Research Reproducibility in Computational Social Science

Aek Palakorn Achananuparp, SMU
Research Integrity Conference 2018, Singapore
INTRODUCTION & DEFINITIONS
COMPUTATIONAL SOCIAL SCIENCE (CSS)

First coined by Lazer et al. (2009) in the Nature article

Modeling human activity, behavior, and relationships through the use of computational methods and large-scale data *(thousands to billions* of data points)
DATA SOURCES
“DIGITAL TRACES”

COMMON STUDY TOPICS

- Predicting friendships in social networks
- Modeling information diffusion process
- Predicting electoral outcomes
- Modeling human activity in offline settings
- Recommending books, papers, articles, movies, songs, etc.
WHAT DOES **REPRODUCIBILITY** MEAN?

<table>
<thead>
<tr>
<th>CONCEPT</th>
<th>TEAM</th>
<th>EXPERIMENT SETUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeatability</td>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Replicability</td>
<td>Different</td>
<td>Same</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Different</td>
<td>Different</td>
</tr>
</tbody>
</table>

Source: ACM
NON-COMPUTATIONAL V.S. COMPUTATIONAL RESEARCH

In non-computational research:

**Replicability = reproducibility**
= different groups can obtain the same result independently by following the original study’s methodology.

In computational research:

**Replicability** = different groups can obtain the same result using the original study's artifacts (datasets, code, and workflows).

**Reproducibility** = different groups can obtain the same result using independently developed artifacts.
We’ll mostly focus on replication and reproduction of computational research, i.e., computational reproducibility, in CSS.
REPRODUCIBILITY CRISIS IN CSS?
For electoral prediction studies using Twitter data, an independent group was not able to reproduce their positive results (Gayo-Avello et al. 2011).

61% of 21 social science studies published in Nature and Science can be reproduced (Camerer et al. 2018).

For 54% of 601 studies published at major computational research conferences, an independent group was able to build the code or the authors stated the code would build with some effort (Collberg et al. 2014).

Out of 400 artificial intelligence papers, 6% provide code for the papers’ algorithm, 30% provide test data, 54% provide pseudocode (Hutson, 2018).
REPRODUCIBILITY CHALLENGES IN CSS
TECHNOLOGICAL IRREPRODUCIBILITY

- Some code and dataset require high-performance or esoteric systems to run.
- Different tools, platforms, & versions may produce different results.
- Some software dependencies are no longer available.
- Is it still possible to run the original artifacts a few years later?
Data privacy is going to be more critical than before after the Cambridge Analytica fiasco.

More difficulty in collecting and sharing online social media data.

Data ownership is not always clear-cut.

Intellectual property prevents code sharing.
EXPERIMENTAL IRREPRODUCIBILITY

- Complex social systems are extremely difficult to study.
- States of the world are irrevocably not the same today compared to the time when the original experiments were conducted.
- Some external influences, e.g., media exposure, are almost impossible to control.
ENABLING REPRODUCIBLE RESEARCH
ENABLING REPRODUCIBLE RESEARCH

Open Research/Data Platforms

- Open Science Framework
- CodaLab
- ReScience
- Jupyter Notebooks
ENABLING REPRODUCIBLE RESEARCH

Open Data Repositories

- Microsoft Research Open Data
- Stanford Network Analysis Project (SNAP)
- UCI Machine Learning Repository
- GroupLens
- LARC Data Repository
LARC Data Repository
“Extraordinary claims require extraordinary evidence and extraordinary transparency.”

Aek Palakorn Achananuparp
palakorna@smu.edu.sg
@aekpalakorn
REFERENCES