Event Data and the Construction of Reality

The past few years have given rise to a number of datasets on social behavior, security, and political conflict that attempt to deliver on the promise and appeal of near real-time, automated event data collection. While these efforts have succeeded in generating large datasets on conflict and political violence, the extent to which these data represent a valid, accurate reflection of reality remains to be seen. With a few isolated exceptions, we have largely failed to challenge these data and critically assess their strengths and limitations.

Developers of automated event data collection platforms offer a panacea, claiming, "because the marginal costs of data collection are so low, it is possible to effectively monitor hot spots from the Straits of Hormuz to the South China Sea, and burgeoning protest movements from Brazil to Burma."[i] Meanwhile, analysts enticed by the appeal of low-cost, near-real-time data often use these datasets with the assumption that the developers implement state-of-the-art techniques to ensure their validity.[ii] One might argue that even if different datasets contain conflicting or invalid records, if the resulting conclusions are generally consistent the resource savings far outweigh the loss of data validity. Users are encouraged to "embrace the suck" and use these event databases to measure relative change, rather than as reliable catalogs of events.[iii] While this may be defensible in some cases, the presumption that it is universally acceptable is certainly problematic.

Given the broad appeal of event data generated at the push of a button, assessing the validity of the resulting data is essential. For example, the number of suicide attacks recorded worldwide over a two-month period varies widely between datasets from a low of 10 to a high of nearly 5,200. Likewise, these data might lead one to conclude that President Obama has been assassinated more than 900 times. Regardless of the sophistication of the analytical methodology employed, the uncritical use of data can lead to dubious conclusions and dangerously misinformed policy recommendations. While no data collection strategy is without flaws,[iv] the scale of datasets produced using fully automated techniques makes the prospect of validation daunting. Although there have been efforts to compare aggregate trends produced by datasets generated using automated strategies, we are unaware of any comprehensive attempts to explicitly evaluate the validity of individual records across multiple datasets.[v]

With the goal of systematically assessing the accuracy and consistency of data produced by automated techniques, we compare five datasets selected for their centrality to the question at hand, their popularity within the larger research and practitioner communities, and their diversity of data collection techniques. These include three datasets identified in the Grand Data Challenge—the Integrated Conflict Early Warning System (ICEWS), the Global Database of Events, Language, and Tone (GDELT), and the Phoenix dataset—each the product of fully automated data collection strategies. In comparison, the Global Terrorism Database (GTD) adopts a hybrid (computer/human) collection strategy in which automated tools,

including Boolean filtering, natural language processing, named entity recognition, and machine learning models winnow a pool of source documents, which are then passed to human analysts for unique event identification, coding, and quality assurance.[vi] Finally, the Suicide Attack Database (SAD) employs a team of human coders for source recognition, event identification, and coding.[vii]

We evaluate these datasets in depth by comparing the identification and event details of suicide attacks that occurred worldwide in January and February of 2015. We isolate this relatively narrow time frame in the interest of feasibility, and focus on suicide attacks for three primary reasons. First, suicide attacks are uniformly included in databases that use automated, hybrid, and manual collection techniques, which allows for systematic comparison across datasets. Second, suicide attacks are newsworthy enough that the media sources used by each of the databases reliably report them. Finally, reporting on suicide attacks is less susceptible to variations in descriptive language that may compromise the accuracy of automated techniques. Suicide attacks are typically described using similar words and phrases, regardless of information source.

In order to quantify the validity of data on suicide attacks in the databases under review, we develop a standardized coding scheme and compare each dataset case by case in terms of event inclusion, event date, and event location. We consult original source materials where practicable, and consider the particular inclusion criteria used for each database to accommodate the possibility that inclusion discrepancies are by design. By manually reviewing the events recorded in each dataset, we produce a new dataset that captures the extent to which suicide attacks are undercounted (false negatives) or overcounted (false positives or duplicates), and whether the date and country of each suicide attack is accurately recorded.

With the resulting data we leverage fundamental tools commonly used in the fields of information retrieval and statistics—precision, recall, F-measures, and inter-rater reliability measures—supplemented with descriptive statistics to characterize key threats to the accuracy of the data. We specifically consider the salience of deficiencies in entity extraction (date/location); automated geo-location; topic segmentation; source selection and quality; and de-duplication. We conclude by discussing existing computational techniques that are typically under-utilized by dataset developers despite the fact that they offer promising solutions to these problems. By systematically evaluating the quality of these data we introduce the means to measure the improvements of any of these techniques, advancing the field, in pursuit of data collection that is both cost-effective and valid.

Team:
Brian Wingenroth, START, University of Maryland, wingenroth@start.umd.edu
Erin Miller, START, University of Maryland, eemiller@gmail.com
Michael Jensen, START, University of Maryland, majensen@umd.edu
Omi Hodwitz, START, University of Maryland, omi@umd.edu
Kieran Quinlan, START, University of Maryland, kieranbquinlan@gmail.com

[i] Beieler, J., Brandt, P. T., Halterman, A., Schrodt, P. A., & Simpson, E. M. (2016). *Generating political event data in near real time: Opportunities and challenges.* In R. M. Alvarez (ed.), *Computational social science: Discovery and prediction* (pp. 98-120), New York: Cambridge.

[ii] Chalabi, M. (2014, May 13). Mapping kidnappings in Nigeria (updated). *FiveThirtyEight*. Retrieved from http://fivethirtyeight.com/datalab/mapping-kidnappings-in-nigeria/

[iii] Ulfelder, J. (2014, January 14). Using GDELT to Monitor Atrocities, Take 2. *Dart-throwing chimp*. Retrieved from https://dartthrowingchimp.wordpress.com/2014/01/14/finding-atrocious-needles-in-the-gdelt-haystack/

[iv] King, G. & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International organization*, 57, 617-642.

[v] Leetaru, K. and Schrodt, P. A. (2013, April). GDELT: Global data on events, location, and tone, 1979-2012. Paper presented at the annual convention of the International Studies Association, San Francisco, CA. Retrieved from http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

Ward, M. D, Beger, A., Cutler, J., Dickenson, M., Dorff, C. & Radford, B. (2013, October 15). Comparing GDELT and ICEWS event data. WardLab. Retrieved from http://mdwardlab.com/sites/default/files/GDELTICEWS_0.pdf

Hanna, A. (2014, February 24). Assessing GDELT with handcoded protest data. Bad Hessian. Retrieved from http://badhessian.org/2014/02/assessing-gdelt-with-handcoded-protest-data/

[vi] National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2015). Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd

[vii] Chicago Project on Security and Terrorism (CPOST). 2016. Suicide Attack Database (April 19, 2016 Release). [Data file]. Retrieved from http://cpostdata.uchicago.edu/