TwoRavens
Statistical Analysis and Data Repositories

Vito D’Orazio, James Honaker, Rohit Bhattacharjee

Assistant Professor of Political Science
University of Texas at Dallas

Center for Research on Computation and Society
Harvard University

Department of Computer Science
University of Texas at Dallas

November 16, 2016
Data Repositories and Analysis

- Massive proliferation of data in recent years
- Relatedly, an increase in number, size, scope of data repositories
  - Dataverse continues to grow as a repository for research data
  - Others include Figshare, Dryad, even GitHub
- This means more data in centralized locations
Data Repositories and Analysis

- Massive proliferation of data in recent years
- Relatedly, an increase in number, size, scope of data repositories
  - Dataverse continues to grow as a repository for research data
  - Others include Figshare, Dryad, even GitHub
- This means more data in centralized locations

TwoRavens is a tool for exploring and analyzing data that sits in a data repository.
Social Science Data

- Survey data
- Experimental data
- Observational data
- Text data
- Qualitative data
Social Science Data

- Survey data
- Experimental data
- Observational data
- Text data
- Qualitative data

TwoRavens functions with **tabular data**.
<table>
<thead>
<tr>
<th>country</th>
<th>year</th>
<th>war</th>
<th>polity2</th>
<th>lpop</th>
<th>lgdpenl1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFGHANISTAN</td>
<td>1996</td>
<td>1</td>
<td>-7</td>
<td>10.09434</td>
<td>NA</td>
</tr>
<tr>
<td>AFGHANISTAN</td>
<td>1997</td>
<td>1</td>
<td>-7</td>
<td>10.12523</td>
<td>NA</td>
</tr>
<tr>
<td>AFGHANISTAN</td>
<td>1998</td>
<td>1</td>
<td>-7</td>
<td>10.15662</td>
<td>NA</td>
</tr>
<tr>
<td>AFGHANISTAN</td>
<td>1999</td>
<td>1</td>
<td>-7</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1991</td>
<td>0</td>
<td>-8</td>
<td>8.232706</td>
<td>8.232714</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1992</td>
<td>0</td>
<td>-9</td>
<td>8.302018</td>
<td>8.232714</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1993</td>
<td>0</td>
<td>-9</td>
<td>8.368229</td>
<td>8.108946</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1994</td>
<td>0</td>
<td>-9</td>
<td>8.390722</td>
<td>7.937374</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1995</td>
<td>0</td>
<td>-9</td>
<td>8.413609</td>
<td>7.72493</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1996</td>
<td>0</td>
<td>-9</td>
<td>8.433376</td>
<td>7.62732</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1997</td>
<td>0</td>
<td>-9</td>
<td>8.446342</td>
<td>7.538202</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1998</td>
<td>0</td>
<td>-9</td>
<td>8.45914</td>
<td>7.405327</td>
</tr>
<tr>
<td>TURKMENISTAN</td>
<td>1999</td>
<td>0</td>
<td>-9</td>
<td>NA</td>
<td>7.460187</td>
</tr>
</tbody>
</table>
What is TwoRavens?

- TwoRavens automatically analyzes tabular data files and produces a metadata file.
- Users explore and interact with the metadata through the interface.
What is TwoRavens?

- TwoRavens automatically analyzes tabular data files and produces a metadata file.
- Users explore and interact with the metadata through the interface.
What is TwoRavens?

- Tool for estimating statistical models
  - Statistical model estimates effect of variable X on variable Y using a tabular data file, for example
  - In this sense, not unlike R, Stata, SPSS, even Excel
What is TwoRavens?

• Tool for estimating statistical models
  ▶ Statistical model estimates effect of variable X on variable Y using a tabular data file, for example
  ▶ In this sense, not unlike R, Stata, SPSS, even Excel
• Users specify relationships among variables using the language of directed graphs
What is TwoRavens?

- Tool for estimating statistical models
  - Statistical model estimates effect of variable X on variable Y using a tabular data file, for example
  - In this sense, not unlike R, Stata, SPSS, even Excel
- Users specify relationships among variables using the language of directed graphs
What is TwoRavens?

- Web-based and runs entirely in the browser
- Metadata sent client-side, data are never local
  - In this sense, not unlike visualization tools like Plotly, Shiny, Google Public Data Explorer
What is TwoRavens?

- Web-based and runs entirely in the browser
- Metadata sent client-side, data are never local
  - In this sense, not unlike visualization tools like Plotly, Shiny, Google Public Data Explorer
TwoRavens Design Features

- Device independent and broadly accessible
  - Only requires an internet connection
  - Does not presuppose expert statistical knowledge or experience with statistical software
- Option for gesture-driven
  - Works on tablets and phones
- Open source software
  - Backend processing using R and the Zelig library
Integration with Dataverse

- 32,842 tabular data files on Dataverse
Integration with Dataverse

- 32,842 tabular data files on Dataverse

Demo – http://2ra.vn/
TwoRavens for Large or Sensitive Datasets

Since the data are never local, TwoRavens is ideal for working with large data or sensitive data.

- Privacy Tools for Sharing Research Data
  - privacytools.seas.harvard.edu
- Modernizing Political Event Data for Big Data Social Science Research
  - http://goo.gl/l2NLwm
Figure: The curator architecture for data privacy.
TwoRavens and Event Data
Event Data

- Type of big data.
- Usually consists of millions of records.
- Five major fields: Date, Source, Target, Action, Location
- Need to manage the data somehow.
Phoenix Event Data

- News Events coded daily from 4th October, 2014 (approx. 771 days)

- 3000 events coded daily

- Approximately **2.3 million entries** and increasing

- Need some way to handle all this data
Handling Event Data

- Subset
- Aggregate
- Analyse
Subset

• Need to select range of data.

• Select according to date, source and target actors, location or action
• Only want events in 2014 where the USA government acted on Syria.

• Select source as “USAGOV”, target as “SYR” and year as “2014”
Aggregation

- Event data raw and unsuitable for model building

- Count events according to temporal or spatial unit

- Count events for each event category across the selected unit of analysis

- Plot the aggregated data for visualization
Unit of Analysis and event types

- A unit of analysis could be a temporal unit (date) or spatial unit (source actor)

- Events are categorized into 5 classes (0-4) called quad classes.

- Quad classes and event types and codes are defined by the CAMEO (Conflict and Mediation Event Observations) dictionary.
### Russia to Iraq events per month

<table>
<thead>
<tr>
<th>Date</th>
<th>Neutral</th>
<th>VCoop</th>
<th>MCoop</th>
<th>VConf</th>
<th>MConf</th>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-01-31</td>
<td>2</td>
<td>27</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-02-28</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-03-31</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-04-30</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-05-31</td>
<td>8</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-06-30</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-07-31</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-08-31</td>
<td>12</td>
<td>36</td>
<td>4</td>
<td>19</td>
<td>10</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-09-30</td>
<td>91</td>
<td>108</td>
<td>80</td>
<td>47</td>
<td>51</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-10-31</td>
<td>159</td>
<td>167</td>
<td>75</td>
<td>75</td>
<td>76</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-11-30</td>
<td>89</td>
<td>94</td>
<td>79</td>
<td>49</td>
<td>61</td>
<td>RUS</td>
<td>SYR</td>
</tr>
<tr>
<td>2015-12-31</td>
<td>73</td>
<td>51</td>
<td>60</td>
<td>20</td>
<td>37</td>
<td>RUS</td>
<td>SYR</td>
</tr>
</tbody>
</table>
Using Event Data in TwoRavens: Subset

- Option to subset by any combination of the five variables
- Can download the data from the tool after obtaining subset data
Using Event Data in TwoRavens: Aggregation

- After subset, user can choose to aggregate by any unit of analysis
- Different degrees of temporal unit maybe chosen for aggregation
Using Event Data in TwoRavens: Analysis

- TwoRavens architecture will allow the aggregated data to be analyzed and modelled in the webpage itself
TwoRavens

- A general purpose tool for analyzing data that sits on data repositories
- Valuable architecture for working with sensitive and large data
- Improve data analysis and proliferate quantitative reasoning by reducing barriers
TwoRavens

- A general purpose tool for analyzing data that sits on data repositories
- Valuable architecture for working with sensitive and large data
- Improve data analysis and proliferate quantitative reasoning by reducing barriers

2ra.vn

Vito D’Orazio
dorazio@utdallas.edu

James Honaker
james@hona.kr

Rohit Bhattacharjee
rxb151030@utdallas.edu