README for the World Cybercrime Index project (2021 survey)

Files

- wci_data.csv
- wci_code.Rdata
- wci_code_studio.R
- wci_2021_markdown.Rmd

About the dataset

This data was collected for the World Cybercrime Index (WCI) project, which is a part of the CRIMGOV research group based in the Department of Sociology at the University of Oxford.

Cybercrime is a major challenge facing the world, with estimated costs in the hundreds of billions. Despite the threat it poses, cybercrime is largely an invisible phenomenon. This project proposes a solution: an expert survey with leading cybercrime professionals from across the world.

The result of the survey is the World Cybercrime Index, a global metric of cybercriminality organised around five types of cybercrime. The results indicate that a relatively small number of countries house the greatest cybercriminal threats. These findings partially remove the veil of anonymity around cybercriminal offenders, may aid law enforcement and policymakers, and contribute to the understanding of cybercrime as a local phenomenon.

We have made our dataset available so that other researchers can better understand our results and build on our analysis. We plan to run this survey multiple times. Future datasets will be published to a GitHub project site and will be named according to the year the survey is run; for example, “World Cybercrime index project (2024 survey)”.

This dataset should be used and read accompanied by our paper, “Mapping the geography of cybercrime with the World Cybercrime Index” (forthcoming). This document was prepared by Miranda Bruce. The project’s full list of collaborators are: Miranda Bruce, Jonathan Lusthaus, Ridhi Kashyap, Nigel Phair, and Federico Varese.

Data collection methodology

Raw data was collected via an online survey hosted on the Qualtrics platform. World leading experts in cybercrime intelligence, investigations, and attribution were invited to participate. The survey ran from March to October 2021 and recorded 92 complete responses.

The survey asked participants to consider five major categories of cybercrime, nominate the countries that they consider to be the most significant sources of each of these cybercrime categories, and then rank each nominated country according to impact, professionalism, and technical skill.

After the survey closed, the results were downloaded as a .csv file and cleaned; the data was anonymised, extraneous information removed, and variables renamed.

The data was then imported to R and analysed using the RStudio interface. Specific software versions and requirements are listed below under “Setup instructions”.

Survey structure

The survey identifies five types of cybercrime:

1) Technical products/services
2) Attacks and extortions
3) Data/identity theft
4) Scams
5) Cashing out/money laundering

Participants nominated up to five countries that they believe are major sources of each cybercrime type. Participants were free to nominate fewer than five countries, and could skip entire sections.

For each country nominated, the participant was prompted with three questions about that specific country. They were asked to rate each country on a 1-10 Likert-type scale against three measures:

- The impact of the cybercrimes produced there
- The professionalism of the offenders based there
- The technical skill of the offenders based there

Participants repeated this process for all five cybercrime types. The survey then collected additional information on the participants’ area(s) of expertise. Finally, participants were presented with a text box to leave comments.

Below is a generic example of the structure of each section of the survey. Wording is not exact. This example may be helpful for understanding how the data was cleaned, and how we performed our analysis in R.

> **Major sources of [cybercrime type]**
>
> [Cybercrime type] is an offence that involves [examples, description, definitions].
>
> Which countries are the major sources of [cybercrime type]? Choose up to 5 countries.
>
> [country 1] #*dropdown menu*
>
> [country 2]
>
> [country 3]
>
> [country 4]
>
> [country 5]
>
> ***#Next page***
>
> These are questions about [cybercrime type] in [country 1].
>
> How impactful is the cybercrime, from 1-10? *#Likert-type scale*

How professional are the cybercrime offenders, from 1-10?

How technically skilled are the cybercrime offenders, from 1-10?

#This question repeats for each country that the participant nominated on the previous page. If no countries were nominated, this page is skipped

How the data is organised

This section will make more sense if you read the "Survey Structure" subsection (above) first.

The wci_df_reprex.Rdata file uses dataframes that are organised both by observations (wide format) and by variables (long format) at different points. Which format was used depended on the type of analysis being done. Further detail is provided in the WCI_2021_markdown.rmd file.

Raw data

wci_data.csv

This document represents a cleaned version of the raw survey data, ready for importing into RStudio. It has been anonymised and cleaned for easier use.

The wci_data.csv file is organised by observations. Each row represents an observation (one row represents one survey participant). Each column represents a variable (one column represents one survey question). Columns are listed in the order in which they appeared in the survey.

Variables were renamed to make analysis in R easier. The five types of cybercrime (each of which constitute a separate section in the survey) were named like so:

- Technical products/services = Technical
- Attacks and extortion = Attack
- Data/identity theft = Data
- Scams = Scams
- Cashing out/money laundering = Cash

For each of these categories, there are five variables to designate the five countries that can be nominated per category. Following this logic, the columns are named like so: Technical1, Technical2, Technical3 … Attack1, Attack2, Attack3… and so on

Each of these variables are associated with three separate scores (impact, professionalism, technical skill). In other words: for each country nominated, there are three scores attached to it. These score variables were named like so:

- Impact = impact
- Professionalism = professional
- Technical skill = techskill

Following the country variable columns (Technical1, Technical2, Technical3…) are their associated scores. These columns are named like so: Technical1_impact, Technical1_professional, Technical1_techskill …

The final variables in the spreadsheet concern the participants' expertise in cybercrime types and/or geographical regions. For cybercrime type expertise, participants could choose one or multiple types, or select "other" and fill in a text box. These variables are named "Expertise_crimetype" and

"Expertise_crimetype_other". For regional expertise, participants were asked to select "Yes" or "No"; if Yes, they were asked to fill in a text box. These variables are named "Expertise_region" and "Expertise_region_other". The final column contains participant comments.

R – data transformation and analysis

wci_code.Rdata

wci_code_studio.R

wci_2021_markdown.Rmd

After being imported into RStudio, the dataset was further cleaned and transformed to make analysis easier. Variables follow a similar naming convention to the wci_data.csv file, but were further simplified. Detailed annotation and guidance can be found in wci_2021_markdown.Rmd

Following this initial cleanup and transformation, the data was used to generate the World Cybercrime Indices and analysed for various patterns in participant rating behaviour. These processes are explained in detail in wci_2021_markdown.Rmd

The wci_code_studio.R file also shows how the analysis was performed, but with minimal notation. It has not been prepared for reproducibility, and contains extraneous/experimental code, only some of which has been annotated out.

Setup instructions

Software needed:

- R-4.3.2 (or later)
- RStudio 2023.12.0+369 (or later)
- Excel (or equivalent .csv reader)

wci_2021_markdown.Rmd contains detailed instructions and explanations of how the data was imported, cleaned, transformed, and analysed. You should be able to reproduce all our analysis by following that document.

R packages used

Packages used are listed below. Click the links for more information on each package and how to install them.

- careless
- formattable
- ggmap
- ggrepel
- ggthemes
- magrittr
- mapproj
- maps
- naniar
- RColorBrewer
- renv
- rio
- scales

- tidyverse
- trend

How to cite

If you use the data in this repository for your own analysis, please cite both the data repository and the journal article it is associated with.

To cite the repository:

> Bruce, M., Lusthaus, J., Kashyap, R., Phair, N., Varese, F. (2024) World Cybercrime Index 2021. GitHub repository. LINK FORTHCOMING

To cite the journal article:

> Bruce, M., Lusthaus, J., Kashyap, R., Phair, N., Varese, F. (2024) Mapping the global geography of cybercrime with the World Cybercrime Index. PlosOne. LINK FORTHCOMING

Acknowledgements and funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 101020598 – CRIMGOV, Federico Varese PI).

The data collection for this project was carried out as part of a partnership between the Department of Sociology at the University of Oxford and UNSW Canberra Cyber. Data collection and analysis was supported by CRIMGOV.

Figure 1 was generated using information from OpenStreetMap and OpenStreetMap Foundation, which is made available under the Open Database License.

We would also like to thank our survey participants for donating their time and insights to our project. The WCI cannot exist without their input.

Contact

For more information about the World Cybercrime Index project, contact:

Dr Miranda Bruce
miranda.bruce@sociology.ox.ac.uk