Supplemental Appendix

MID4 Search Parameters and Source Identification

Unlike MID3, which relied on a dyadic/region specific identification of MID-relevant news stories, MID4 uses a consistent and uniform approach for establishing the initial set of documents. We use the approach discussed in Schrodt, Palmer and Haptipoglu (2008) for constructing global search parameters, which allows researchers to search across multiple reporting news agencies, providing global coverage for each day of the year. To do so, we compiled the following MID3 related search words in a comprehensive, MID-specific, LN boolean search string:

> (air base OR air strike OR airbase OR aircraft OR airstrike OR alert OR antiaircraft OR armed OR armo! OR arms OR army OR artillery OR attack OR batteries OR battery OR battle OR battleship OR block! OR bomb OR border OR buildup OR carrier OR casualties OR casualty OR cease OR ceasefire OR cease-fire OR clash! OR combat OR conflict OR crisis OR cruiser OR damage OR declare war OR defence OR defense OR defensive measures OR defian! OR deploy! OR destroy OR detained OR dispatch! OR display of force OR dispute! OR embargo OR erupt! OR fight! OR fire OR fired OR forc! OR fortification OR hit OR hostile OR incursion! OR infantry OR interstate OR invasion OR jet OR kill! OR launch! OR liberate OR line of control OR maneuver OR milit! OR missile! OR mobiliz! OR mortar OR naval OR nuclear OR occup! OR offensive OR operation OR patrol! OR peace declaration OR pullback OR radar OR raid! OR recon! OR reinforcement OR reprisal OR retali! OR rocket OR security OR seiz! OR shell! OR shoot OR shot down OR show of force OR shrapnel OR skirmish OR soldier! OR squadron OR stronghold OR subside! OR target OR tension! OR territ! OR threat! OR trade fire OR troop OR truce OR ultimatum OR USS OR vessel OR violat! OR violence OR vows to OR war OR warn! OR warplane OR warship OR weapon! OR weapons OR withdraw!)

To improve the initial precision from LN, every query in this part of the data collection process also includes a set of "AND NOT" exclusion parameters. Their purpose is to improve the "true" to "false" ratio of returns by systematically removing stories with the following words contained in the headline:

> AND NOT (sports OR business OR lifestyle OR tax cuts OR entertainment OR

Table 1: MID4 News Sources

Associated Press	Deutsche Presse Agentur	London Times
United Press International	Japan Economic Newswire	New York Times
Agence France Presse	ITAR-TASS News Agency	Interfax
British Broadcasting Corporation	Montreal Gazette	AFX News
Xinhua General News Service	Jerusalem Post	CNN
(All sources are English versions)		

Wall Street OR budget OR baseball OR food OR weather OR health OR natural disasters)

Although these parameters are intended to be all-inclusive, source selection remains an important concern. Including too many sources can lead to redundancy and loss of precision. However, too few sources can lead to loss of recall due to the sources' scope of coverage and various types of reporting biases. Our goal is to use the fewest number of sources capable of collecting all MID-relevant information.

We begin our source selection process with all news sources that are found to contain MID-relevant information in MID3. Sources were initially kept or rejected based on temporal coverage. Because the LN database contains subscriptions to news agencies on a year-to-year basis, some sources in the initial list were removed because they did not have the required nine year coverage. This validation process results in a pared down list of thirty different candidate sources with coverage that matched the temporal domain of the project.

We then group sources based on geographic coverage, here referred to as Lists A, B, and C. List A contains the sources with the most international coverage, comprising of agencies across multiple continents. Examples include the Associated Press and British Broadcasting Corporation; the complete list of sources is found in Table 1. List B is more restrictive, containing agencies confined to a continent or hemisphere. Examples include the *Boston Globe* and *Toronto Star*. List C is the most specific, having agencies with coverage constrained to a particular region of a continent or area of the world. Examples include *Straits Times* (Singapore) and *New Straits Times* (Malaysia).

These lists were then evaluated to assess two distinct concerns. The first is the cost, in terms of unnecessary additional stories, of using all thirty potential sources, rather than only the global sources on List A. To assess this, we queried the LN database using our search string on fifteen randomly generated dates between the years 2003-2004 to examine the number of returns A, B, and C yield. The averaged results of these tests are displayed below:

- List A: Daily average of 2,365
- List B: Daily average of 266
- List C: Daily average of 320

Lists B and C would add an additional 550-600 stories per day, which we then evaluated to determine whether these contained new reports relevant to the coding. Ten MIIs from MID3 were selected at random for the year 2001. Using our global search parameters and querying on the day *after* each MII began, we assessed each list on its ability to catch these incidents. The List A candidates caught six of ten, B one of ten, and C zero. Next, we looked at dates before, on, *and* after the start of the ten randomly chosen MIIs. This subsequent test improved A's performance to ten of ten, with B and C catching no additional MIIs. Based on these results, we concluded that List A is sufficient.

This result runs counter to the intuition of many researchers that local sources contain more detail on local events. That may well have been the case historically, prior to the advent of low-cost electronic news sources, but in the contemporary environment, there are two reasons that the international sources are likely to be more comprehensive. First, the major international sources now have arrangements with the local papers – or individual "stringers" reading these – to pick up any information likely to be of interest (and MIIs would almost always be in this category). Second, the Web-based version of local papers is often the printed version of that paper, which is subject to space limitations which are not found in the all-electronic wire services. The absence of significant valueadded that we found in our experiments was consistent with the findings of the Integrated Conflict Early Warning System project (O'Brien 2010).

References

O’Brien, Sean. 2010. “Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research.” *International Studies Review* 12(1):87–104.

Schrodt, Philip A., Glenn Palmer and Mehmet Emre Haptipoglu. 2008. “Automated Detection of Reports of Militarized Interstate Disputes: The SVM Document Classification Algorithm.” Presented at the Annual Meeting of the American Political Science Association, Toronto, Canada.