The Importance of Identifying and Accommodating E-Resource Usage Data for the Presence of Outliers.

Alain R. Lamothe

ABSTRACT

This article presents the results of a quantitative analysis examining the effects of abnormal and extreme values on e-journal usage statistics. Detailed are the step-by-step procedures designed specifically to identify and remove these values, termed outliers*. By greatly deviating from other values in a sample, outliers distort and contaminate that data. Between 2010 and 2011, e-journal usage at Laurentian University's J. N. Desmarais Library spiked because of illegal downloading. The identification and removal of outliers had a noticeable effect on e-journal usage levels. They represented more than 100,000 erroneous articles downloaded in 2010 and nearly 200,000 erroneous downloaded in 2011.*

INTRODUCTION

This article was written with two purposes in mind. First, it presents and discusses the results of a quantitative analysis that assessed how outlier values can influence usage statistics. Second, and more important, it details the step-by-step procedures designed specifically to identify outliers and reduce their impact on the data.

Outliers are abnormal values that result in the corruption or contamination of data by artificially increasing or reducing average values.[1] An outlier can thus be defined as a value that appears to greatly deviate from all other values in the sample,[2] as an observation that seems to be inconsistent with the rest of the dataset,[3] or as a very extreme observation requiring special attention because of potential impacts it may have on a summary of the data.[4] They occur frequently in measurement data.[5]

The presence of outliers in usage data can significantly and negatively impact libraries. For libraries having e-resource subscription pricing based on usage statistics, the presence of outliers can contribute to unwarranted increases in subscription rates. For libraries that integrate e-resource usage statistics into their collection development and management practices, the presence of outliers can affect decisions on purchase, retention, or elimination of particular e-resources. Evaluators can be fooled into thinking that a particular e-resource is heavily used and must be kept. Further, the presence of extreme outliers is often the result of a malicious system

Alain R. Lamothe (alomothe@laurentian.ca) is Associate Librarian, Department of Library and Archives, Laurentian University, Sudbury, Ontario, Canada.

intrusion,[6] as was experienced by the J. N. Desmarais Library of Laurentian University in Sudbury, Ontario, Canada.[7]

Between June 2010 and May 2011, e-journal usage at the J. N. Desmarais Library spiked after a four-year period of stable annual usage levels.[8] Between 2006 and 2010, the total number of full-text articles downloaded from the library's e-journal collection ranged between 640,000 and 720,000 annually, with an average of 700,000 articles downloaded per year. But in 2010 that number dramatically increased to more than 857,000 full-text articles downloaded. This was followed by an additional 870,000 full-text articles downloaded in 2011. Then, as suddenly and inexplicably as the increase had occurred, usage levels returned to the same quantities recorded in the years prior to 2010. A total of 716,000 full-text articles were downloaded in 2012.

During this period of spiking usage the library received notifications and warnings from certain e-journal vendors of abnormally large numbers of full-text articles being downloaded over a relatively short period of time from the Laurentian University EZProxy server's IP address. This level of usage was a breach of license agreements. These vendors then proceeded to temporarily block Laurentian University's EZProxy access until they obtained assurances from the university that the offending accounts were no longer active. This action on the vendors' part prevented any further suspected illegal downloading from occurring but also barred Laurentian University students, staff, and faculty from authorized off-campus access. But not all vendors operated in this fashion and, unknown to the library at the time, full-text articles continued to be downloaded from other vendor sites in excessive amounts. Either they were not monitoring excessive usage or they did not have the technical means to do so. Regardless, in some cases certain e-journal titles recorded downloads thousands of times higher than normal. In some cases dozens of articles were being downloaded in seconds. The situation continued until late spring 2011, at which point it was discovered that confidential proxy account login information had been posted illegally on the web. With the login information of all compromised accounts now available, proxy managers were able to block their access at once, thereby ending the period of illegal downloading of Laurentian University licensed material.

Web robots were suspected to have been involved. Web robots, also referred to as Internet bots or WWW robots, are automated software applications that run tasks on the web much as search engines do.[9] They send requests to web servers to procure resources.[10] Some robots are developed with malicious intent and are designed to download entire websites for the purpose of copying the site,[11] for autonomous logins to send spam,[12] or for autonomous logins to steal confidential or copyright protected material.[13] Web robots specifically designed for the illegal procurement of copyright protected content are obviously of particular concern for libraries.

Unlawful downloading of full-text content occurs for many reasons. Studies have clearly demonstrated that excessively high prices of digital content is a major drive for illegal downloads.[14] Misunderstanding and misinterpretation of copyright laws in addition to

unfamiliarity with and general apathy toward these same copyright laws further contribute to unlawful downloading of protected material.[15]

Many students are unaware that the transmission of downloaded articles violates copyright laws and license agreements and often misunderstand the fair use aspect of copyright as meaning that the acquisition and distribution of licensed content for the purpose of education is allowed.[16] In the minds of these students, distribution is permitted provided it is not for profit. Librarians have also reported students systematically downloading all articles from recent journal issues not for the purpose of distribution or sale but rather to build their own personal collection.[17] They are more concerned with obtaining resources quickly and completely rather than legally.[18]

Aggravating the situation are students who firmly believe that by paying tuition they have permission to do whatever they wish with their institutions' e-resources.[19] Some of these same students even use web robots to download as much as possible thereby saving them time and energy.[20] They consider the downloaded item as their personal property. In fact, Calluzzo and Cante found that students displayed an ethical sense to personal property but became neutral if the property belonged to an enterprise.[21] And Solomon and O'Brien found that 71 percent of students believed illegal copying to be a socially and ethically acceptable behavior.[22]

The J. N. Desmarais Library integrates e-resource usage into its collection development policy. As stated in the library's Collection Development Policy, "if the cost-per-use of an online resource is greater than the cost of an interlibrary loan for three consecutive years, this resource will be reviewed for cancellation."[23] In fact, this practice has been enforced for the past several years and has saved the library a considerable sum of money.[24] For this reason, it is extremely important not to assume the accuracy of usage values without carefully examining the data. The artificial inflation of usage numbers could substantially cost the library if it was believed that an e-resource was beginning to experience an improvement in usage when, in actuality, it was not the case. The decision to keep this resource could cost the library tens of thousands of dollars before it was realized that the high number of searches or downloads recorded were not reflective of actual usage but were rather the result of data recording errors or illegal activity.

Regrettably, libraries will continue to deal with the consequences of copyright infringement, even if the library itself is not at fault. It is, however, important to recognize and understand that publishers are businesses and like any business, expect financial gain.[25] Even though e-resource piracy is currently very small, the risk of it becoming the single greatest threat to the industry is quite real. Both music and film industries have been greatly affected by piracy for nearly two decades, and everyone witnessed the damaging effect it had. Publishers have learned from this and will not allow it to happen to them as well.[26] Unfortunately for all parties involved, the nature of e-resources has made them extremely easy to copy.[27]

METHOD

The following methodology will detail the step-by-step procedures to identify and deal with suspected outliers. All data manipulation and calculations were executed in Microsoft Excel for Mac 2011 (version 14.3.2). All tables and figures were generated using the same version of Excel.

The first step is to identify suspected outliers by visually examining an entire usage dataset. A dataset is defined as a collection of related data corresponding to the contents of a single database table in which each column represents a particular variable and each row, a given member of the dataset in question.[28] For this reason, the term dataset will be referred to in this paper as a grouping of data from any single spreadsheet. Each spreadsheet contains the number of full-text articles downloaded per year per vendor.

Each dataset was downloaded from vendors' sites as JR1 COUNTER-compliant reports, which detail the number of successful full-text articles downloaded per month and per journal for a given year. All vendors provided JR1 COUNTER-compliant reports that were downloaded as Excel spreadsheets. Each spreadsheet, or dataset, contained the list of e-journal title and the number of articles downloaded for each title per month (see table 1). Each dataset was then visually inspected in its entirety for suspected outliers.

	January	February	March	April	May	June	July	August	September	October	November	December
Polymer	12	15	26	33	38	64	39	5	13	**15,123**	109	44
Surface and Coatings Technology	3	1	2	1	22	17	17	0	12	**3,771**	**5,428**	601
International Journal of Radiation Oncology	11	18	35	22	17	**6,436**	176	13	25	29	24	19
Journal of Catalysis	0	1	5	1	2	2	16	4	0	2	**6,693**	1

Table 1. Sample from a 2010 JR1 COUNTER-Compliant Report Indicating the Number of Articles Downloaded per Journal Over a Twelve-Month Period. Suspected Outliers Are Highlighted in Bold.

Since it was impractical to include the entire spreadsheet, table 1 provides an excerpt from a 2010 JR1 COUNTER-Compliant report containing five suspected outliers that have been marked for identification. The suspected outliers are highlighted in bold. The first of these extreme values belongs to the title *Polymer* and was recorded in October. Compared to the other values for *Polymer*, it stands out dramatically at 15,123 articles downloaded. The second and third extreme values belong to *Surface and Coatings Technology* and are recorded for the months of October (3,771 downloads) and November (5,428 downloads). The fourth is the 6,436 articles downloaded in June from *International Journal of Radiation Oncology* and the fifth from *Journal of Catalysis* in November (6,693 downloads). These five values greatly deviate from the other values recorded

for each e-journal title. For *Polymer*, the next highest value is 109 articles downloaded in November 2010, making the suspected outlier almost 14,000 percent greater.

Now that the suspected outliers have been identified, they must be compared quantitatively to the rest of the values recorded for their corresponding titles and only for their corresponding titles. For example, to test the probability that the value of 15,123 downloads recorded in October 2010 for *Polymer* is indeed an outlier, the comparison must include all other 2010 *Polymer* monthly values plus all other available *Polymer* values. This is achieved by copying all 2010 *Polymer* monthly values into a separate blank spreadsheet and then adding all other *Polymer* monthly values from all other available years to that same spreadsheet (see table 2). This new spreadsheet can be labeled Dataset 2, with Dataset 1 being the original JR1 report downloaded from the vendor. Suspected usage outliers from an e-journal need to be compared to other usage values of that particular title because each e-journal tends to be used differently. It would be inaccurate to test for an outlier by comparing it to the values of all other e-journals included in a collection and would be like comparing apples to oranges.

		January	February	March	April	May	June	July	August	September	October	November	December
Polymer	**2009**	27	14	35	22	15	28	24	19	11	8	13	7
Polymer	**2010**	12	15	26	33	38	64	39	5	13	**15,123**	109	44
Polymer	**2011**	113	159	638	345	52	57	94	70	39	36	221	65
Polymer	**2012**	130	4	98	24	27	18	13	16	18	25	9	5

Table 2. Combining *Polymer*'s Usage Values from all Available JR1 COUNTER-Compliant Reports. The Suspected Outlier is Highlighted in Bold.

Table 2 provides the number of articles downloaded for the title *Polymer* over a four-year period. These were the only JR1 reports available from the vendor. The suspected outlier is highlighted in bold. When visually comparing the suspected outlier of 15,123 downloads to the rest of the values in Dataset 2, it again appears to be an extreme. The next highest value being 638 articles downloaded during March 2011, making the suspected outlier 2,200 percent greater than the next highest value in the dataset. All further outlier testing and accommodating will be performed on this table.

The Dixon Q Test was chosen to test for outliers. It is simple to use and designed to test for a small number of outliers in a dataset.[29] The Q value is calculated by measuring the difference in the gap between the suspected outlier and the next value over the range of values in the dataset (e.g., outlier—next value/largest—smallest). The gap is the absolute difference between the outlier and the closest number to it.

To facilitate the calculation, the data should be arranged in order of increasing value with the smallest value at the front of the sequence and the largest value at the end of the sequence. For example, using the data in table 2 each value is be arranged beginning with $4, 5, 5, 7, \ldots, 345, 638,$

and finally ending with 15,123. The calculation would thus be (15,123−638) / (15,123−4) = 0.9581. The calculated Q value will also be represented by the symbol of Q_{value} from this point onward, making $Q_{value} = 0.9581$.

The next step is to compare the calculated Q_{value} to the critical values for Q determined by Verma and Quiroz-Ruiz.[30] Critical values correspond to a particular significance level and represent cut-off values that lead to the acceptance or rejection of a null hypothesis.[31] The null hypothesis refers to the position in which there is no statistically significant relationship between two variables.[32] The alternate hypothesis would thus be the existence of a relationship between two variables.[33] If the calculated value is less than the critical value, the null hypothesis is accepted.[34] On the other hand, if the calculated value is greater than the critical value, the null hypothesis is rejected.[35] If the null hypothesis is rejected, then the alternate hypothesis must be accepted. Here, the null hypothesis can be stated as "the suspected outlier *is not* an outlier." The alternate hypothesis can then be stated as "the suspected outlier *is* an outlier." Therefore if the null hypothesis is rejected. then the suspected outlier is to be considered, in fact, to be an outlier.

Verma and Quiroz-Ruiz have calculated the critical value for Q for a sample size of 48 and at a 95 percent confidence level to be $Q_{critical} = 0.2241$.[36] Although operating at a 99 percent confidence level is a more conservative approach, it increases the likelihood of retaining a value that contains an error.[37] Operating at a 95 percent confidence level provides a reasonable compromise.[38] If the calculated value is greater than the critical value, then the suspected outlier is confirmed to be an outlier. Therefore, testing for the suspected outlier of 15,124, the Q value was calculated to be $Q_{value} = 0.9581$. With $Q_{0.9581} > Q_{0.2241,}$ the null hypothesis is rejected and it must be accepted that 15,123 is an outlier.

Once it is determined with statistical certainty that the suspected outlier is indeed an outlier, it needs to be replaced with the median calculated from all values found in Dataset 2. For the case of *Polymer*, the median was calculated to be 27 from all values in table 2. Replacing an outlier with the median to accommodate the data has been proven to be quite effective in dealing with outliers by introducing less distortion to that dataset.[39] Extreme values are therefore replaced with values more consistent with the rest of the data.[40]

		January	February	March	April	May	June	July	August	September	October	November	December
Polymer	**2009**	27	14	35	22	15	28	24	19	11	8	13	7
Polymer	**2010**	12	15	26	33	38	64	39	5	13	**27**	109	44
Polymer	**2011**	113	159	638	345	52	57	94	70	39	36	221	65
Polymer	**2012**	130	4	98	24	27	18	13	16	18	25	9	5

Table 3. The Identified Outlier is Replaced with the Median (Highlighted in Bold).

Table 3 represents the number of full-text articles downloaded for *Polymer* after the outlier had been replaced with the median. The confirmed outlier of 15,123 articles downloaded recorded in October 2010 is replaced with the median of 27, highlighted in bold. This then becomes the accepted value for the number of articles downloaded from *Polymer* in October 2010. The outlier is discarded. The new value of 27 articles downloaded in October 2010 replaces the extreme value of 15,123 in the original 2010 JR1 Report (see table 4). This is the final step.

	January	February	March	April	May	June	July	August	September	October	November	December
Polymer	12	15	26	33	38	64	39	5	13	**27**	109	44
Surface and Coatings Technology	3	1	2	1	22	17	17	0	12	3,771	5,428	601
International Journal of Radiation Oncology	11	18	35	22	17	6,436	176	13	25	29	24	19
Journal of Catalysis	0	1	5	1	2	2	16	4	0	2	6,693	1

Table 4. Sample from a 2010 JR1 COUNTER-Compliant Report Indicating the Number of Articles Downloaded per Journal Over a Twelve-Month Period. *Polymer*'s Identified Outlier Is Replaced with the Median Calculated from Table 2 (Highlighted in Bold).

Once the first outlier is corrected, the same procedures need to be followed for the other suspected outliers highlighted in table 1. If it is determined that they are outliers, they are replaced with their associated median values. Although the steps and calculations used to identify and correct for outliers are relatively simple to follow, it is admittedly a very lengthy and time-consuming process. But in the end, it is well worth the effort.

RESULTS AND DISCUSSION

Table 5 details the changes in the overall number of articles downloaded from J. N. Desmarais Library e-journals that resulted from the elimination of outliers. The column titled "Recorded Downloads" details the number of articles downloaded between 2000 and 2012, inclusively, prior to outlier testing. The column titled "Corrected Downloads" represents the number of articles downloaded during the same period of time but after the outliers had been positively identified and the data cleaned. The affected values are highlighted in bold.

Year	Recorded Downloads	Corrected Downloads
2000	806	806
2001	1034	1034
2002	1015	1015
2003	4890	4890
2004	72841	72841
2005	251335	251335
2006	640759	640759
2007	731334	731334
2008	710043	710043
2009	725019	725019
2010	**857360**	**757564**
2011	**869651**	**696973**
2012	716890	716890

Table 5. Comparison of the Recorded Number of Articles Downloaded to the Corrected Number of Articles Downloaded, Over a Thirteen-Year Period.

All data from all available years were tested for outliers. Only data recorded in 2010 and 2011 tested positive for outliers. Replacing outliers with the median values for those affected journal titles dramatically reduced the total number of downloaded articles (see table 5). Between 2007 and 2009, inclusively, the actual number of full-text articles downloaded recorded from the library's e-journal collection totaled between 731,334 and 725,019 annually (see table 5). The annual average for those three years is 722,132 articles downloaded. But in 2010 that number dramatically increased to 857,360 downloaded articles, which was followed by 869,651 downloaded articles in 2011 (see table 5). The elimination of outliers from the 2010 data resulted in the number of downloads dropping from 857,360 to 757,564, a difference of nearly 99,796 downloads, or 12 percent. Similarly, in 2011, the number of articles downloaded decreased from 869,651 to 696,973 once outliers were replaced with median values. This represents a reduction of over 172,678 downloaded articles, or 20 percent. A staggering 20 percent of articles downloaded in 2011 can therefore be considered as erroneous and, in all likelihood, the result of illicit downloading.

Figure 1 is a graphical representation of the change in the number of articles downloaded before and after the identification of outliers and their replacement by median values. The line "Recorded Downloads" clearly indicates a surge in usage between 2010 and 2011 with usage returning to levels recorded prior to the 2010 increase. The line "Corrected Downloads" depicts a very different picture. The plateau in usage that began in 2007 continues through 2012.

Evidently, the observed spike in usage was artificial and the result of the presence of outliers in certain datasets. If the data had not been tested for outliers, it would have appeared that usage

had substantially increased in 2010 and it would have been incorrectly assumed that usage was on the rise once more. Instead, the corrected data bring usage levels for 2010 and 2011 back in line with the plateau that had begun in 2007 and reflects a more realistic picture of usage rates at Laurentian University.

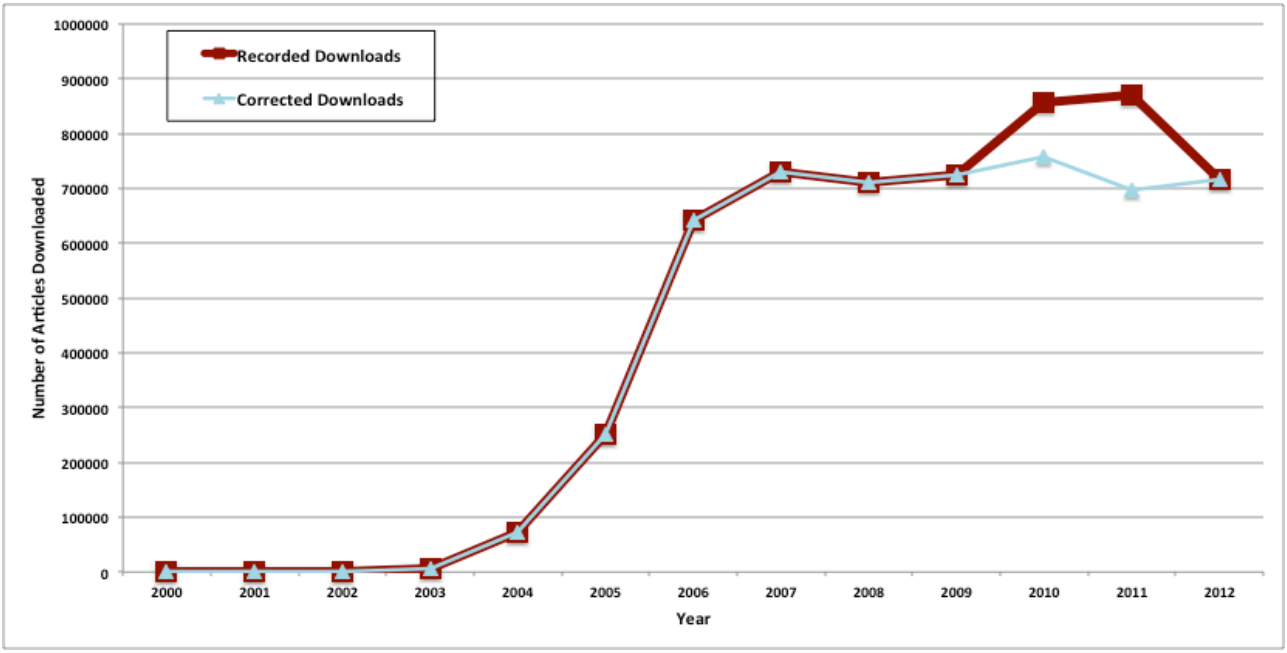

Figure 1. Comparing the Recorded Number of Articles Downloaded to the Corrected Number of Articles Downloaded Over a Thirteen-Year Period**.**

Accuracy in any data gathering is always extremely important, but accuracy in e-resource usage levels is critical for academic libraries. Academic libraries having e-journal subscription rates based either entirely or partly on usage can be greatly affected if usage numbers have been artificially inflated. It can lead to unnecessary increases in cost. Since it was determined that outliers were present only during the period in which the library had found itself under "attack," it can be assumed that the vast majority, if not all, of the extreme usage values were a result of illegal downloading. It would therefore be a shame to need to pay higher costs because of inappropriate or illegal downloading of licensed content.

Accurate usage data is also important for academic libraries that integrate usage statistics into their collection development policy for the purpose of justifying the retention or cancellation of a particular subscription. The J. N. Desmarais Library is such a library. As indicated earlier, if the cost-per-download of a subscription is consistently greater than the cost of an interlibrary loan for three or more years, it is marked for cancellation. At the J. N. Desmarais Library, the average cost of an interlibrary loan had been previously calculated to be approximately Can$15.00.[42] Therefore, subscriptions recording a "cost-per-download" greater than the Can$15.00 target for more than three years can be eliminated from the collection.

Any artificial increase in the number of downloads would have as result to artificially lower the cost-per-use ratio. This would reinforce the illusion that a particular subscription was used far more than it really was and lead to the false belief that it would be less expensive to retain rather than rely on interlibrary loan services. The true cost-per-use ratio may be far greater than initially calculated. The unnecessary retention of a subscription could prevent the acquisition of another, more relevant, one. For example, after adjusting the number of articles downloaded from ScienceDirect in 2011, the cost-per-download ratio increased from Can$0.74 to Can$1.59, a 53 percent increase. For the J. N. Desmarais Library, this package was obviously not in jeopardy of being cancelled. but a 53 percent change in the cost-per-use ratio for borderline subscriptions would definitely have been affected. It must also be stated that none of the library's subscriptions having experienced extreme downloading found themselves in the position of being cancelled after the usage data had been corrected for outliers.

Regardless, it is important to verify all usage data prior to any data analysis to identify and correct for outliers. Once the outlier detection investigation has been completed and any extreme values replaced by the median, there would be no further need to manipulate the data in such a fashion. The identification of outliers is a one-time procedure. The corrected or cleaned datasets would then become the official datasets to be used for any further usage analyses.

CONCLUSIONS

Outliers can have a dramatic effect on the analysis of any dataset. As demonstrated here, the presence of outliers can lead to the misrepresentation of usage patterns. They can artificially inflate average values and introduce severe distortion to any dataset. Fortunately, they are fairly easy to identify and remove. The following steps were used to identify outliers in JR1 COUNTER-Compliant reports:

1. Identify possible outliers: Visually inspect the values recorded in a JR1 report dataset (Dataset 1) and mark any extreme values.
2. For each suspected outlier identified, take the usage values for the affected e-journal title and incorporate them into a separate blank spreadsheet (Dataset 2). Incorporate into Dataset 2 all other usage values for the affected journal from all available years. It is important that Dataset 2 contain only those values for the affected journal.
3. Test for the outlier: Perform Dixon Q Test on the suspected outlier to confirm or disprove existence of the outlier.
4. If the suspected outlier tests as positive, calculate the median of Dataset 2.
5. Replace the outlier in Dataset 1 with the median calculated from Dataset 2.
6. Perform steps 1 through 5 for any other suspected outliers in Dataset 1.
7. The corrected values in Dataset 1 will become the official values and will be used for all subsequent usage data analysis.

The identification and removal of outliers had a noticeable effect on the usage statistics for J. N. Desmarais Library's e-journal collection. Outliers represented over 100,000 erroneous downloaded articles in 2010 and nearly 200,000 in 2011. A total of 20 percent of recorded downloads in 2011 were anomalous, and in all likelihood a result of illicit downloading after Laurentian University's EZProxy server was breached.

New technologies have made digital content easily available on the web, which has caused serious concern for both publishers[43] and institutions of higher learning, which have been experiencing an increase is illicit attacks.[44] The history of Napster supports the argument that users "will freely steal content when given the opportunity."[45] Since web robot traffic will continue to grow in pace with the Internet, it is critical that this traffic be factored into the performance and protection of any web servers.[46]

REFERENCES

1. Victoria J. Hodge and Jim Austin, "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review* 85 (2004): 85–126, http://dx.doi.org/10.1023/B:AIRE.0000045502.10941.a9; Patrick H. Menold, Ronald K. Pearson, and Frank Allgöwer, "Online Outlier Detection and Removal," in *Proceedings of the 7th Mediterranean Conference on Control and Automation (MED99) Haifa, Israel—June 28-30, 1999* (Haifa, Israel: IEEE, 1999): 1110–30.
2. Hodge and Austin, "A Survey of Outlier Detection Methodologies," 85–126.
3. Vic Barnett and Toby Lewis, *Outliers in Statistical Data* (New York: Wiley, 1994).
4. Hodge and Austin, "A Survey of Outlier Detection Methodologies," 85–126; R. S. Witte and J. S. Witte, *Statistics* (New York: Wiley, 2004); Menold et al., "Online Outlier Detection and Removal," 1110–30.
5. Menold et al., "Online Outlier Detection and Removal," 1110–30.
6. Hodge and Austin, "A Survey of Outlier Detection Methodologies," 85–126.
7. Laurentian University (Sudbury, Canada) is classified as a medium multi-campus university. Total 2012 full-time student population was 6,863, of which 403 were enrolled in graduate programs. In addition, 2012 part-time student population was 2,652 with 428 enrolled in graduate programs. Also in 2012, the university employed 399 full-time teaching and research faculty members. Academic programs cover a multiple of fields in the sciences, social sciences, and humanities and offers 60 undergraduate, 17 master's, and 7 doctoral degrees.
8. Alain R. Lamothe, "Factors Influencing Usage of an Electronic Journal Collection at a Medium-Size University: An Eleven-Year Study," *Partnership: The Canadian Journal of Library and Information Practice and Research* 7, no. 1 (2012), https://journal.lib.uoguelph.ca/index.php/perj/article/view/1472#.U36phvmSy0J.

9. Ben Tremblay, "Web Bot—What is it? Can It Predict Stuff?" *Daily Common Sense: Scams, Science and More* (blog), January 24, 2008, http://www.dailycommonsense.com/web-bot-what-is-it-can-it-predict-stuff/.
10. Derek Doran and Swapna S. Gokhale, "Web Robot Detection Techniques: Overview and Limitations," *Data Mining and Knowledge Discovery* 22 (2011): 183–210, http://dx.doi.org/10.1007/s10618-010-0180-z.
11. C. Lee Giles, Yang Sun, and Isaac G. Councill, "Measuring the Web Crawler Ethics," in *WWW 2010 Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC: International World Wide Web Conferences Steering Committee, 2010): 1101–2, http://dx.doi.org/10.1145/17772690.1772824.
12. Shinil Kwon, Kim Young-Gab, and Sungdeok Cha, "Web Robot Detection Based on Pattern-Matching Technique," *Journal of Information Science* 38 (2012): 118–26, http://dx.doi.org/10.1177/0165551511435969.
13. David Watson, "The Evolution of Web Application Attacks," *Network Security* (2007): 7–12, http://dx.doi.org/10.1016/S1353-4858(08)70039-4.
14. Eric Kin-wai Lau, "Factors Motivating People toward Pirated Software," *Qualitative Market Research* 9 (2006): 404–19, http://dx.doi.org/1108/13522750610689113.
15. Huan-Chueh Wu et al., "College Students' Misunderstanding about Copyright Laws for Digital Library Resources," *Electronic Library* 28 (2010): 197–209, http://dx.doi.org/10.1108/02640471011033576.
16. Ibid.
17. Ibid.
18. Emma McCulloch, "Taking Stock of Open Access: Progress and Issues," *Library Review* 55 (2006): 337–43; C. Patra, "Introducing E-journal Services: An Experience," *Electronic Library* 24 (2006): 820–31.
19. Wu et al., "College Students' Misunderstanding about Copyright Laws for Digital Library Resources," 197–209.
20. Ibid.
21. Vincent J. Calluzzo and Charles J. Cante, "Ethics in Information Technology and Software Use," *Journal of Business Ethics* 51 (2004): 301–12, http://dx.doi.org/10.1023/B:BUSI.0000032658.12032.4e.
22. S. L. Solomon and J. A. O'Brien "The Effect of Demographic Factors on Attitudes toward Software Piracy," *Journal of Computer Information Systems* 30 (1990): 41–46.
23. J. N. Desmarais Library, "Collection Development Policy" (Sudbury, ON: Laurentian University, 2013),

http://biblio.laurentian.ca/research/sites/default/files/pictures/Collection%20Development%20Policy.pdf.

24. Lamothe, "Factors Influencing Usage"; Alain R. Lamothe, "Electronic Serials Usage Patterns as Observed at a Medium-Size University: Searches and Full-Text Downloads," *Partnership: The Canadian Journal of Library and Information Practice and Research* 3, no. 1 (2008), https://journal.lib.uoguelph.ca/index.php/perj/article/view/416#.U364KvmSy0I.
25. Martin Zimerman, "E-books and Piracy: Implications/Issues for Academic Libraries," *New Library World* 112 (2011): 67–75, http://dx.doi.org/10.1108/03074801111100463.
26. Ibid.
27. Peggy Hageman, "Ebooks and the Long Arm of the Law," *EContent* (June 2012), http://www.econtentmag.com/Articles/Column/Ebookworm/Ebooks-and-the-Long-Arm-of-the-Law--82976.htm.
28. "dataset, n.," *OED Online*, (Oxford, UK: Oxford University Press, 2013), http://www.oed.com/view/Entry/261122?redirectedFrom=dataset; "Dataset—Definition," OntoText, http://www.ontotext.com/factforge/dataset-definition; W. Paul Vogt, "Data Set," *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences* (London, UK: Sage, 2005); Allan G. Bluman, *Elementary Statistics—A Step by Step Approach* (Boston: McGraw-Hill, 2000).
29. David B. Rorabacher, "Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's 'Q' Parameter and Related Subrange Ratios at the 95% Confidence Level," *Analytical Chemistry* 63 (1991): 139–45; R. B. Dean and W. J. Dixon, "Simplified Statistics for Small Numbers of Observations," *Analytical Chemistry* 23 (1951): 636–38, http://dx.doi.org/10.1021/ac00002a010.
30. Surenda P. Verma and Alfredo Quiroz-Ruiz, "Critical Values for Six Dixon Tests for Outliers in Normal Samples up to Sizes 100, and Applications in Science and Engineering," *Revista Mexicana de Ciencias Geologicas* 23 (2006): 133–61.
31. Robert R. Sokal and F. James Rohlf, *Biometry* (New York: Freeman, 2012); J. H. Zar, *Biostatistical Analysis* (Upper Saddle River, NJ: Prentice Hall, 2010).
32. "null hypothesis," *AccessScience* (New York: McGraw-Hill Education, 2002), http://www.accessscience.com.
33. Ibid.
34. "critical value," *AccessScience*, (New York: McGraw-Hill Education, 2002), http://www.accessscience.com.
35. Ibid.
36. Verma and Quiroz-Ruiz, "Critical Values for Six Dixon Tests for Outliers," 133–61.

37. Rorabacher, "Statistical Treatment for Rejection of Deviant Values," 139–45.

38. Ibid.

39. Jaakko Astola and Pauli Kuosmanen, *Fundamentals of Nonlinear Digital Filtering* (New York: CRC, 1997); Jaakko Astola, Pekka Heinonen, and Yrjö Neuvo, "On Root Structures of Median and Median-Type Filters," *IEEE Transactions of Acoustics, Speech, and Signal Processing* 35 (1987): 1199–201; L. Ling, R. Yin, and X. Wang, "Nonlinear Filters for Reducing Spiky Noise: 2-Dimensions," *IEEE International Conference on Acoustics, Speech, and Signal Processing* 9 (1984): 646–49; N. J. Gallagher and G. Wise, "A Theoretical Analysis of the Oroperties of Median Filters," *IEEE Transactions of Acoustics, Speech, and Signal Processing* 29 (1981): 1136–41.

40. Menold et al., "Online Outlier Detection and Removal," 1110–30.

41. Ibid.

42. Lamothe, "Factors Influencing Usage"; Lamothe, "Electronic Serials Usage Patterns."

43. Paul Gleason, "Copyright and Electronic Publishing: Background and Recent Developments," *Acquisitions Librarian* 13 (2001): 5–26, http://dx.doi.org/10.1300/J101v13n26_02.

44. Tena McQueen and Robert Fleck Jr., "Changing Patterns of Internet Usage and Challenges at Colleges and Universities," *First Monday* 9 (2004), http://firstmonday.org/issues/issue9_12/mcqueen/index.html.

45. Robin Peek, "Controlling the Threat of E-Book Piracy," *Information Today* 18, no. 6 (2001): 42.

46. Gleason, "Copyright and Electronic Publishing," 5–26.