) = (1 - p_i)$ and $f(1) = p_i$. Hence, $p_i = \Pr(Y_{T_i} = 1) = F(\beta(\mathbf{W}_i + \mathbf{U}_i))$. Thus, the likelihood function is $L = f(Y_{T_1}, Y_{T_2}, \cdots, Y_{T_n})$. Even if, for $i \neq j$, each Y_{T_i} were independent of each Y_{T_j} (which is not the case since multiple dyads contain the same country) so that the log likelihood is $\ln L = \sum_{i=1}^{n} Y_{T_i} \ln p_i + (1 - Y_{T_i}) \ln(1 - p_i)$, it would still be the case that $p_i = F(\beta(\mathbf{W}_i + \mathbf{U}_i))$. Consequently,

$$\ln L = \sum_{i=1}^{n} Y_{T_i} \ln F(\beta(\mathbf{W}_i + \mathbf{U}_i)) + (1 - Y_{T_i}) \ln(1 - F(\beta(\mathbf{W}_i + \mathbf{U}_i))). \tag{6}$$

One will note that the presence of $\mathbf{U}_i$, where $E[\mathbf{U}_i] \geqslant 0$, will inflate the value of $\mathbf{X}_i$. Because the values of Y_{T_i} are fixed, the larger value of $\mathbf{X}$ must necessarily reduce $\hat{\beta}$, the estimate of the true β. Thus, $\hat{\beta} \neq \beta$.

[6]It should be noted that $U_1, \ldots, U_n$ are not independent as some observations contain the same third country.

3 Modeling *k*-adic Data Using Choice-Based Sampling

Scholars have relied upon dyadic data to analyze international events because it provides intuitive tractability, is computationally convenient, and simplifies the collection of data. Moreover, scholars have made great strides in devising estimation adjustments that account for temporal, spatial, monadic, and strategic violations of the independence assumption present in many dyadic data sets. However, none of these adjustments, nor any estimation correction, can account for the bias produced by evaluating *k*-adic events with dyadic data.

The simulation in the previous section shows that estimating the formation of trilateral alliances with triadic data will produce unbiased estimates of the parameter on *capratio*. Therefore, the solution seems obvious: use a data set with all possible *k*-ads. Fortunately, by all "possible" *k*-ads, I do not mean to suggest that if there are *n* countries, then one needs to include all *k*-ads of size *n* or less. Instead, I will show that if the *k*-adic event of interest contains, at most, $k < n$ countries, then one need only to estimate a data set with all combinations of states up to size *k*. Unfortunately, creating a data set of all combinations of size $k < n$ has a major downside: it still dramatically expands the data set's observations. For a system of 100 countries, just as in the above simulations, a dyad-only data set contains 4950 observations, but a triad-only data set contains 161,700 observations. If one were to consider a data set of only four country *k*-ads, the data set size would explode to 3,921,225 observations. Consequently, it would be infeasible or impossible to estimate a data set capable of explaining the creation of an alliance the size of North Atlantic Treaty Organization (NATO), which was formed by 12 countries (where all combinations of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 out of 100 countries will lead to over $1.211475 \times 10^{+15}$ observations)!

3.1 *Choice-Based Sample of Triadic Data*

Choice-based sampling on the dependent variable (see King and Zeng 2001a, 2001b) offers a means of creating a computationally manageable data set appropriate for estimating *k*-adic data. Because so few triadic observations, relative to the total number of triads, contain the formation of a military alliance, one is left with a classic "rare events" data set (binary dependent variable characterized by dozens to thousands of times fewer events [coded with a positive value] than nonevents [coded with a 0]). When presented with data of this type, King and Zeng (2001a, 2001b) recommend sampling on the dependent variable as it avoids the issues commonly associated with rare events data such as underestimating the probability of an event. The sampling method entails constructing a data set containing all observations for which the dependent variable is coded with a positive value, along with a random sample of observations for which the dependent variable is coded 0. According to King and Zeng (2001a, 702), it is acceptable to collect anywhere from two to five times more 0s than positive values, though one should attempt to collect as many zero values as is computationally feasible. Thus, in the simulations that follow, I collect 10 times more 0s than 1s. Even then, this creates a data set between two- and three-thousand observations, which is highly manageable from a computational perspective.

When drawing the sample, it is important to stratify by *k*-ad. This means if a data set with a binary dependent variable has 100 dyads where $Y = 1$ and 50 triads where $Y = 1$, one should attempt to draw 1000 dyads where $Y = 0$ and 500 triads where $Y = 0$. Why should one sample in this manner rather than simply draw a nonstratified random sample from the full population of possible outcomes? The reason is that higher order *k*-ads quickly overwhelm lower order *k*-ads with respect to quantity. For instance, if a data set contains all five-ads, quad-ads, tri-ads, and dyads of 100 actors, then the full population contains 75,287,520 five-ads and only 4950 dyads. A nonstratified random sample of the $Y = 0$ k-ads would contain virtually no $Y = 0$ dyads, even if dyads comprise the majority of $Y = 1$ observations!

Having obtained a stratified choice-based random sample of k-ads where the dependent variable equals zero, this sample can be combined with the k-ads where the dependent variable equals 1. This combined data set can then be estimated using a *rare events logit model*, which, by and large, is a logit model that applies a postestimation correction to the constant term (called prior correction) to account for the fact that sampling on the dependent variable has artificially inflated the prominence of observations where the dependent variable equals 1. Since the data set on which the model is estimated is a stratified sample, one must weight the observations from each strata by the inverse probability of being drawn from the sample. For example, if there are 4950 total dyads, and $Y = 1$ for 100 of these dyads, then each $Y = 0$ dyad has a $\frac{1}{4850}$ probability of being drawn. When estimating the model, each $Y = 0$ observation in the sample should be multiplied by $\frac{1}{\frac{1}{4850}}$.

After applying choice-based sampling, I have a cross-sectional data set of approximately 55,000 triad observations (where approximately 5000 of which *ALLY* equals 1). The column 3 of Table 2 presents the results obtained from estimating a choice-based sample of the triadic data set using a King and Zeng (2001a, 2001b) rare events logit model. Comparing this parameter estimate to that obtained from estimating the full triadic data set, one can see that the parameter estimates are nearly identical, with the Overconfidence measure suggesting that the rare events estimate produces slightly more variance (which is expected, given that it contains fewer observations).

3.2 *Accounting for All k-ads with Choice-Based Sampling: Proof of Concept*

Though this solution works for the triadic data set, what about a data set in which the largest k-lateral alliance contains four countries or a data set containing k-lateral alliances of multiple sizes? The latter is of particular importance since this is the shape of actual data sets in international relations. For example, the Alliance Treaty Obligations and Provisions (ATOP) data set contains 648 military alliance treaties formed between 1815 and 2005. Of these, 536 are bilateral alliances, 47 are trilateral alliances, 23 are quadrilateral alliances, 11 alliances have five members, and 38 have 6 or more members (with the largest alliance containing 50 members).[7] As mentioned above, estimating a data set with all possible combinations of all possible alliance sizes is computationally infeasible. For instance, a data set with all possible dyadic, triadic, and quadratic combinations of 100 countries contains 4950 + 161,700 + 3,921,225 = 4,087,775 observations! However, one can still sample on the dependent variable in order to obtain parameter estimates for such data. To show this is the case, this subsection provides a "proof of concept" focusing on a simulated data set containing bilateral and trilateral alliances.

3.2.1 Bilateral-trilateral simulation

For this simulation, I place the 100 countries into all possible dyadic and triadic combinations. This generates a combined data set of 166,550 observations (where an observation is any grouping of states, dyadic or triadic). The variable *capratio* is the ratio of the capabilities of a k-ad's largest member over that k-ad's total capabilities. Next, I set β_1, the parameter on *capratio*, to 25 and the constant term is set to -25. With these parameter

[7]See Leeds et al. (2002) for further details on the ATOP data set.

Table 3 Bilateral-trilateral alliance simulation results (true $\beta_1 = 25$)

	(1) Triadic-dyadic DGP estimated with triadic-dyadic data	*(2) Triadic-dyadic DGP estimated with dyadic data*	*(3) Triadic-dyadic DGP estimated with trilateral alliances removed*	*(4) Triadic-dyadic DGP estimated with choice-based sample*
Average $\hat{\beta}_1$	25.03	22.54	25.18	25.21
Bias	0.03	−2.46	0.18	0.21
Root Mean Squared Error	1.39	4.05	2.17	2.07
Overconfidence	2.35	8.27	3.17	3.32

values, a typical simulation produces approximately equal numbers of bilateral and trilateral alliances (typically 120–130 each). As in the above simulations, the DGP also includes a logistically distributed error term.

Table 3 reports the results from 500 Monte Carlo simulations of this DGP. One can see from column 1 that applying logit estimation to the full triad-dyad data set produces, on average, parameter estimates close to the true parameter value. Column 2 of Table 3 reports the average parameter estimates from 500 Monte Carlo simulations where the data are converted back into dyadic data. As in the above simulation with triadic data, this produces large bias in the parameter estimate.

In column 3 of Table 3, I present the results from a "quick fix" one might be tempted to apply when faced with a multilateral event: dropping the multilateral events. Given that dyads, by definition, capture bilateral relations, some readers may decide a simpler solution for obtaining unbiased estimates lies in simply excluding multilateral events from the data. Therefore, I rerun the simulation with the triads removed from the data set prior to estimation.[8] This does produce dramatically less bias in the estimates and, therefore, is not an unreasonable approach. However, if one wishes to model multilateral observations, dropping the multilateral cases is obviously not an option. Moreover, as the next simulation will show, the bias produced by this approach increases with the number and variety of *k*-ads in the data set.

Column 4 of Table 3 reports the results from estimation with choice-based sampling on the dependent variable. The bias is substantially reduced compared to converting the data into dyadic data. Therefore, choice-based sampling appears to offer a viable and computationally feasible means of obtaining relatively unbiased parameter estimates of actual *k*-adic processes.

3.2.2 A FIVE-adic simulation

To more fully illustrate the ability of choice-based sampling to create a feasibly sized *k*-adic data set that reduces estimation bias, I consider a final simulation in which the maximum size of a *k*-ad is five countries. I focus on a FIVE-adic data set because, as the above description of the ATOP data set illustrates, there are very few *k*-adic alliances with more than five participants.

[8]This approach is adopted in some studies, such as Remmer (1998), with the explicit desire to avoid placing multilateral events into dyadic data (Remmer 1998, 35).

Table 4 FIVE-adic simulation results (true $\beta_1 = 26$)

	(1) FIVE-adic DGP estimated with choice-based sample	*(2) FIVE-adic DGP estimated with dyadic data*	*(3) FIVE-adic DGP estimated with non-dyads removed*
Average $\hat{\beta}_1$	28.3	13.95	34.31
Bias	2.3	−12.05	8.31
Root Mean Squared Error	0.01	1.34	7.73
Overconfidence	0.05	2.53	18.72

For this simulation, I place the 100 countries into all possible combinations of 2, 3, 4, and 5 countries. This generates a combined data set of 79,375,395 observations (4950 + 161,700 + 3,921,225 + 75,287,520). The variable *capratio* is defined the same as in the earlier simulations, and the DGP is the same as in the above simulations, except *capratio* is computed with the capabilities of two, three, four, or five states, depending on whether an observation is a dyad, triad, "4"-ad, or "5"-ad.

Next, I set β_1, the parameter on *capratio*, to 26 and the constant term is set to −27. With these parameter values, a typical simulation produces about 190 bilateral alliances, 180 trilateral alliances, 110 quadrilateral alliances, and 70 five-member alliances. Table 4 reports the results from estimating this data set when applying choice-based sampling to the full data set (column 1), dividing the data into dyadic combinations (column 2), or dropping the *k*-adic observations in which an alliance formed (column 3). One can immediately see that the bias produced by analyzing the FIVE-adic DGP with dyadic data is dramatically more pronounced than in the previous simulation. Moreover, estimation with the choice-based sample outperforms both estimation with a data set in which all *k*-adic alliances are split into their dyadic combinations and estimation with a data set that simply dropped the *k*-adic observations in which an alliance formed.

4 Application: Alliance Formation in International Relations

I will now illustrate how one can apply choice-based sampling to actual data that follows a *k*-adic DGP. I will do so using the study of alliance formation by Gibler and Wolford (2006), who draw on the dyad-year research design of Lai and Reiter (2000). This is one of the only studies to conduct multivariate estimation of alliance formation. Their dependent variable, *alliance formation*, equals 1 the year two states become alliance members, 0 otherwise. Because this study is especially interested in the relationship between regime type and alliance formation, Gibler and Wolford (2006) mention how dyadic data could overstate the role of democracy on alliance formation. In particular, Gibler and Wolford (2006, 139) highlight how the bulk of democratic dyads that formed alliances are contained in an incredibly small number of alliances. For example, NATO accounts for more than 55% of the jointly democratic allied dyad-years. Nevertheless, Gibler and Wolford (2006), in order to match as closely as possible previous research designs, test their model using all dyad-year data drawn from all alliances.

I am not setting out to nullify the results of Gibler and Wolford (2006). Instead, I am using their study because, in addition to alliance formation exemplifying a *k*-adic DGP that has typically been tested using dyadic data, their study is useful for illustrating new measurements of covariates that were previously coded only in the dyadic context. For example, *geographic distance* is easy to conceptualize for a dyad, but what does it mean in a *k*-adic data set? Is it the maximum distance between any two of k members or is it the average distance between the k members? Similarly, what is a joint democracy k-ad? If a k-ad contains five states, is it a joint democracy k-ad only when all five states are democracies? If so, does that not treat a k-ad where four of the five states are democracies as equivalent to a k-ad where one of five states are democracies? Alternatively, perhaps one could construct a continuous measure of joint democracy such as the proportion of states in the k-ad that are democracies. My goal is not to rectify such measurement issues as answers will depend on the particular research question. Instead, by illustrating how one can properly construct and test a data set for an event that is inherently k-adic, I will propose and apply reasonable codings for such variables.

4.1 *Real Alliance Formation and Capabilities Data*

Before more fully applying the Gibler and Wolford (2006) model of alliance formation, I begin with a simple model that closely follows the above simulations. Specifically, I test a single covariate model where the dependent variable is the formation of a Correlates of War military alliance (similar to Gibler and Wolford 2006) and the independent variable is the capability ratio (the capabilities of the largest state over the sum of the k-ad's capabilities), where capabilities is measured using the Correlates of War composite index of national capabilities score. The dependent variable *ally* equals 1 the year an alliance forms, 0 otherwise. Because this replication focuses on the decision to form a new alliance, I consider states that join an alliance after the year of its initial formation as having not joined the alliance (the decision to join an existing alliance is a worthy research question but is treated here as distinct from the decision to create a new alliance).

Column 1 of Table 5 reports the results from applying logit estimation with clustered SEs to a dyadic data set where $ally = 1$ if any alliance is formed (bilateral or multilateral alliance). Column 2 of Table 5 reports the results from applying logit estimation with clustered SEs to a dyadic data set where $ally = 1$ only when bilateral alliances are formed. Column 3 of Table 5 reports the results when using a rare events logit to estimate a choice-based data set that directly measures all k-ads that formed alliances.

It is important to make two notes regarding the results in column 3. First, for $ally = 0$ observations, I use eight times the number of $ally = 1$ observations. Second, the data set used to produce the results does not include alliances with six or more members. This is for two reasons. First, the data set only contains four alliances of such size. Hence, if alliance formation is a "rare event," then the formation of alliances with six or more members is an "unusual" event. Second, the set of possible $ally = 0$ k-ads of six or more members is simply enormous. For example, with 196 countries from which to choose (the number of countries in the Gibler and Wolford [2006] data set), all combinations with 6 countries lead to 72,887,293,024 observations. Since there is only a single six-member alliance, I would draw only eight 6-ads where $ally = 0$. This is problematic because when using a stratified choice-based sample, one must weight each observation by the inverse probability of that observation being drawn from its stratum. In the case of 6-ads where $ally = 0$ (where the $ally = 0$, 6-ad stratum contains $72, 887, 293, 024 - 1$ observations), this produces a probability of $\frac{8}{72{,}887{,}293{,}023} = 0.0000000011$ or an inverse probability weight of approximately 9,111,000,000. Placing such a massive weight on a single observation renders the SEs of the point estimates uninformative. Since some data sets will have more than 196 actors and some will have less, the decision of what constitutes an "unusual" versus simply a "rare" k-adic event must be left to the analyst.

Table 5 *K*-ad year alliance formation regressed on capability ratio

Data set	*(1) Dyadic*	*(2) Remove k-adic alliance*	*(3) k-adic choice-based sample*
Capability ratio	−1.79*** (0.157)	−1.08** (0.39)	7.21** (3.43)
Constant	−3.69*** (0.132)	−6.13*** (0.33)	−49.81*** (2.69)
N	570,390	570,390	215
Estimation technique	Logit with SEs clustered on the dyad	Logit with SEs clustered on the dyad	Rare events logit

Note. SEs reported in parentheses (nonclustered SEs produce similar results for models 1 and 2).
** indicates statistical significance at the 0.05 level.
*** indicates statistical significance at the 0.01 level.

Column 1 of Table 5 shows that splitting all *k*-ads into dyadic relations leads to a negative and statistically significant value on the coefficient for capability ratio. In contrast, the coefficient is positive when one accounts for all *k*-adic combinations.

4.2 *Reconsidering Gibler and Wolford (2006)*

Gibler and Wolford (2006), drawing upon Lai and Reiter (2000), model alliance formation as a function of several variables. For my illustrative application, I will only use a subset of their variables: *Common Threat*, *Geographic Distance*, and *Joint Democracy*. I will include with these variables the capability ratio of each *k*-ad. These variables are chosen because they are (1) consistently found to be important determinants of alliance formation and (2) are examples of variables problematic to code in *k*-adic data.

In dyadic data, *Common Threat* is a dichotomous variable coded 1 if both states participated in a Militarized Interstate Dispute (MID) against the same third state sometime in the previous 10 years, 0 otherwise. Coding this variable in the *k*-adic context creates similar difficulties to coding *Joint Democracy*: if a *k*-ad contains five states, does it not face a joint threat if only four of the five states participated in a MID against the same third state? Given that the variable *Common Threat* is intended to capture the idea that a group of states will have a strong incentive to form an alliance when *all* members of that group face the same threat, I will adopt the following coding rule: *Common Threat* equals 1 if each state has participated in a MID against the same third state sometime in the previous 10 years, 0 otherwise.

In dyadic data, *Geographic Distance* gives the square root of the capitol-to-capitol distance, unless states are contiguous, in which case distance is set to 0. The potential complications with coding *Geographic Distance* in *k*-adic data were discussed above. I will code *Geographic Distance* in *k*-adic data by applying the "weakest link" principle of Oneal and Russett (1997).[9] This means I will represent the geographic distance of the entire *k*-ad using the geographic distance of the most distant pair of states.

In dyadic data, *Joint Democracy* is a dichotomous variable coded 1 when both members of the dyad are democracies, 0 otherwise. As mentioned above, this coding rule is a bit problematic when applied to *k*-adic data: is a FIVE-ad where four of five states are

[9]Oneal and Russett (1997) use as a measure for the entire dyad the minimum state level value for the dyad. For example, they measure a dyad's overall level of trade integration by using the lower of the two state levels of trade integration (if state A has trade integration of 40% and state B has trade integration of 30%, then the trade integration for the dyad is 30%).

Table 6 Descriptive statistics

	Observations	*Mean*	*SD*	*Minimum*	*Maximum*
k-adic data					
Distance	287	66.51	27.87	0	109.29
Joint Threat	299	0.04	0.16	0	1
Joint Democracy (continuous)	295	0.29	0.33	0	1
Joint Democracy (dummy)	299	0.12	0.32	0	1
Capability Ratio	296	0.73	0.21	0.18	0.99
Dyadic data					
Distance	515,753	63.57	24.38	0	111.33
Joint Threat	516,914	0.05	0.22	0	1
Joint Democracy	411,476	0.10	0.30	0	1
Capability Ratio	570,390	0.83	0.15	0.5	0.99

democracies equivalent to a FIVE-ad where one of the five states is a democracy? Above, I suggested using a continuous measure of democracy, such as the proportion of states in the *k*-ad that are democracies. Therefore, I code *Joint Democracy* using two approaches: as the proportion of states in a *k*-ad that are democracies and as a dichotomous variable coded 1 when all members of the *k*-ad are democracies, 0 otherwise. This will allow me to compare how the results change when using an alternative coding rule. Descriptive statistics for these variables are reported in Table 6, along with a comparison to the typical dyadic values of these variables.

The results are reported in Table 7. Comparing column 3 to column 4, one can see that using a continuous or dichotomous measure for *Joint Democracy* does not drastically alter the coefficient estimates. Comparing column 1 of Table 7 (splitting all *k*-ads into dyadic relations) to column 3 of Table 7 (estimation with choice-based sampling of *k*-ads) reveals two major changes in the results. First, the sign on the variable for capability ratio flips from negative to positive. Although the model estimated with dyadic data identifies capability ratio as having a significant and negative effect on the probability of alliance formation, the model estimated with *k*-adic data finds that the effect is positive and insignificant. Second, the coefficients on the remaining variables are dramatically larger in the *k*-adic model. To illustrate the substantive impact of these larger coefficients, consider a change in the relative risk associated with going from having no common threat (*Common Threat* = 0 in both the dyadic and *k*-adic models) to having a common threat (*Common Threat* = 1 in both the dyadic and *k*-adic models).[10] Estimating this model with dyadic data shows that such a change increases the risk of forming an alliance by 5.02 times. However, estimating this model with k-adic data shows that such a change increases the risk of forming an alliance by 420 times.[11]

Upon seeing these results, some scholars may wonder if a simpler approach for modeling a *k*-adic process without bias would be to simply incorporate into dyadic data a dummy variable that accounts for the *k*-adic concept. For instance, consider again the Belgium-Turkey example that opened the paper. Given that the presence of the

[10]Bennett and Stam (2007) suggest using the risk ratio to substantively evaluate logit coefficients as the rare occurrence of many international events render their predicted probabilities to be exceedingly small (Bennett and Stam 2007, 67–9).

[11]The probabilities and relative risk ratios are computed using prior correction by applying the *relogitq* command in STATA. Replication do files are available upon request.

Table 7 *k*-ad year alliance formation with Gibler and Wolford (2006) data

Data set	*(1) Dyadic*	*(2) Remove k-adic alliance*	*(3) k-adic choice-based sample*	*(4) k-adic choice-based sample*
Distance	−0.042*** (0.001)	−0.055*** (0.004)	−0.10*** (0.02)	−0.11*** (0.03)
Common Threat	1.615*** (0.079)	1.341*** (0.169)	6.02*** (0.95)	5.42*** (0.97)
Joint Democracy (continuous)	−0.406*** (0.136)	−0.439 (0.293)	−2.78** (1.16)	
Joint Democracy (dichotomous)				−2.56* (1.38)
Capability Ratio	−0.576** (0.250)	0.358 (0.639)	1.33 (3.49)	0.63 (4.21)
Constant	−3.46*** (0.216)	−5.591*** (0.56)	−23.65** (3.69)	−22.68 (4.45)
N	411,476	411,476	202	203
Estimation technique	Logit with SEs clustered on the dyad	Logit with SEs clustered on the dyad	Rare events logit	Rare events logit

Note. SEs reported in parentheses (nonclustered SEs produce similar results for models 1 and 2).
* indicates statistical significance at the 0.10 level.
** indicates statistical significance at the 0.05 level.
*** indicates statistical significance at the 0.01 level.

United States induced both to join NATO, could one not simply add a variable for ‘‘Alliance Formation with the United States’’ or even ‘‘Alliance Formation with a Super Power’’?

Depending on the research question, such a reasonable ‘‘quick fix’’ may be appropriate (i.e., if the scholar is studying the influence of the United States in the formation of alliances). However, it is important to note that not all multilateral alliances include a major power. Additionally, a dummy variable does not capture the reason why the presence of a major power leads to the formation of an alliance. Is it because the major power poses a threat, offers security, or creates the ‘‘correct’’ balance in the capability ratio? This is not made clear by the simple inclusion of a dummy variable.

5 Alternative Approaches and Their Limitations

Though the inclusion of a dummy variable will not address the misconceptualization of multilateral events as fitting a dyadic DGP, scholars may still wish to model a *k*-adic process using dyadic data. Therefore, it is worth discussing some approaches that attempt to retain dyadic data and why these approaches, though quite useful in other contexts, are not yet suitable for modeling *k*-adic processes: bilinear mixed-effects hierarchical models, spatial interdependence regression models, and evolving network models.

5.1 *Bilinear Mixed-Effects Hierarchical Model*

Ward, Siverson, and Cao (2007) and Hoff and Ward (2004) use the bilinear mixed-effects model developed by Hoff (2005) to address monadic dependency in dyadic data. In essence, this model enables scholars to overcome a problem that is the mirror image of the issue I raise: standard approaches to analyze nondirected dyadic data (i.e., movement from state i to state j is considered the same as from j to i) hold that the dependence of observations having a common sender and the dependence of observations having a common receiver are both zero. This is a problem as it seems unreasonable to assume, for example, that all dyads containing the United States are independent from one another.

The bilinear mixed-effects model can account for this country-specific dependency by explicitly incorporating both dyadic and monadic (country specific) characteristics into the regression model.

Formally, suppose there is a binary outcome, $y_{i,j}$, which is either 0 or 1, indicating the presence or absence of a "link" from i to j.[12] Suppose we are interested only in estimating the linear relationships between responses $y_{i,j}$ and a vector of variables $\mathbf{x}_{i,j}$, which could include characteristics of unit i, characteristics of unit j, or characteristics specific to the pair. Thus, we can consider the regression model

$$y_{i,j} = \beta' \mathbf{x}_{i,j} + \varepsilon_{i,j}. \tag{7}$$

The generalized least squares estimate of $\hat{\beta}$ and its covariance matrix depend on the joint distribution of the $\varepsilon_{i,j}$'s only through their covariance. Next, two key assumptions are made. First, it is commonly assumed in regression problems that the regressors $\mathbf{x}_{i,j}$ contain enough information so that the distribution of the errors is invariant under any combination/arrangement of i and j. This is known as "weak row-and-column exchangeability" of an array. Second, it is assumed that $\varepsilon_{i,j}$ is Gaussian with mean 0. For undirected dyadic data (in which $y_{i,j} = y_{j,i}$), the first assumption implies that $\varepsilon_{i,j}$ is equal in distribution to $f(u, \alpha_i, \alpha_j, \gamma_{i,j})$, where u, α_i, α_j, and $\gamma_{i,j}$ are independent random variables and f is a function to be specified. When combined with the second assumption, we can now express $\varepsilon_{i,j}$ as

$$\varepsilon_{i,j} = \alpha_i + \alpha_i + \gamma_{i,j}, \tag{8}$$

where (α_i, α_j) are distributed multivariate normal with mean zero and variance $\Sigma_{\alpha_i,\alpha_j}$ and $(\gamma_{i,j}, \gamma_{j,i})$ are distributed multivariate normal with mean zero and variance $\Sigma_{\gamma_{i,j},\gamma_{j,i}}$. Because α_i, α_j, and $\gamma_{i,j}$ are independent random variables, then

$$\Sigma_{\alpha_i,\alpha_j} = \begin{pmatrix} \sigma^2_{\alpha_i} & 0 \\ 0 & \sigma^2_{\alpha_j} \end{pmatrix} \tag{9}$$

and

$$\Sigma_{\gamma_i,\gamma_j} = \begin{pmatrix} \sigma^2_{\gamma_{i,j}} & 0 \\ 0 & \sigma^2_{\gamma_{i,j}} \end{pmatrix}. \tag{10}$$

This means the covariance structure of the errors (and thus the observations) is

$$\begin{aligned} E\left(\varepsilon^2_{i,j}\right) &= \sigma^2_{\alpha_i} + \sigma^2_{\alpha_j} + \gamma^2_{i,j} \\ E\left(\varepsilon_{i,j}, \varepsilon_{j,i}\right) &= 0 \\ E\left(\varepsilon_{i,j}, \varepsilon_{i,k}\right) &= \sigma^2_{\alpha_i} \\ E\left(\varepsilon_{i,j}, \varepsilon_{k,j}\right) &= \sigma^2_{\alpha_j} \\ E\left(\varepsilon_{i,j}, \varepsilon_{k,i}\right) &= 0 \\ E\left(\varepsilon_{i,j}, \varepsilon_{k,l}\right) &= 0 \end{aligned} \quad , \tag{11}$$

[12]The formal discussion is adopted from Hoff (2005, 286–7).

so that $\sigma^2_{\alpha_i}$ represents the dependence of observations having a common sender and $\sigma^2_{\alpha_j}$ represents the dependence of observations having a common receiver. Standard approaches to analyzing dyadic data assume that both of these values are zero, but this is theoretically unlikely (i.e., it seems unreasonable to assume that all dyads containing the United States are independent from one another). Thus, the Ward, Siverson, and Cao (2007) approach provides a better way of analyzing dyadic data that accounts for the characteristics of the individual members of the dyad. This is useful when the dyad is the appropriate unit of observation and could be extended to account for individual-level characteristics in *k*-adic data, but it does not allow one to avoid using the *k*-ad as the unit of observation.

5.2 *Spatial Interdependence Regression Model*

The spatial interdependence specification of Franzese and Hays (2007a, 2007b, 2007c, 2008) is applied whenever an outcome in country *i* is influenced by the connection country *i* has with country *j*. For example, spatial interdependence regression models have been widely applied in political science to the study of interstate capital tax rate competition, pejoratively referred to as the "race to the bottom." Specifically, a common practice in statistical models of international capital tax competition is to control for the influence that tax rates of countries neighboring state *i* have on the tax rate of state *i*. Unfortunately, Franzese and Hays (2007b) point out how previous studies on capital tax competition, such as Hays (2003) and Basinger and Hallerberg (2004), improperly specify such spatial interdependence. For instance, Hays (2003) fails to weight the importance of the tax rate of one country relative to another.

Formally, if N is the number of countries and T is the number of time periods, then the connection between country i and country j is typically captured by $\mathbf{W}$, an $NT \times NT$ block diagonal matrix where element w_{ij} reflects the degree of connection from i to j according to some metric (such as the similarity or complimentarity between i's and j's economies or their trade bundles). For instance, a standard $\mathbf{W}$ matrix is expressed as

$$\left[\begin{array}{ccccc:c:ccccc}
w_{11_1} & w_{12_1} & w_{13_1} & \cdots & w_{1n_1} & & & & & & \\
w_{21_1} & w_{22_1} & w_{23_1} & \cdots & w_{2n_1} & 0 & & & 0 & & \\
\vdots & \vdots & \vdots & \ddots & \vdots & & & & & & \\
w_{n1_1} & w_{n2_1} & w_{n3_1} & \cdots & w_{nn_1} & & & & & & \\
\hdashline
 & & 0 & & & \ddots & & & 0 & & \\
\hdashline
 & & & & & & w_{11_T} & w_{12_T} & w_{13_T} & \cdots & w_{1n_T} \\
 & & 0 & & & 0 & w_{21_T} & w_{22_T} & w_{23_T} & \cdots & w_{2n_T} \\
 & & & & & & \vdots & \vdots & \vdots & \ddots & \vdots \\
 & & & & & & w_{n1_T} & w_{n2_T} & w_{n3_T} & \cdots & w_{nn_T}
\end{array}\right]$$

Suppose "connection" in this instance is a shared border. Therefore, if country i and country j share a border, then $w_{ij} = 1$. Otherwise $w_{ij} = 0$. Hence, this block-diagonal matrix can be rewritten as

$$\left[\begin{array}{c:c:c} \begin{matrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 1 & \cdots 0 & \end{matrix} & 0 & 0 \\ \hdashline 0 & \ddots & 0 \\ \hdashline 0 & 0 & \begin{matrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 1 & \cdots 0 & \end{matrix} \end{array}\right]$$

where the diagonal in each matrix is zero (as it contains self-referencing entries; e.g., w_{11} contains the impact of country 1 on country 1). Having constructed this spatial weighting matrix, the impact on policy y of country i by country j is captured with

$$y = \rho \cdot \mathbf{Wy}, \tag{14}$$

where $\mathbf{y}$ is an $NT \times 1$ vector of outcome observations stacked by time (i.e., time 1, country 1 to N, then time 2, country 1 to N, through time T). Combined, $\mathbf{Wy}$ reduces to a vector, where the parameter ρ captures the impact the spatially weighted outcome of countries $-i$ has on the outcome of country i. Franzese and Hays (2007b) then recommend row standardizing the resulting spatial weighting matrix by replacing the ones in each country's row in the weight matrix with $1/N$, where N is the number of countries the country borders. This procedure normalizes the sums across rows of cell entries to 1 and creates a nonuniform weighting matrix.

Given the emphasis spatial interdependence regression models place on the joint determination of policy outcomes, one might suspect that this approach offers a way to properly estimate the creation of multilateral alliances. Specifically, one could conceptualize membership in the same alliance as analogous to "neighboring" states in the capital taxation context. Returning to the notation from the previous section, $\mathbf{y}$ is an $NT \times 1$ vector of capability observations stacked by country and the spatial weighting matrix is computed by coding $w_{ij} = 1$ for countries i and j that are members of the same *multilateral* alliance, $w_{ij} = 0$ for countries that are not part of the same multilateral alliance. However, one must keep in mind that the weighting matrix captures the ability of state i to influence state j on policy y. In other words, it captures whether or not i and j have a connection, which is precisely what one needs to estimate! This means that whereas the spatial regression model treats $\mathbf{W}$ as a prespecified right-hand side variable, $\mathbf{W}$ is actually the element one needs to estimate as a left-hand side variable. It is for this reason that spatial interdependence regression will be inappropriate for modeling k-adic data.

5.3 *Evolving Network Models*

Another approach is offered by longitudinal network (or ‘‘evolving networks’’) models, which attempt to directly model the interdependence of states as a network and then statistically estimate the network data.[13] Though scholars such as Warren (Forthcoming 2010) have used these models to study multilateral events (specifically military alliances), evolving networks models currently cannot properly model a *k*-adic DGP. To understand why this is the case, one need only to briefly consider the setup of these models.

Formally, let N actors be connected according to an observed, binary endogenous, and time-variant connectivity matrix, $\mathbf{x}$, with elements $\mathbf{x}_{ij}(t)$, representing the connection between actor i and j at time t (which is analogous to the weighting matrix, $\mathbf{W}$, of Franzese and Hays). Let $\mathbf{z}$, be a vector of N observed, binary behaviors at time t (analogous to $\mathbf{y}(t)$ in Franzese and Hays). Actors have opportunities to make changes in their network connections, switching on or off one at a time or doing nothing. When the opportunity to change network connections arrives for some i, this actor chooses to change the status on one of his/her $N - 1$ connections, turning it on or off, or leaving them all unchanged. The actor makes this choice by comparing the values of some objective function specified by

$$f_i^{\text{net}}\left(x, x' z\right) + \varepsilon_i^{\text{net}}\left(x, x', z\right), \tag{15}$$

where f^{net} is a deterministic objective function that can be interpreted as a measure of the actor’s satisfaction with the result of the network decision, and ε^{net}, is a random disturbance term representing unexplained change that is assumed extreme value distributed. This, coupled with the additional assumption that the data are IIA, allows the objective function to take on multinomial logit shape of categorical choice (where each category is a relation with another actor).

This specification illustrates two reasons why the evolving networks approach will not properly model *k*-adic data. First, the notation $\mathbf{x}_{ij}(t)$ illustrates that the presence of a connection is *dyadically* measured. More concretely, even though this approach can identify the presence of a connection between any two of n countries and even determine if these links serve to ‘‘close’’ a triangular relationship, it cannot distinguish between a triangular relationship that is closed due to the presence of a single trilateral alliance and a triangular relationship that is closed due to the presence of interlinking bilateral alliances.

Second, the underlying IIA assumption means these models treat each node’s decision regarding which ties to form as independent of every other nodes’ decisions. Thus, one does not directly model, for the specific edges between specific i, j, and k, the probability of i and j being connected as a function of the probability that j and k are connected.

Moreover, current methods for statistically estimating network data, such as the Simulation Investigation for Empirical Network Analysis (*SIENA*) software package developed by Ripley and Snijders (2010), can estimate the presence of ties and the similarity in covariate values between no more than two states. For example, in the above simulations, all countries have a country-specific explanatory variable (i.e., the level of capabilities). The values of this explanatory variable can be entered into a network-analytic program, which then estimates a summary statistic (such as capability ‘‘similarity’’ between two states). This summary statistic is some function of the edges and/or nodes—which is to say, it is some function of the 1s or 0s that indicate a connected or

[13]The discussion that follows is drawn from Franzese, Hays, and Kachi (2009) and Steglich, Snijders, and Pearson (Forthcoming 2010).

a nonconnected pair of nodes and/or of characteristics of those nodes. For instance, *SIENA* uses the following formula to compute "similarity":

$$1-\left(\frac{|v_i-v_j|}{r_V}\right), \tag{16}$$

where v_i is the capability score of state i, v_j is the capability score of state j, and r_V is the difference between the highest and lowest capability scores in the data set. Hence, because this formula only measures the similarity in capabilities between two states, *SIENA*, in essence, only estimates how this dyadic statistic impacts the probability of two states forming an alliance. Consequently, network-analytic models cannot circumvent the essentially dyadic nature of the information in the data as recorded and used (e.g., they cannot distinguish from *i*-*j*-*k* connected in three binary treaties from *i*-*j*-*k* connected in one trilateral agreement).[14]

6 Conclusion

The following events did not occur: Belgium and Turkey formed a bilateral military alliance during the second half of the 20th century, Greece and Germany fought an isolated bilateral war in the 1940s, and the French Greens and Communists formed a bilateral electoral coalition. However, empirical scholars have widely used data suggesting otherwise: when analyzing multilateral events, they often divide the actors involved into a series of dyadic relations, thereby creating observations that disregard the dyad's relations with outside actors. Though this practice has some benefits, such as making the operationalization of a variety of concepts (e.g., geographic distance) intuitive and straightforward, this practice also leads to flawed statistical inference. Specifically, through a series of simulations, I show that one cannot recover a k-adic DGP (with $k > 2$) using *dy*-adic data. Instead, one must analyze k-adic events using k-adic data.

Unfortunately, creating and estimating data sets containing all k-adic combinations of countries can be computationally infeasible. Therefore, I also show that choice-based sampling offers a means of creating k-adic data sets of manageable size.[15] Specifically, suppose there are n actors and suppose that $q < n$ is the largest combination of these n actors that experienced an event. In this case, one must first collect all k-ads that witnessed the event. Next, one should obtain a sample of k-ads that did not witness the event, where the largest of these nonevent k-ads is of size q (i.e., no larger than the largest k-ad that witnessed the event). One can then apply a rare events logit to this choice-based k-adic data set to obtain unbiased parameter estimates.

Where should scholars go from here? Two steps seem immediately clear. First, researchers must ensure that ongoing and future research on multilateral events use k-adic data.

Second, scholars should return to the numerous studies that have used dyadic data to analyze multilateral events, such as (but not limited to) the influential interstate conflict studies

[14]Two other methods for estimating network data include exponential random graph models (see Robins and Morris [2007] for a primer on these models) and neural network models, as applied by Beck, King, and Zeng (2000). However, both are greatly limited in their ability to capture k-adic processes as they require a dyadic-based measure of connectivity between nodes.

[15]I have written software for converting dyadic observations into the appropriate number of k-adic observations. This software is available through the author's Web site (http://sitemaker.umich.edu/poast.paul/home) or by typing "findit kadcreate" into the command line of STATA. Code and files to assist readers through the process of coding independent variables in k-adic data and estimating the resulting data set are also available on the author's Web site.

of Bremer (1992), Oneal and Russett (1997), and Reiter and Stam (2002). Specifically, scholars should reevaluate previous studies in a manner consistent with the recommendations of this paper. The simplest reevaluation would require limiting the analysis to just bilateral events as this would retain the original dyadic structure of the data. However, because we must still seek to understand multilateral events, one should also restructure the data along the lines recommended in this paper. This will entail identifying the size of the largest multilateral event, identifying the distribution of multilateral events (i.e., how many involved 5 actors, 4 actors, etc.), and then creating a sample of nonevent *k*-adic observations that are stratified according to the distribution of *k*-adic observations that witnessed the event. Once the *k*-adic data sets are created, it will be critical for scholars to code the independent variables. I used the Gibler and Wolford (2006) study of military alliance formation to illustrate some of the various ways scholars can code variables that, like geographic distance, have an intuitive interpretation in dyadic data but whose coding is less clear in a *k*-adic setting. Of course, the manner of coding these variables that will best capture the concept the scholar wishes to operationalize can ultimately only be chosen by that scholar.

There are two important caveats one must keep in mind when working with *k*-adic data. First, some multilateral events of exceptional size are simply too rare to be estimated using quantitative methods. In particular, recall that estimating a *k*-adic data set after applying choice-based sampling requires weighting each observation by the inverse probability of it being drawn from its respective stratum. Consequently, events such as the formation of NATO, with 12 countries at the time of formation, require placing such a massive weight on a single observation that the SEs are rendered uninformative. What constitutes such an "unusual" (versus simply a "rare") *k*-adic event will vary according to the number of possible actors from which the event could have been formed. However, the analyst must be aware that some multilateral events will simply not be amenable to quantitative analysis.

Second, as emphasized in the introduction, using *k*-adic data will not address the many other sources of nonindependence pervasive in dyadic data sets. Spatial, strategic, temporal, or monadic interdependencies are still features of the data that must be modeled. However, if the process the scholar wishes to evaluate is *k*-adic in nature, then, regardless of the estimation technique employed, that scholar cannot use dyadic data. Instead, multilateral events must be modeled using *k*-adic data.

References

Achen, Christopher. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22:327–39.

Basinger, Scott, and Mark Hallerberg. 2004. Remodeling the competition for capital: How domestic politics erases the race to the bottom. *The American Political Science Review* 98:261–7.

Beck, Nathaniel, and Jonathan Katz. 2001. Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon. *International Organization* 55:487–95.

Beck, Nathaniel, Jonathan Katz, and Richard Tucker. 1998. Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42:1260–88.

Beck, Nathaniel, Gary King, and Lanche Zeng. 2000. Improving quantitative studies of international conflict: A conjecture. *The American Political Science Review* 94:21–35.

Bennett, Scott, and Allan Stam. 2000. Research design and estimator choices in the analysis of interstate dyads: When decisions matter. *The Journal of Conflict Resolution* 44:653–85.

———. 2007. *The behavioral origins of war.* Ann Arbor: University of Michigan Press.

Bianco, William T., and Robert H. Bates. 1990. Cooperation by design: Leadership, structure, and collective dilemmas. *The American Political Science Review* 84:133–47.

Bremer, Stuart A. 1992. Dangerous dyads: Conditions affecting the likelihood of interstate war, 1816–1965. *The Journal of Conflict Resolution* 36:309–41.

Clarke, Kevin. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22:341–52.

Croco, Sarah, and Tze Kwang Teo. 2005. Assessing the dyadic approach to interstate conflict processes: A.k.a. ''dangerous'' dyad-years. *Conflict Management and Peace Science* 22:5–18.

Findley, Michael, and Tze Kwang Teo. 2006. Rethinking third party interventions into civil wars: An actor-centric approach. *Journal of Politics* 68:828–37.

Franklin, M., and T. Mackie. 1984. Reassessing the importance of size and ideology in the formation of governing coalitions in parliamentary democracies. *American Journal of Political Science* 28:671–92.

Franzese, Robert, and Jude Hays. 2007a. Empirical models of spatial interdependence. In *Oxford handbook of political methodology*, ed. Colin Read and Greg N. Gregorious, 570–604. Oxford: Oxford University Press.

———. 2007b. Empirical models of international capital-tax competition. In *International taxation handbook*. 43–72.

———. 2007c. Spatial-econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis* 15:140–64.

———. 2008. Interdependence in comparative politics. *Comparative Political Studies* 41:742–80.

Franzese, Robert, Jude Hays, and Aya Kachi. 2009. *The m-STAR model as an approach to modeled, dynamic, endogenous interdependence in comparative and international political economy.* Working paper.

Gent, Stephen. 2007. Strange bedfellows: The strategic dynamics of major power military interventions. *Journal of Politics* 69:1089–102.

Gibler, Douglas, Toby Rider, and Mare Hutchison. 2005. Taking arms against a sea of troubles: Conventional arms races during periods of rivalry. *Journal of Peace Research* 42:131–47.

Gibler, Douglas, and Scott Wolford. 2006. Alliances, then democracy: An examination of the relationship between regime type and alliance formation. *The Journal of Conflict Resolution* 50:1–25.

Green, Donald, Soo Yeon Kim, and David H. Yoon. 2001. Dirty pool. *International Organization* 441–68.

Golder, Sona. 2006. *The logic of pre-electoral coalition formation*. Columbus: Ohio State University Press.

Hays, Jude. 2003. Globalization and capital taxation in consensus and majoritarian democracies. *World Politics* 56(3):79–113.

Hoff, Peter. 2005. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association* 100:286–95.

Hoff, Peter, and Michael Ward. 2004. Modeling dependencies in international relations networks. *Political Analysis* 12:160–75.

Kadera, Kelly, and Sara McLaughlin Mitchell. 2005. Manna from heaven or forbidden fruit? The (Ab)use of control variables in research on international conflict. *Conflict Management and Peace Science* 22:273–5.

King, Gary. 2001. Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization* 55:497–507.

King, Gary, and Langche Zeng. 2001a. Explaining rare events in international relations. *International Organization* 55:693–715.

————. 2001b. Logistic regression in rare events data. *Political Analysis* 9:137–63.

Lai, Brian, and Dan Reiter. 2000. Democracy, political similarity, and international alliances, 1816–1992. *The Journal of Conflict Resolution* 44:203–27.

Leeds, Brett Ashley, Jeffrey M. Ritter, Sara McLaughlin Mitchell, and Andrew Long. 2002. Alliance treaty obligations and provisions, 1815–1944. *International Interactions* 28:237–60.

Mansfield, Edward, Helen Milner, and Edward Rosendorff. 2002. Why democracies cooperate more: Electoral control and international trade agreements. *International Organization* 56:477–513.

Mansfield, Edward, and Eric Reinhardt. 2003. Multilateral determinants of regionalism: The effects of GATT/WTO on the formation of preferential trading arrangements. *International Organization* 57:829–62.

Martin, Lanny, and Randolph Stevenson. 2001. Government formation in parliamentary democracies. *American Journal of Political Science* 45:33–50.

Morrow, James. 1991. Alliances and asymmetry: An alternative to the capability aggregation model of alliances. *American Journal of Political Science* 35:904–33.

Oneal, John, and Bruce Russett. 1997. The classic liberals were right: Democracy, interdependence, and conflict, 1950–1985. *International Studies Quarterly* 41:267–93.

Peceny, Mark, Caroline Beer, and Shannon Sanchez-Terry. 2002. Dictatorial peace? *The American Political Science Review* 96:15–26.

Ray, James Lee 2005. Constructing multivariate analyses (of dangerous dyads). *Conflict Management and Peace Science* 22:277–92.

Reiter, Dan, and Allan Stam. 2002. *Democracies at war*. Princeton, NJ: Princeton University Press.

———. 2003. Indentifying the culprit: Democracy, dictatorship, and dispute initiation. *The American Political Science Review* 97:333–7.

Remmer, Karen. 1998. Does democracy promote interstate cooperation? Lessons from the Mercosur region. *International Studies Quarterly* 42:25–51.

Ripley, Ruth, and Tom A. B. Snijders. 2010. *Manual for SIENA version 4.0*. Department of Statistics, Nuffield College, University of Oxford. http://stat.gamma.rug.nl/s_man400.pdf (accessed July 15, 2010).

Robins, Garry, and Martina Morris. 2007. Advances in exponential random graph models. *Social Networks* 29:169–72.

Signorino, Curtis. 1999. Strategic interaction and the statistical analysis of international conflict. *The American Political Science Review* 93:279–97.

Starr, Harvey. 2005. Cumulation from proper specification: Theory, logic, research design, and 'nice' laws. *Conflict Management and Peace Science* 22:353–63.

Steglich, Christian, Tom Snijders, and Michael Pearson. Forthcoming 2010. Dynamic networks behavior: Separating selection from influence. *Sociological Methodology*.

Stone, Randall, Branislav Slantchev, and London Tamar. 2008. Choosing how to cooperate: A repeated public-goods model of international relations. *International Studies Quarterly* 52:335–62.

Ward, Michael, Randolph Siverson, and Xun Cao. 2007. Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science* 51:583–601.

Wallace, Michael. 1976. Arms races and the balance of power: A preliminary model. *Applied Mathematical Modeling* 1:83–92.

———. 1979. Arms races and escalation: Some new evidence. *The Journal of Conflict Resolution* 23:3–16.

Warren, T. Camber. Forthcoming 2010. The geometry of security: Modeling interstate alliances as evolving networks. *Journal of Peace Research*.

Weede, Erich. 1980. Arms races and escalation: Some persisting doubts. *The Journal of Conflict Resolution* 24:285–7.