URL Threat Score Analysis

Brandon Julian
University of Colorado Boulder
Boulder, United States
brju4682@colorado.edu
https://github.com/bdjulian/url-threatscore-project

***Abstract*—The Hybrid-Analysis website could possibly be a great resource for dataset creation as the number of threats is extensive (URL, exe, netflows, cmd, etc.). In my attempt to interpret/reverse-engineer Hybrid Analysis™ Falcon Sandbox© incident response threat assessments for malicious URL's; I was successful in determining some feature importances. Using a balance of features that weigh complexity and simplicty I was also successful in rather accurate predictions of H.A URL threat scores on a self created set of over X URL's. Hybrid Analysis's Falcon Sandbox incident response software weighs complexity and verbosity disproportionately compared to other features, and if a URL spawns unnecessary Network Activity it is typically a highly malicious domain. The final model used was a simple Random Forest Regressor tuned by GridSearch and model introspection was done using a variety of methods. If a user wanted to deploy this model they would need to collect their URL traffic information similarly to Hybrid Analysis.**

***Index Terms*—Malicious URL detection, features, static-analysis, Network Activity, Payload Delivery, Mutex, partial dependence**

I. Introduction

As the internet continues to grow, malicious intent also grows with it. The potential for users to be scammed or taken advantage of increases with the ever-evolving attack methods of the unscrupulous. The information and digital security space is an uneven arms race between malicious actors and their victims. By utilizing resources like Hybrid Analysis more tools can be created to help protect ourselves.

II. Business Understanding

The goal of this project is to observe what is important in a cutting-edge analysis algorithm of URL's which is otherwise obfuscated as a proprietary product. This modeling process can either be used to deliver on a prediction or inform research and security individuals. By utilizing data already extracted by Hybrid Analysis Falcon Sandbox and constructing it into features for machine learning (ML) it is possible to see what information is most important to an extremely successful product like Falcon Sandbox. If the top features are known it can inform other forensic teams what to look for in their own analysis without needing to deploy a proprietary product.

III. Data Understanding

A. Hybrid Analysis

To quote their FAQ, "This webpage is a free malware analysis service for the community. Using this service one can submit files for in-depth static and dynamic analysis." Files can be submitted to their site for analysis. However, Hybrid Analysis also produces a product they call Falcon Sandbox. Which is a much more verbose system that does most of the dynamic analysis behind the scenes. What is very cool about their website-product integration is when items are submitted for analysis by a user of their product Falcon Sandbox; it by default generates a high-level report of the item in question. This report include interesting features of the item, whether that be signs of attempted access escalation, known malicious artifacts are present, or process spawning. It also includes a threat score out of one-hundred and a final label in the list of malicious, suspicious, no specific threat, no verdict, or whitelisted. The information is stored in two parts - both accessible through their API with a CURL request, a more in-depth explanation on the dataset generation can be found on my repository

B. Target Variable

The target for each URL is a continuous number from 0-100 titled 'Threat Score' on the Hybrid Analysis website. All collected URL samples are classified as malicious by Hybrid Analysis intentionally as the intent of the project is to determine what makes something *more dangerous* not necessarily malicious overall.

C. Features

28 features were created utilizing the Falcon Sandbox reports, the code to create them as well as in-depth explanations of them are on my repository. Brief explanations will be here.

- avg-val-length: Average length of value key data in report. Total value divided by number of key value pairs. The value is representative of its type, a payload will have a filename and a hash of the file. While a peristence mechanism will have a registry key and its accompanying value.
- total-val-len: Total length of value key data.
- category-count: The number of categories found in the report. A larger category-count means a URL is more complex, possibly possessing multiple types of spawned network activity or multiple methods of payload delivery and installation. The possible unqiue values are: Artifacts Dropped, External Analysis, Network Activity, Payload

Delivery, Payload Installation, Persistence Mechanism.

- Each category type is also a one-hot encoded feature for each row to identify if a malicious URL possesss at least one method, file, or action of the above listed categories. (6 total columns).

- Each type of category is also one hot encoded: filename-md5, user-agent, domain-ip, mutex, ip-dst, regkey-value, comment, filename-sha512, domain, filename-sha256, link, filename-sha1. (12 total columns)

- Frequency values were also created form each individual category: this creates 6 more columns for each row which shows how many instances of each category make up the total.

D. Hybrid Analysis API - Search Query

At the beginning of the project I spent most of my time attempting to get their API and associated python wrapper up and running on my machine. That ended up being a tumultuous process due to extremely poor documentation on H.A.'s side and lack of my own patience. With the help of Dr. David Eargle of the Leeds MS. BA. program, he pointed out that the API was simply processing a POST request. And therefore, could be used outside of their python wrapper and web API interface. Using the requests package, we wrote a function to submit a post request. This is an upside even though I sunk an enormous amount of time into fiddling with their API because now using requests the process can be automated to return hundreds of searches. To be clear, the search query returns basic information of a file that was submitted to their platform for analysis in a nested JSON format. It includes some metadata, a verdict, threat score, and the most important piece which is a job-id. The job-id can then be used in their report download get request to download the full report of each file returned in the search. It is this full report that contains H.A's features and analysis on the item. See screenshots in the appendix (section IV) for examples of the code and return.

E. Hybrid Analysis API - Report Download

After the struggle that was implementing a scalable way to search their database, a scalable way to extract the features and therefore the reports needed to be made. Using the previous POST request format, I was able to create an extremely simple version for their report download. The only hang up I had was realizing that the report download was a GET request and not a POST. With that slight adjustment it worked as intended.

F. Next Steps

Now with the initial framework developed I need to create a pipeline that (1) queries their database (2) saves the query results as there is necessary target information there (3) extracts job-id from the query (4) uses the job-id to get the report for the file (5) generate feature extraction code and decide on ways to reduce dimensionality as well as possibly implement my own lexical feature creation. I suspect (5) will soak up the next chunk of my time as I also need to draw connections between their web platform and the return from the report download as there are some slight nuances to how features are described (e.g "mutex" vs "mutant").

IV. Previous Work

Since I recently decided on the project pivot, I did not collect 3 outside publications just yet. However, I did find this: Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi. 2019. Malicious URL Detection using Machine Learning: A Survey. 1, 1 (August 2019). That paper is an extremely readable and well documented approach to URL detection and will inspire the possible lexical feature creation I hope to implement in my final project. It also very well could influence my choices in feature extraction, but my hopes there lie in using all of the available information from H.A's reports. What I find to be most beneficial of Sahoo et al. is the education on the domain of URL detection.

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections IV-A–IV-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and,

conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.