Zeta Avarikioti*, Roman Brunner, Aggelos Kiayias, Roger Wattenhofer, and Dionysis Zindros

The iceberg below the surface: An analysis of dark net content

Abstract: We analyze the type of content present on the "dark web", the set of websites accessible via Tor. We create a darknet spider and crawl the whole darknet by starting from a bootstrap list and recursively following links so that the whole connected component of more than X websites and Y base URLs is explored. We publish our spider as open source software. We find that the darknet is well-connected through hub websites such as wikis and forums. We perform comprehensive measurements on the content found using machine learning to analyze and categorize the various types of content. We observe that the majority of darknet content belongs to P and Q. We close by discussing the political and ethical implications of these results.

Keywords: anonymity, tor, machine learning, scraping, spider

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

1 Introduction

Describe the scope of the project, refer to previous work (e.g., contrast with Nicolas Christin's work on Silk Road as well as the work where they ran their own relays to collect content). Clearly state our contributions. Summarize the methodology and results. Introduce visible vs invisible part of the darknet

2 Scraping the Darknet

In order to get a complete scrape of the visible part of the Darknet, we created a scraper, which follows links recursively until it does not find new links.

***Corresponding Author: Zeta Avarikioti:** Affil, E-mail: zetavar@ethz.ch
Roman Brunner: Affil, E-mail: robrunne@student.ethz.ch
Aggelos Kiayias: Affil, E-mail: aggelos.kiayias@ed.ac.uk
Roger Wattenhofer: Affil, E-mail: wattenhofer@ethz.ch
Dionysis Zindros: Affil, E-mail: dionyziz@gmail.com

2.1 Structure of the spider

The spider itself is written in Javascript and relies on NodeJS as the runtime environment. NodeJS was used since the spider is waiting most of the time for requests to return, be that to the database or to a hidden website. Therefor the NodeJS event loop came in handy and did not add a lot of overhead in development due to task or process handling. To store the data, we used PostgresDB Insert schema, which not only had to store the downloaded content but also all the links that were not yet visited. To access the Tor network we needed to add a Tor proxy in front of the spider itself.

FORMATTING To speed up the scraping process, we ran 100 concurrent requests to the Tor network. In order to prevent overload on the Tor nodes within the circuit, we used a pool of Tor of equal size, such that per request one Tor instance was used. The Tor instances were scheduled in a round robin manner

Describe the architectural decisions put in designing the spider. Mention the technologies we're using (node.js, postgres, the tor libraries). Include a diagram for our software architecture. Discuss how we bypass rate limits (e.g. by randomization of visits and by using multiple tor circuits in parallel) and what the rate limiting problems are (including rate limits by individual websites as well as by the tor circuit / network itself). How we avoid downloading illegal content (whitelists and blacklists, content types). Talk about bootstrapping the list, exploration depth. Show the numbers such as how many sites exist, how many base URLs exist, how many links exist, summarize them nicely in a table. Try to aggregate data from our database to create insights for the reader. Extract the "hubs" and state that, contrary to popular opinion, the darknet is well connected; specifically discuss which this hubs are, what their purpose is, and how many links they have for the top 10 hubs.

3 Analyzing content

Describe the types of content that we found. Talk about our methodology in how we analyzed the content us-

ing machine learning and NLP. How were the categories chosen? Discuss how we rank base URLs as well as individual paths. Show diagrams (a pie chart or bars?) with the popularity of various content categories.

4 Political and Ethical Discussion

Discuss the political implications of tor. Is it mostly illegal content? Does it matter? Discuss the ethical considerations of our research and make sure you mention that we didn't subvert the protocol itself but only collected public data.

5 Conclusion

Summarize our results and discuss directions for future work.

6 Editorial Policy

6.1 Choice of reviewers

The Editor responsible for a given area of physics, turns to experts of the subject for opinion. Research articles and communications are reviewed by minimum two reviewers, review papers by at least three.

6.2 Suggestions from authors

Authors are requested to suggest persons competent to review their manuscript. However, please note that this will be treated only as a suggestion, the final selection of reviewers is exclusively the Editor's decision. The reviewers remain anonymous in any case.

The Editor is fully responsible for decision about the manuscript. The final decision, whether to accept or reject a paper, rests with him/her. The Managing Editor only communicates the final decision and informs the author about further processing.

6.3 Revised manuscript submission

When revision of a manuscript is requested, authors should return the revised version of their manuscript as soon as possible. Prompt action may ensure fast publication, if the paper is finally accepted for publication in .... If it is the first revision of an article authors need to return their revised manuscript within 60 days. If it is the second revision authors need to return their revised manuscript within 14 days. If these deadlines are no met, and no specific arrangements for completion have been made with the Editor, the manuscript will be treated as a new one and will receive a new identification code along with a new registration date.

6.4 Final proofreading

Authors will receive a PDF file with the edited version of their manuscript for final proofreading. This is the last opportunity to view an article before its publication on the journal's web site. No changes or modifications can be introduced once it is published. Thus authors are requested to check their proof pages carefully against manuscript within 3 working days and prepare a separate document containing list of all the changes that should be introduced. Authors are sometimes asked to provide additional comments and explanations in response to remarks and queries from the language and technical editors. In case the authors do not deliver the list of corrections to proofs in the requested time the manuscript will be published as is.

6.5 Reprints

Because the journal is published in an Open Access model, and has no printed version, the authors receive no reprints.

6.6 Erratum

If any errors are detected in the published material they should be reported to the Managing Editor. The corresponding authors should send appropriate corrected material to the Managing Editor via email. This material will be considered for publication in form of erratum in the earliest available issue of ....

6.7 Copyright

All authors retain copyright, unless – due to their local circumstances – their work is not copyrighted. The non-commercial use of each article will be governed by the Creative Commons Attribution-NonCommercial-

7 Paper writing guide

7.1 Paper elements

1. title page with:
   (a) title (short title),
   (b) full name(s) of author(s),
   (c) name and address of workplace(s),
   (d) personal e-mail address(es),
2. abstract,
3. up-to five keywords,
4. text,
5. reference lists.

7.1.1 Abstract

An abstract must accompany every article. It should be a brief summary of the significant items of the main paper. An abstract should give concise information about the content of the core idea of your paper. It should be informative and not only present the general scope of the paper but also indicate the main results and conclusions. An abstract should not normally exceed 200 words. It should not contain literature citations or allusions to the tables or illustrations. All non-standard symbols and abbreviations should be defined.

In combination with the title and key-words, the abstract is an indicator of the content of the paper. Authors should remember that on-line systems rely heavily on the content of titles and abstracts to identify articles in electronic bibliographic databases and search engines. They are therefore requested to take great care in preparing these elements.

7.1.2 Text

7.1.2.1 General rules for writing

- use simple and declarative sentences, avoid long sentences, in which the meaning may be lost by complicated construction;
- be concise, avoid idle words;
- make your argumentation complete; use commonly understood terms; define all non-standard symbols and abbreviations when you introduce them;
- explain all acronyms and abbreviations when they first appear in the text;
- use all units consistently throughout the article;
- be self-critical as you review your drafts.

7.1.2.2 Structure of a paper

Research papers and review articles should follow a strict structure. Generally a standard scientific paper is divided into:

- introduction: you present the subject of your paper clearly, you indicate the scope of the subject, you present the goals of your paper and finally the organization of your paper;
- main text: you present all important elements of your scientific message;
- conclusion: you summarize your paper.

Experimental part and/or calculations should be presented in sufficient details to enable reader to repeat the original work.

7.1.2.3 Footnotes/End-notes/Acknowledgments

We encourage authors to restrict the use of footnotes. If necessary, please make end-notes rather than footnotes. Allowable footnotes/end-notes may include:

- the designation of the corresponding author of the paper;
- the current address of an author (if different from that shown in the affiliation);
- traditional footnote content.

7.1.2.4 Tables

Authors should use tables only to achieve concise presentation, or where the information cannot be given satisfactorily in other ways. Tables should be numbered consecutively using Arabic numerals and referred to in the text by number. Each table should have an explanatory caption which should be as concise as possible.

7.1.2.5 Figures

Authors may use line diagrams and photographs to illustrate theses from their text. The figures should be clear, easy to read and of good quality. Styles and fonts should match those in the main body of the article. All

Fig. 1. A figure caption should be placed **below** the figure.

Fig. 2. A figure caption for Fig. 2.

figures must be mentioned in the text in consecutive order and be numbered with Arabic numerals.

7.1.2.6 Typesetting

Type main text in roman (upright) font. The chemical symbols and compounds, units of measure, most multi-letter operators and functions should are written in roman upright as well. The variables, constants, symbols for particles, most single-letter operators, axes and planes, channels, types (e.g., n, p), bands, geometric points, angles, lines, chemical prefixes, symmetry designations, transitions, critical points, color centers, quantum-state symbols in spectroscopy, and most single-letter abbreviations should be written in roman italic. Boldface roman type is reserved for indicating vectors and in some special cases matrices.

7.1.2.7 Mathematical symbols

The multiplication signs are reserved for a vector product ($\mathbf{A} \times \mathbf{B}$) and simple dot product ($\mathbf{A} \cdot \mathbf{B}$). The only exception are numbers expressed in scientific notation (9.7×10^3 MeV).

7.1.2.8 Units

Units and dimensions should be expressed according to the metric system and SI units. This system is based on: meter (m), second (s), kilogram (kg), ampere (A), kelvin (K), mole (mol), and candela (cd). Most units are spaced off from the number, e.g. 12 mV. The only exceptions are:

$$1\%,\ 1‰,\ 1°\mathrm{C},\ 1°,\ 1',\ 1".$$

Decimal multiples or sub-multiples of units are indicated by the use of prefixes

$$\mu{=}10^{-6},\ \mathrm{m}{=}\ 10^{-3},\ \mathrm{c}{=}\ 10^{-2},\ \mathrm{d}{=}\ 10^{-1},\ \mathrm{da}{=}\ 10^{1},\\ \mathrm{h}{=}\ 10^{2},\ \mathrm{k}{=}\ 10^{3},\ \mathrm{M}{=}\ 10^{6},\ \mathrm{G}{=}\ 10^{9},\ \textit{etc.}$$

Compound units are written as

$$4221.9\ \mathrm{J\ kg^{-1}\ K^{-1}}\ \text{or}\ 4221.9\ \mathrm{J/(kg\ K)},$$

with a thin space between unit parts.

Authors should indicate precisely in the main text **where tables and figures should be inserted**, if these elements are given at the end in the original version of the manuscript (or supplied in separate files). If this information is not provided along with the manuscript, we will assume that the figures and/or tables should be insert at the closest position to first reference to them in the published paper.

7.1.2.9 Multimedia and images

Authors can attach files in most popular formats, including (for example):

- images in BMP, GIF, JPEG formats,
- multimedia files in MPEG or AVI formats.

However please keep to file types that are read by standard media players (e.g. RealPlayer, Quicktime, Windows Media Player) and/or standard office applications (Adobe Acrobat Reader, Microsoft Office etc.).

Your attachments may be accessible through links to external locations or to our internal locations (if you choose the second option, please remember to send us your attachments).

Please remember that your images, video and animation clips are intended for Internet use and we need to consider the needs of users with slow Internet connections. Please try to minimize file sizes by using a lower resolution or number of colors for images and animations (as long as the material is still clear). To help you in formatting your images (including tables and figures) or multimedia files, please submit your paper with separate attachments, which are used in your paper.

7.1.2.10 English language

Journal is published only in English. Make sure that your manuscripts are clearly and grammatically written. Please note that authors who are not native-speakers of English can be provided with help in rewriting their contribution in correct English. Try to prepare your manuscript in an easily readable style; this will help avoid severe misunderstandings which might lead to rejection of the paper.

7.1.3 Reference list

A complete reference should give the reader enough information to find the relevant article. All authors (unless there are six or more) should be named in the citation. If there are six or more, list the name of the first one followed by "et al". Please pay particular attention to spelling, capitalization and punctuation here. Completeness of references is the responsibility of the authors. A complete reference should comprise the following:

7.1.3.1 Reference to an article in a journal

Elements to cite: Author's Initials. Surname, – if more authors, see examples below, Title of journal – abbreviated according to the ISI standards[1], volume number, page or article number (year of publication). Please supply DOI or URL for e-version of the papers. See Refs. [1–8] for example.

7.1.3.2 Reference to a book

Elements to cite: Author's Initials. Surname, Title, Edition – if not the first (Publisher, Place of publication, Year of publication) [9].

7.1.3.3 Reference to a part/chapter book

Elements to cite: Author's Initials. Surname, In: Editor's Initials. Editor's Surname (Ed.), Book Title, Edition – if not the first, (Publisher, Place of publication, Year of publication) page number [10].

7.1.3.4 Reference to a preprint

Elements to cite: Author's Initials. Surname, arXiv:preprint-number and version [11, 12].

7.1.3.5 Reference to a conference proceedings

Elements to cite: Author's Initials. Surname, In: Editor's Initials. Editor's Surname (Ed.), Conference, date, place (town and country) of conference (Publisher, place of publication, year of publication) page number [13].

7.1.3.6 Reference to a thesis

Elements to cite: Author's Initials. Surname, D.Sc./Ph.D./M.Sc./B.Sc. thesis, University, (town, country, year of publication) [14].

7.1.3.7 Reference to an article in a newspaper

Elements to cite: Author's Initials. Surname, Newspaper Title, date of publication, page number [15, 16].

7.1.3.8 Reference to a patent

Elements to cite: Originator, Series designation which may include full date [17].

7.1.3.9 Reference to a standard

Elements to cite: Standard symbol and number, Title [18, 19].

Please add language of publication for materials which are not written in English. Indicate materials accepting for publications by adding "(in press)". Please avoid references to unpublished materials, private communication and web pages.

You should make sure the information is correct so that the linking reference service may link abstracts electronically. For the same reason please separate each reference from the others.

Before submitting your article, please ensure you have checked your paper for any relevant references you may have missed.

7.1.4 Submission formats

Manuscripts for ... should be submitted in the LaTeX format with figures in EPS, PDF or PNG format. Authors are strongly encouraged to register their manuscript in arXiv preprint server and submit it to our Editorial Manager using arXiv's paper ID.

7.1.5 Supplementary data

You can also submit any supplementary data files as well. These may include long tables (in HTML or plain TXT format) or movies (preferably in AVI format).

1 http://images.isiknowledge.com/WOK46/help/WOS/0-9_abrvjt.html

References

[1] A. P. Raposo, H. J. Weber, D. E. Alvarez–Castillo, M. Kirchbach, Cent. Eur. J. Phys. 5, 253 (2007)

[2] J. Barth et al. (SAPHIR Collaboration), Phys. Lett. B 572, 127 (2003)

[3] S. Chekanov et al., Eur. Phys. J. C 51, 289 (2007)

[4] K. Malarz, Postepy Fizyki 57, 235 (2006) (in Polish)

[5] G. Meng, Cent. Eur. J. Phys., DOI:10.2478/s11534-007-0038-1

[6] R. Hegselmann, U. Krause, Journal of Artificial Societies and Social Simulation (2006), http://jasss.soc.surrey.ac.uk/9/3/10.html

[7] A. Dybala, Cent. Eur. J. Chem. (in press)

[8] A. Dybala, Przeglad chemiczny (in Polish, in press)

[9] M. Lister, Fundamentals of Operating Systems, 3rd edition (Springer-Verlag, New York, 1984)

[10] C. K. Clenshaw, K. Lord, In: B. K. P. Scaife (Ed.), Studies in Numerical Analysis (Academic Press, London and New York, 1974) 95

[11] M. Majewski, K. Malarz, arXiv:cond-mat/0609635v2 [cond-mat.stat-mech]

[12] J. A. C. E. Solano, arXiv:0707.1343v1 [astro-ph]

[13] A. Kaczanowski, K. Malarz, K. Kulakowski, In: T. E. Simos (Ed.), International Conference of Computational Methods in Science and Engineering, Sep. 12-16, 2003, Kastoria, Greece (World Scientific, Singapore 2003) 258

[14] A. J. Agutter, Ph.D. thesis, Edinburgh University (Edinburgh, UK, 1995)

[15] A. Sherwin, The Times, Jul. 13, 2007, 1

[16] M. Dzierzanowski, Wprost, Jul. 8, 2007, 18 (in Polish)

[17] Philip Morris Inc., European patent application 0021165 A1, Jan. 7, 1981

[18] ISO 2108:1992, Information and documentation — International standard book numbering (ISBN)

[19] ISO/TR 9544:1988, Information processing — Computer-assisted publishing — Vocabulary