International Journal of Drug Policy xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Drug Policy

journal homepage: www.elsevier.com/locate/drugpo

Commentary

A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network"

Rasmus Munksgaard, Jakob Demant *, Gwern Branwen

Department of Sociology, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form 15 February 2016
Accepted 28 February 2016

Keywords:
Silk Road
Cryptomarkets
Methodology
Web crawling
Digital analysis
Web-o-metrics: Dark web
Reproducibility

ABSTRACT

The development of cryptomarkets has gained increasing attention from academics, including growing scientific literature on the distribution of illegal goods using cryptomarkets. Dolliver's 2015 article "Evaluating drug trafficking on the Tor Network: Silk Road 2, the Sequel" addresses this theme by evaluating drug trafficking on one of the most well-known cryptomarkets, Silk Road 2.0. The research on cryptomarkets in general—particularly in Dolliver's article—poses a number of new questions for methodologies. This commentary is structured around a replication of Dolliver's original study. The replication study is not based on Dolliver's original dataset, but on a second dataset collected applying the same methodology. We have found that the results produced by Dolliver differ greatly from our replicated study. While a margin of error is to be expected, the inconsistencies we found are too great to attribute to anything other than methodological issues. The analysis and conclusions drawn from studies using these methods are promising and insightful. However, based on the replication of Dolliver's study, we suggest that researchers using these methodologies consider and that datasets be made available for other researchers, and that methodology and dataset metrics (e.g. number of downloaded pages, error logs) are described thoroughly in the context of web-o-metrics and web crawling.

© 2016 Elsevier B.V. All rights reserved.

In October of 2013, the first cryptomarket, Silk Road, was seized by authorities. In spite of this seeming victory for law enforcement, the dark web drug economy bounced back stronger than before. In a few weeks, users had migrated to competing marketplaces, such as Sheep Marketplace and Black Market Reloaded (Van Buskirk, Roxburgh, Naicker, & Burns, 2015). In November 2013, key persons from the old SR relaunched Silk Road as Silk Road 2.0 (SR2). The apparent success was short, as the new incarnation of the infamous Dread Pirate Roberts allegedly fled from law enforcement with user funds (though these funds were later returned). Shortly after, SR2 allegedly suffered a hack, and thieves made off with 2.7 million USD (DeepDotWeb, 2014). In spite of these incidents, SR2 still played a key role in the cryptomarket economy. It operated as one of the largest cryptomarkets up until its seizure in 2014 (Soska & Christin, 2015).

Since the fall of Silk Road, the cryptomarket economy has continued to grow (Soska & Christin, 2015). As this special issue of IJDP shows, the scientific literature on cryptomarkets is growing and much of the research focuses on the distribution of drugs and other illegal goods through these markets. Consequently, this research agenda has posed a number of new questions for methodologies. Dolliver's (2015a) article "Evaluating drug trafficking on the Tor Network: Silk Road 2, the Sequel" was the first study of drug trafficking on Silk Road 2.0. It was subsequently met with criticism from other researchers (Aldridge & Décary-Hétu, 2015; Van Buskirk et al., 2015). The following replication of and commentary on Dolliver's (2015a) study addresses a number of these new questions about methods based on a replication of key findings presented in the article. We find it relevant to address new questions in order to refine the methods in this novel field. The use of a more refined methodological framework will help ensure that findings of distribution of drugs on the cryptomarkets are accurate and that forthcoming research does not rely on misleading results that may also have consequences outside of academia, especially in terms of policing and public discourse.

Cryptomarkets

Martin (2014, p. 356) defines the cryptomarket as an "online forum where goods and services are exchanged between parties

* Corresponding author at: Department of Sociology, University of Copenhagen, Øster Farimagsgade 5, Bld. 16, DK-1014 Copenhagen K, Denmark.
Tel.: +45 35321584; mobile: +45 81742074.
E-mail address: jd@soc.ku.dk (J. Demant).
URL: http://www.soc.ku.dk/english

http://dx.doi.org/10.1016/j.drugpo.2016.02.027
0955-3959/© 2016 Elsevier B.V. All rights reserved.

Please cite this article in press as: Munksgaard, R., et al. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy* (2016), http://dx.doi.org/10.1016/j.drugpo.2016.02.027

who use digital encryption to conceal their identities." Essentially cryptomarkets operate as websites similar to eBay with a range of privacy and security-enhancing technologies (Barratt, 2012). First, they are located on the dark web (typically on the Tor network) as *hidden services*. Therefore, it is impossible to locate the server using traditional means and the sites can therefore operate without the risk of shutdowns due to law enforcement intervention. This requires the user to download and use special software (e.g. Tor) to access the site. Typically, the cryptocurrency bitcoin is used for transactions, as this allows a higher degree of anonymity than other means of transactions (e.g. Western Union, credit cards, PayPal).

The similarity between cryptomarkets and eBay-like websites is to be found in the decentralized structure in which many different vendors operate on the site, with the site only providing the means for transactions and review system. The review system, in which users post a rating of the service provided (e.g. product, shipping), provides a good measure of the economics of cryptomarkets and has been used in a number of studies of the cryptomarket economy (Aldridge & Décary-Hétu, 2014; Christin, 2013; Dolliver 2015a, 2015b; Soska & Christin, 2015). Thus, due to the public nature of cryptomarkets, it is possible to ascertain information on the scale of the economy.

In spite of some success by law enforcement to seize sites and arrest operators (Europol, 2014; Van Buskirk, Roxburgh, Farrell, & Burns, 2014) the cryptomarket economy has grown rapidly since 2011 in which the first site, Silk Road, launched. Christin (2013) estimated that Silk Road grossed 1.2 million USD monthly in 2012 and Soska and Christin (2015) found that daily revenues on cryptomarkets in 2014 reached as high as 600,000 USD daily.

Web-o-metrics and web crawling

Dolliver's study is based on what is termed "web-o-metrics." Björneborn and Ingwersen (2004 p. 1216) defines web-o-metrics as "(1) Web page content analysis; (2) Web link structure analysis; (3) Web usage analysis (including log files of users' searching and browsing behavior); (4) Web technology analysis (including search engine performance)." This broad definition draws upon bibliometric and scientometric applications of online data. More broadly, web-o-metrics can be thought of as "the quantitative study of Web-related phenomena" (Thelwall et al., 2005). The common use of online data for cryptomarket research consists of downloading, or "mirroring," an entire website (Aldridge & Décary-Hétu, 2014; Dolliver 2015a, 2015b; Soska & Christin, 2015). This form of data collection can more specifically be referred to as web crawling (Olston & Najork, 2010, p. 176). Web crawling is, for example, employed by search engines that crawl and index pages that may be queried by a user. As web crawling is the central method for the analysis of data within this new field of cryptomarket research, we will focus specifically on this aspect of the methods.

When deploying, the web crawler first visits and downloads a page. From this page, links to other pages are downloaded. These are then subsequently crawled and downloaded. Depending on the instructions the crawler is given, it may download the entire website or crawl to a certain depth. For example, the crawler may be instructed to download and follow the links available on the first page but not to follow links on subsequent pages. When the available pages on a website are downloaded (which are identified by following links available on it) the researcher is in possession of a "mirror" of the website. This mirror may be browsed on a personal computer or utilized for data extraction using programming. The mirror is a copy of the website on the date that the individual pages were downloaded, and it is therefore a static image in time. The application collecting the data may be designed by a researcher (Aldridge & Décary-Hétu, 2014; Soska & Christin, 2015) or a pre-packaged solution such as HTTrack (Christin, 2013). The researcher defines some rules—for example, to exclude images from the download (Christin, 2013; Dolliver 2015a, 2015b) or a time period before each request. The web crawls, or mirrors, of Silk Road 2.0 that we utilize for this replication are downloaded using *wget*, an application for website mirroring.

Dolliver (2015) study

Dolliver crawled and downloaded SR2 in August and September 2014. The study produced the following findings: Of the 1834 unique items for sale, 348 were drug items, 145 distinct vendors shipped drugs from 19 countries, and the U.S. was the primary origin and destination country for drugs. Dolliver further concluded that SR2 did not primarily deal in drugs but in eBooks and other non-drug items and that drug-related items only accounted for 1% of the number of transactions (Dolliver, 2015b, p. 1117). After publication of the article, criticisms of the results were voiced. Aldridge and Décary-Hétu (2015) describe the results as expressing a "radical discontinuity" between SR2 and its predecessor, arguing that the data seemed flawed, was not in agreement with measurements by other researchers, and that it should have been subject to extensive validity checks given this radical discontinuity. This argument is supported by the measurements of Van Buskirk et al. (2015) who found that in this period, SR2 offered 9103 drug items for sale from 519 vendors. Dolliver (2015a) subsequently responded that the data had been subject to extensive validity checks, and that Aldridge and Décary-Hétu (2015) relied on data that had not been subject to peer-review. Based upon the discussion following the publication of the article, we attempt to reproduce the most basic findings of the study: The number of vendors and items for sale.

The methodology of Dolliver's study is described as:

> To capture publicly available data from Silk Road 2, HTTrack (a free, offline browser) was employed in order to obtain a mirrored copy of the website maintaining the site's original link structure. [...] This was fairly time-consuming as the crawler had to be pointed multiple levels deep before it was able to reach individual advertisements; simply instructing the software to mirror the site was not enough. HTTrack was instructed to only mirror HTML text and not to include images (capturing images substantially slows down the progress of the crawling software and produces unnecessary data). On September 3rd, 2014 a complete crawl of Silk Road 2 was conducted, which required approximately 7 h and 18 min and produced a file size of 84.4 MB. (Dolliver, 2015b, p. 1117)

This description of the crawls clarifies a number of points. We do, however, find that it also lacks some information, which raises the following questions: (1) How many unique pages were downloaded? (2) How many errors were produced (e.g. "Page not found")? (3) How were these data extracted? (4) How were the frequent disruptions in the availability of the site handled?

In particular, we would like to see a description of handling errors (pages not found due to server downtime), dataset size metrics, and data extraction. This information is vital in order to evaluate and review this methodology, especially given this paper's surprising finding that 19% of the items on SR2 were drug-related and that these only constituted 1% of the number of transactions. This absent information, especially regarding how these data were checked for inconsistencies, makes it difficult to understand how these results were produced and what considerations should be taken in evaluating them. It might not be necessary to provide answers to all of these questions, as it is not a "methods-centric" paper (Dolliver, 2015a); however, as the results

Please cite this article in press as: Munksgaard, R., et al. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy* (2016), http://dx.doi.org/10.1016/j.drugpo.2016.02.027

are highly surprising, a more thorough description of methodology is vital. In particular, it is unclear what is meant by the paragraph on tuning the crawler to go to a certain "multiple levels deep" (Dolliver, 2015b, p. 1117). The crawler should download all pages of interest on the website to ensure that all listings of products are downloaded.

Replication of study

It is not fully clear what data Dolliver bases the analysis on. It is unclear whether it was based on crawls in both August and September or only the full crawl in September. We therefore assume that it is only based on the crawl on September 3 and that data from the test crawls were not included. Dolliver has not been able to provide us the original dataset due to a non-disclosure agreement, and we have therefore replicated the study using a publicly available dataset gathered by independent researcher Gwern Branwen which is part of a larger collection of crypto-market-related datasets contributed to by a number of individuals (Branwen et al., 2015). Though Dolliver (2015a) and the research team has argued that the data collected by Branwen is "not a credible source to use as evidence to discredit peer-reviewed work," because of the non-disclosure agreement, it is the only way to publicly assess the validity of the study. While the results presented by Dolliver may have been subjected to peer-review (Dolliver, 2015a), the data is not available for public scrutiny, and it is not clear whether reviewers have had access to the data (Dolliver, 2015a). Branwen et al.'s (2015) data, however, is publicly available with configuration files, description of its shortcomings, and contributions from other researchers.

We acknowledge that our replication is based on data not entirely similar to that used by Dolliver. However, we believe conclusions can still be drawn from this replication. Ideally, we measure the same phenomena, SR2 from August to September 2014, using the same methodology. As such, the results should not differ to a great extent. We reproduced Dolliver's study based on nine crawls collected between August 4 and September 10 by Branwen. All of these crawls are partial, meaning that the whole website has not been downloaded or mirrored, which we attribute to the frequent downtime of SR2. Our data should therefore be considered as large samples of varying size over a given period, rather than absolute images of the site, whereas Dolliver's study is presented as a complete crawl (Dolliver, 2015b, p. 1116). We further note that errors were not logged during the crawls we study.

Dolliver's study is primarily concerned with the different types and numbers of items for sale on SR2. In order to ensure accuracy, it is important to critically review the material that is crawled. In our review of the data used for the analysis, we noticed that some items for sale on SR2 had very similar URLs; we found these items to be highly similar (identical product descriptions and prices), though differing in regards to certain properties (shipping destination and reviews). For example, we observed a URL ending with "*/items/1-gram-organic-og-kush-chocolate-kush-ships-from-finland*" and one with "*/items/1-gram-organic-og-kush-chocolate-kush-ships-from-finland-ananas_xpress*". These items should be considered different listings (because the website presents them as such) and were therefore not removed from the dataset.

As the crawls are partial, this replication study follows a methodology that allows for extracting as much data as possible. In the case of a complete crawl, we could have identified the number of items for sale simply by counting pages matching a particular URL structure (e.g. "/items/1-gr-cannabis" designating an item or "/users/swazibudbud888" designating a vendor). However, as the dataset is partial, we searched all downloaded pages, including both listings by categories and sub-categories, vendor profiles, and so forth to produce a more complete analysis. To do this, we first identify the total number of pages of interest (pages containing raw HTML). From these, we extract all links present on each page. Finally, we identify the number of unique vendors and items by counting identified links referring to pages for these vendors and items.

Note that we do not merely count the number of downloaded pages matching the URL structure, but the actual links present on all pages. This gives a fuller picture of available items while simultaneously providing an idea as to how complete a crawl is. This method is comparing the number of downloaded pages to the number of links found in those that have not been crawled. Table 1 describes the crawls. It does not provide an account for pages identified in the pages that have not been downloaded. Therefore, it is likely that even more links, and thus vendors and items, are present though not identified because of the partialness of the crawls.

To properly assess the validity of the results, we also applied other methods of measuring the number of items. We extracted items listed under categories, and we identified items and vendors through successfully downloaded pages using the URL ("/items/", "/users/"). Both of these showed roughly the same results, though identified fewer items and vendors, which we attribute to the partialness of the crawl.

Results

The replication based on the partial crawls described above does not find numbers of items for sale on SR2 that resemble those presented by Dolliver. Both in August and September, we find that the number of items and vendors is much higher than those measured by Dolliver. In September, at which time Dolliver conducted a "complete crawl" of Silk Road (Dolliver, 2015b, p. 1116), we find a much higher number of vendors and items both on September 2 and 10. Combining these data, we find in total 581 unique vendors and 12,259 unique items for sale in these two crawls, and in total, we observe 18,513 unique items in the dataset. These are the items and vendors which appear in either one or more crawls. None of these are complete crawls of SR2. Yet, in spite of this, the number of items and vendors found is much greater

Table 1
Description of crawls.

Date	Links found in downloaded pages	Downloaded pages	Crawl completeness	Items	Vendors
2014-08-04	206,866	63,522	30.71%	12,558	560
2014-08-09	198,402	84,266	42.47%	9283	463
2014-08-11	25,116	3539	14.09%	5588	414
2014-08-17	305,838	103,854	33.96%	13,987	561
2014-08-23	200,711	115,976	57.78%	9072	478
2014-08-27	226,997	100,546	44.29%	14,534	560
2014-08-30	171,256	67,507	39.42%	13,701	544
2014-09-02	59,779	13,093	21.90%	9455	517
2014-09-10	108,269	60,209	55.61%	8682	489

Please cite this article in press as: Munksgaard, R., et al. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy* (2016), http://dx.doi.org/10.1016/j.drugpo.2016.02.027

than Dolliver's findings. Thus, though our data should identify fewer items and vendors than Dolliver because the crawls are partial, at no point do we observe anything that resembles the results presented by Dolliver.

Dolliver argues that the number of available items on SR2 was lower than advertised by SR2 administrators in the sidebar. When logged in on SR2, users would be able to view the number of items in a particular category. According to Dolliver, these were inflated. To review whether this was the case, we can compare the total number of unique items observed in the period (18,513) to the number of items listed in the sidebar. These are the items that have appeared at least once in a crawl. The lowest number we observed in the sidebar (the sum of all items in categories, excluding subcategories) at the start of a crawl was 15,723, with a maximum of 16,549. From August to September, we observed more unique items (18,513) in nine partial crawls than SR2 listed during this period in its sidebar. However, we know from previous studies of the original SR that listings (on average) were taken down after a number of weeks (Christin, 2013). This means that this measure is not precise, as we are comparing the number of items observed throughout the period with the number of items listed in the sidebar. With only partial crawls at our disposal, it is our best estimate. A preferred strategy would have been to compare a *complete crawl* to the number of listings in the sidebar. As such, we cannot dismiss the conclusion that SR2 manipulated the number of items for sale based on only partial crawls. To fully replicate this result, we would need a complete crawl of the site. However, in reviewing our data, we found no indication that SR2 inflated numbers to the degree Dolliver (2015b, p. 1117) found. It should be noted, however, that an unintended inflation of listings may have occurred in the form of the almost identical listings we found. This means that some items may have been listed multiple times, thus inflating the number of listings.

Discussion

The findings from the replication of Dolliver's study suggest that the results are misleading due to incomplete crawls. Dolliver (2015a) has refuted this, arguing that the data were subjected to validity checks, but our results are in agreement with the criticism expressed (Aldridge & Décary-Hétu, 2015; Van Buskirk et al., 2015) and the most recent research (Soska & Christin, 2015). We stress that we have only analyzed the *number of items and vendors* and have not reviewed the other findings in Dolliver's paper. It is possible that the conclusions reached by Dolliver are correct, though evidence points to the data being flawed. As pointed out by critics (Aldridge & Décary-Hétu, 2015; Van Buskirk et al., 2015), shown in recent research (Soska & Christin, 2015), and shown in our replication, the results of Dolliver's study are in conflict with all other measurements of SR2.

This could be the result of the crawler being bogged down in the eBooks category, which, at the time, was overflowing with listings by one vendor. If the crawler then encountered an error page attempting to download listings in drug categories (e.g. "/categories/drugs-stimulants") containing links to items in this particular category, these would not be downloaded. This would have severely skewed the data towards the finding that SR2 did not deal primarily in drugs, which is the most controversial claim of Dolliver's study. Whether a crawl is incomplete due to website outages is simple to determine by reviewing error pages. If a web server is unavailable or experiences problems, it will produce an error page, the most well-known being "404 – Page not found." Tools for crawling and downloading a site like wget and HTTrack (used by Dolliver) do not, by default, log these events; instead, they skip the pages. For this replication we used crawls conducted by Branwen which did not incorporate error logging. While the data was still useful, an approach to web crawling, which we suggest to other researchers, is to log errors and assess the quality of the collected data. For example, with logged errors, it is simple to produce a table of HTTP responses. Another way to assess the quality of a crawl is to examine the links that have not been downloaded. We would argue that if Dolliver had extracted all of the links from SR2 and then compared them to the downloaded pages on which the study is based (as we present in Table 1 for the replication study), it would have shown that the crawl was (as we concluded) only partial.

It is admirable that researchers outside computer and information science use new tools such as web crawling for their research. These tools allow us to follow a number of new research agendas within social science as evidenced in the growing literature on cryptomarkets utilizing these methodologies. The novelty of these methods within social science does, however, call for a discussion of methods. Soska and Christin (2015) provide a good discussion of the difficulties of web crawling. Such refinement in methods and their descriptions would make it possible for other researchers to evaluate (or reproduce) results. We believe it is vital that the research community not omit methodological details, especially when we move over to a new set of tools. The magnitude of Soska and Christin's study, however, does require a thorough approach. For small-scale studies utilizing web crawling, we suggest four measures for quality assessment inspired by this replication study. We believe that these four measures would have served to avoid flawed results from web crawling methodologies and suggest that the following considerations be undertaken in future studies using web crawling:

(1) **Error logging and quality assessment:** When crawling websites, researchers should examine the quality of their data. For example, cryptomarkets will often experience downtime, which can halt a crawl. If the errors encountered by the crawler (e.g., "Page not found") are logged, researchers can examine the number of pages that were not downloaded. These can be downloaded later when the site is online again. If errors are not logged, an alternative approach is to identify all internal links in the downloaded pages and compare them to the list of downloaded pages. This would identify pages that have not been downloaded.
(2) **Validity checks of downloaded content:** Downloaded content should be subject to qualitative assessment to ensure the validity. As an example, when crawling cryptomarkets, the crawler may suddenly log out due to technical issues, thus leading it to download pages that simply prompt the crawler to log in. However, it is not certain that the crawler will detect this. The researcher, however, can easily assess whether this is the case by identifying all of these pages. For example, the researcher may search the pages for the text string "Please log in." Alternatively, custom-built crawlers may alert the operator upon login expiration, HTTP errors, or other errors.
(3) **Peer-reviewed quality assessment:** Peer review could be strengthened by supporting open access to the dataset. While it can be preferable not to share entire datasets for ethical reasons (Décary-Hétu & Aldridge, 2015, p. 132), we suggest at least providing an error log (e.g. a table of HTTP responses), the number of downloaded pages (for comparison against other crawls), and the configuration of the tools employed. In a case such as this, where several people studied and crawled markets, it may also prove fruitful to consult both colleagues in academia and community insiders to ensure the validity of the results.
(4) **Methodological skepticism:** When the conclusions reached are surprising (such as SR2 primarily selling eBooks), this should spur discussion and critical examination of the data. In particular, when employing new tools and methodologies such

Please cite this article in press as: Munksgaard, R., et al. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy* (2016), http://dx.doi.org/10.1016/j.drugpo.2016.02.027

as web crawling, both skepticism and some digital literacy are critical. Whether results are surprising or expected, however, the data should be thoroughly reviewed. Qualitative assessment of data is particularly fruitful, as our example with the seemingly duplicate URLs and Dolliver (2015b, p. 1116) observation of miscategorized items both show. These are observations that would not have appeared without reviewing data qualitatively.

Conclusion

We have found that the results produced by Dolliver differ greatly from our replicated study. Our replication is based on a dataset collected at the same time, though it is partial. Therefore, it should be the case that we find fewer items and vendors as the data is partial and Dolliver analyzes a complete crawl. Contrary to this, we observe a much higher number of both vendors and items in our dataset. Whereas Dolliver found 1834 items for sale and 145 vendors on SR2 in the beginning of September 2014, we found 12,259 items for sale and 581 vendors. This discrepancy is well outside what should be considered a reasonable margin of error. While it is to be expected that results would differ in the exact numbers, as listings would be pulled down and pages may not have been crawled, the inconsistencies we found are too great to attribute to anything other than methodological issues. Our results and the criticism levelled at Dolliver (Aldridge & Décary-Hétu, 2015; Van Buskirk et al., 2015) suggest that the results of the study are flawed. However, as the original dataset, which consists of data gathered from an easily accessible website, remains under a non-disclosure-agreement, we cannot conclusively state whether the results are flawed. Based on the replication study, we suggest that the methodology should be described in detail incorporating (1) error logging and quality assessment, (2) validity checks of downloaded content, (3) peer-reviewed quality assessment, and (4) methodological skepticism. Furthermore, we would like to call for a greater openness in terms of reproducible codes and datasets. This could include generating lists of downloaded pages, configuration settings, or error logs readily available in case the data is sensitive or under a non-disclosure-agreement. In this spirit, we invite our colleagues to reproduce our findings. Our code and datasets for reproduction are available upon contacting the authors.

Conflict of interest statement

The authors declare no conflicts of interest.

References

Aldridge, J., & Décary-Hétu, D. (2015). A response to Dolliver's "Evaluating drug trafficking on the Tor Network: Silk Road 2, the sequel". *The International Journal on Drug Policy, 26*(11), 1124–1125.

Aldridge, J., & Décary-Hétu, D. (2014). *Not an 'Ebay for Drugs': The Cryptomarket 'Silk Road' as a paradigm shifting criminal innovation*. Available at SSRN 2436643.

Barratt, M. J. (2012). Silk Road: eBay for drugs. *Addiction, 107*(3), 683.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216–1227.

Branwen, G., Christin, N., Décary-Hétu, D., Andersen, R. M., StExo, Presidente, E., et al. (2015). *Dark net market archives, 2011–2015*. Retrieved from www.gwern.net/Black-market%20archives

Christin, N. (2013). Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. *Proceedings of the 22nd International Conference on World Wide Web*.

Décary-Hétu, D., & Aldridge, J. (2015). Sifting through the net: Monitoring of online offenders by researchers. *The European Review of Organized Crime, 2*(2), 122–141.

DeepDotWeb (2014). *Silk Road 2.0 Hacked, All Bitcoins Stolen – $2,7 Million*. DeepDotWeb Retrieved from http://www.deepdotweb.com/2014/02/13/silk-road-2-hacked-bitcoins-stolen-unknown-amount/

Dolliver, D. S. (2015a). A rejoinder to authors: Data collection on Tor. *International Journal of Drug Policy, 26*(11), 1128–1129.

Dolliver, D. S. (2015b). Evaluating drug trafficking on the Tor Network: Silk Road 2: The sequel. *International Journal of Drug Policy, 26*(11), 1113–1123.

Europol (2014). *Global action against dark market on Tor network*. Retrieved from https://www.europol.europa.eu/content/global-action-against-dark-markets-tor-network

Martin, J. (2014). Lost on the Silk Road: Online drug distribution and the 'cryptomarket'. *Criminology and Criminal Justice, 14*(3), 351–367.

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval, 4*(3), 175–246.

Soska, K., & Christin, N. (2015). Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. *24th USENIX Security Symposium (USENIX Security 15)*. USENIX Association.

Thelwall, M. et al. (2005). Webometrics. *Annual Review of Information Science and Technology, 39*(1), 81–135.

Van Buskirk, J., Roxburgh, A., Farrell, M., & Burns, L. (2014). The closure of the Silk Road: What has this meant for online drug trading? *Addiction, 109*(4), 517–518.

Van Buskirk, J., Roxburgh, A., Naicker, S., & Burns, L. (2015). A response to Dolliver's "Evaluating drug trafficking on the Tor network". *International Journal of Drug Policy, 26*(11), 1126–1127.

Please cite this article in press as: Munksgaard, R., et al. A replication and methodological critique of the study "Evaluating drug trafficking on the Tor Network". *International Journal of Drug Policy* (2016), http://dx.doi.org/10.1016/j.drugpo.2016.02.027