Deviations from Power Law in Citation Distributions

Technical report originally posted in May 2010

Richard Barker

Brian E. Carpenter
brian@cs.auckland.ac.nz
Ewan Tempero
e.tempero@cs.auckland.ac.nz

Department of Computer Science
The University of Auckland

Computer Science Technical Reports (2010-001)
ISSN 1173-3500

Deviations from Power Law in Citation Distributions

(TECHNICAL REPORT ON WORK IN PROGRESS, MAY 2010)

Richard Barker
richardb@cs.auckland.ac.nz

Brian E. Carpenter
brian@cs.auckland.ac.nz

Ewan Tempero
e.tempero@cs.auckland.ac.nz

Department of Computer Science, University of Auckland
Private Bag 92019, Auckland, New Zealand

ABSTRACT

It is commonly asserted that citations within a corpus of material such as scholarly articles or web pages show a power law distribution. While this appears to be approximately true for incoming citations, we describe cases where there is substantial deviation from a power law for outgoing citations.

Categories and Subject Descriptors

TBD [**TBD**]

General Terms

TBD

Keywords

power law, citations

1. INTRODUCTION

Since an initial observation by Barabási and Albert[2], it has been common to assert that large and apparently random networks, whether natural or artificial, tend to follow a power law distribution when one ranks the cardinalities of their connections. Specifically, the assertion is that if a node has k links, then the probability distribution of k is given by $P(k) \sim k^{-\gamma}$ (for large k). The value of γ is a constant characteristic of the network studied, but is typically in the range 2 to 4.

Indeed there is a body of evidence, of which some early examples are cited in [2], showing this for artificial networks, including those composed of citations of scholarly articles or of web links. The topic has been comprehensively reviewed in [5].

In software engineering, and particularly when considering large software systems constructed using object-oriented techniques, there is considerable interest in mutual dependencies between software components. Baxter *et al.*[4] studied such dependencies for a corpus of 56 Java applications. They considered the Java classes occurring in this corpus and counted their mutual dependencies. It is important to stress the word *mutual*; a class may invoke a certain number of other classes, and it may be invoked by a different number of other classes. These two numbers are independent. In this paper, we will use the term *out-degree* to denote the number of outgoing invocations or citations *from* a given class, and *in-degree* to denote incoming invocations or citations *of* a given class

Baxter *et al.* investigated whether both the out-degree and the in-degree distributions matched the power law expectation. In fact, they observed a clear power law behaviour for the in-degree data. However, the results were not the same for the out-degree data; they could draw no clear conclusion about the mathematical nature of the distribution. The most that can be said is that low out-degree counts are much more frequent than a power law distribution would require.

Indeed there is no intrinsic reason why the out-degree distribution should be the same as the in-degree distribution; invocations from and invocations of a given class are completely independent choices made, very often, by different programmers. It is nevertheless interesting that the two types of distribution do not even appear to follow the same type of law.

We note that if we consider a Java corpus as a network or graph, its links or arcs are not bidirectional; if A invokes B, that does not mean that B invokes A. Similarly, in a corpus of scholarly articles, if A cites B, B is rather unlikely to cite A. In a corpus of web pages, mutual hyperlinks are certainly possible, but far from the norm. There is therefore no more reason for in-degree and out-degree distributions to be similar for document citations or web page citations than for a Java corpus. This article describes some preliminary observations of such distributions.

As far as we could determine, none of the major scholarly citation indexes collect out-degree data; published data and studies only consider the in-degree of each paper ("How many people have cited my paper?"). We therefore analysed a specific corpus of technical documents for which both in-degree and out-degree data were readily available. For web pages, we chose to use a locally accessible corpus of pages to ensure complete consistency of the in-degree and out-degree data. The following two sections describe our results.

2. THE RFC CORPUS

The Internet "Request for Comments" (RFC) series of documents has existed for forty years[9]. The large majority of RFCs are technical specifications or some form of discussion of such specifications. The series embodies many citations of RFCs by other RFCs. Almost all RFCs are available on-line in plain text format[8], and there is at least one database that tracks mutual references between RFCs, which made gathering the in-degree and out-degree data very straightforward. The data were collected on April 15, 2008 when there were 4309 RFCs with at least one citation in the database.

The in-degree results are shown in Fig. 1 and the out-degree results in Fig. 2, both as log-log plots, which would be straight lines for perfect power-law behaviour. Fig. 1 is indeed consistent with a power law, despite some outliers (one RFC, for example, being cited 1827 times). On the other hand, Fig. 2 shows somewhat the same type of deviation from a power law observed for out-degree plots in [4], with what can be viewed as an excess of RFCs making a very small number of outgoing references. Again there is an outlier, which is actually a bibliography of the first thousand RFCs, after which the bibliography was held on-line instead. However, what is interesting is the "flat top" of the curve, which represents the fact that 3131 of the 4309 RFCs contain ten or fewer references to other RFCs. The deviation from a power law is apparent.

3. WEB SITES

We collected data for the web sites *www.cs.auckland.ac.nz* and *www.auckland.ac.nz*, the latter being a superset of the former. We captured the sites using the WGET utility on DATE1 and DATE2 respectively. The first site contained XXX pages and the second site contained YYYY pages. We then analysed the contents in order to count both the links pointing to each page in each site, and the outgoing links in each page.

The two entire web sites were each downloaded using WGET, which was directed to start at the top level domain of each site and to download all pages which were linked to and whose link included the main domain, e.g., *link.auckland.ac.nz* and *sub.link.auckland.ac.nz* for *www.auckland.ac.nz*. In other words, only links within the site were considered. Only HTML pages were downloaded. These sites were then analyzed using a Java HTML parser. Counting the number of outgoing links on each page was trivial. *Mailto:* links were ignored. Counting incoming links was performed by creating a list of all outgoing links across all pages, breaking the list up into sections small enough to be sorted in memory, merging the sorted lists to form a complete sorted list and then generating a frequency count for all links. This process provided the in-degree and out-degree data.

The in-degree and out-degree results for the smaller site are shown in Fig. 3 and Fig. 4, and those for the larger site in Fig. 5 and Fig. 6.

Both in-degree plots appear to conform to a power-law result plus outliers, but the out-degree plots deviate substantially from this and exhibit a "flat top" shape, although the actual distributions in Figs 2, 4 and 6 are all different. At present we have not computed additional statistics as was done in [4].

4. DISCUSSION AND CONCLUSION

The observation that out-degree distributions deviate from a power law for a corpus of Java code[4] appears also to apply both to a corpus of technical documents (the RFCs) and to web sites of significant size. The fact that in all cases, the in-degree distributions conform to the power law expectation suggests that the data sets studied are not unusual or atypical, although of course this should be checked by studying additional data sets. Although the out-degree distributions concerned are generally similar in having a flat top, they do not appear to fit a particular mathematical law. It is for this reason that we have not carried out a statistical analysis – we have no hypothesis to test.

The conventional explanation for power-law behaviour in natural and artificial networks is based on popularity and herding (more formally called *preferential attachment*). A popular node is considered to be one with many incoming links, and such popularity is hypothesised to attract even more incoming links as a network grows. Mathematically, preferential attachment can lead to a power law behaviour[2],[5]. Indeed, one can readily understand that this model could apply to nodes such as Java classes, technical specifications, or web pages. A useful class, specification or page will be preferentially cited by new classes, specifications or pages, and its usefulness can be measured by the number of incoming citations.

However, there is no such rational explanation for outgoing citations. Consider a programmer writing a new Java class. The decision to invoke a particular class depends on its usefulness, but the *number* of classes she chooses to invoke is not a result of their usefulness; it is a result of the requirements of the new class being written. Exactly the same applies to the number of citations made by a technical specification such as an RFC, or the number of links from a web page. The rationale of the preferential attachment explanation does not apply.

The preferential attachment model is commonly qualified as applying for large k. Indeed, some of the example distributions shown in [2] (Fig. 1, A and C), [7] (Fig. 3, a, b and c), and [5] (Fig. 10.6 and Fig. 10.11) visibly deviate from a power law for small in-degrees. The last case is interesting because the author asserts that for Wikipedia articles, the "distribution of both in-degree and out-degree is a power law." In fact the cited figure shows only the in-degree and the exact data collection method is not explained; however, the observed pattern is rather similar to our first out-degree distribution in Fig. 2. We are not aware of a systematic explanation for this small k behaviour in some, but not all, in-degree plots. However, the original note drawing attention to power-law behaviour in the World-Wide Web[1] (Fig. 1a), expanded in [3], shows an out-degree plot which appears consistent with our flat top observations.

Some investigators of complex networks have considered an alternative to the classical power law known as the *exponentially truncated power law*, expressed as $P(k) \sim k^{-\gamma}e^{\alpha k}$ where α is an additional constant[6]. This produces a distribution which is not a straight line and indeed can be read as having a flat top. It is generally interpreted as fitting a real-world situation where the degree is naturally limited. For example, in social networks, people who can manage more than a few hundred social relationships are rather rare. We can hypothesise that for outgoing citations, the authors of software modules or of technical documents can only manage a limited number of external references.

We conclude that the deviation from simple power law behaviour for out-degree distributions found in [4] is not an artefact of the particular Java corpus studied. We have found such deviations in RFC citation data and in web hyperlink data. We do not propose a mathematical model for out-degree distributions, since the various distributions appear quite different, with only a "flat top" shape in common. We believe that observations of out-degree distributions from other sources would be of interest, to see how general this phenomenon may be.

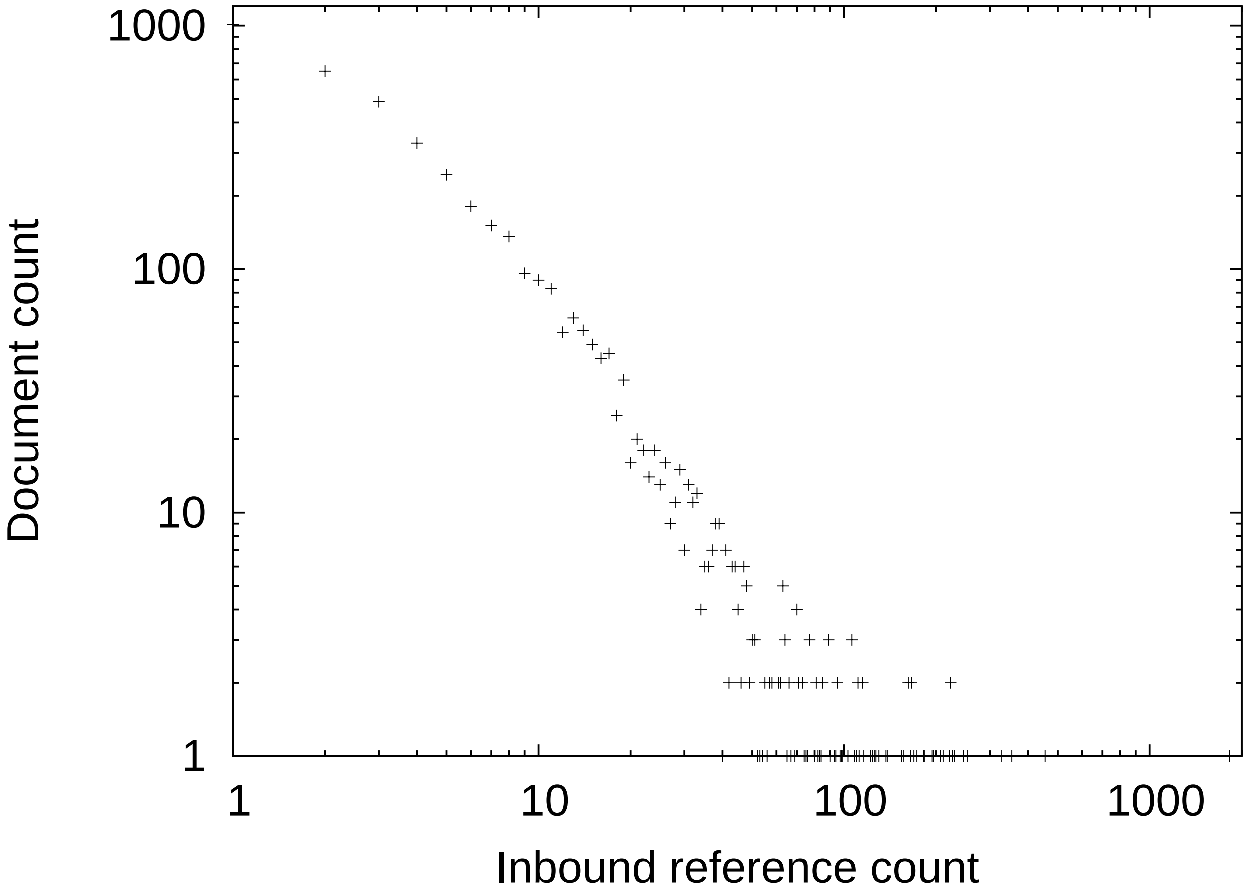

Figure 1: In-degree distribution for RFCs

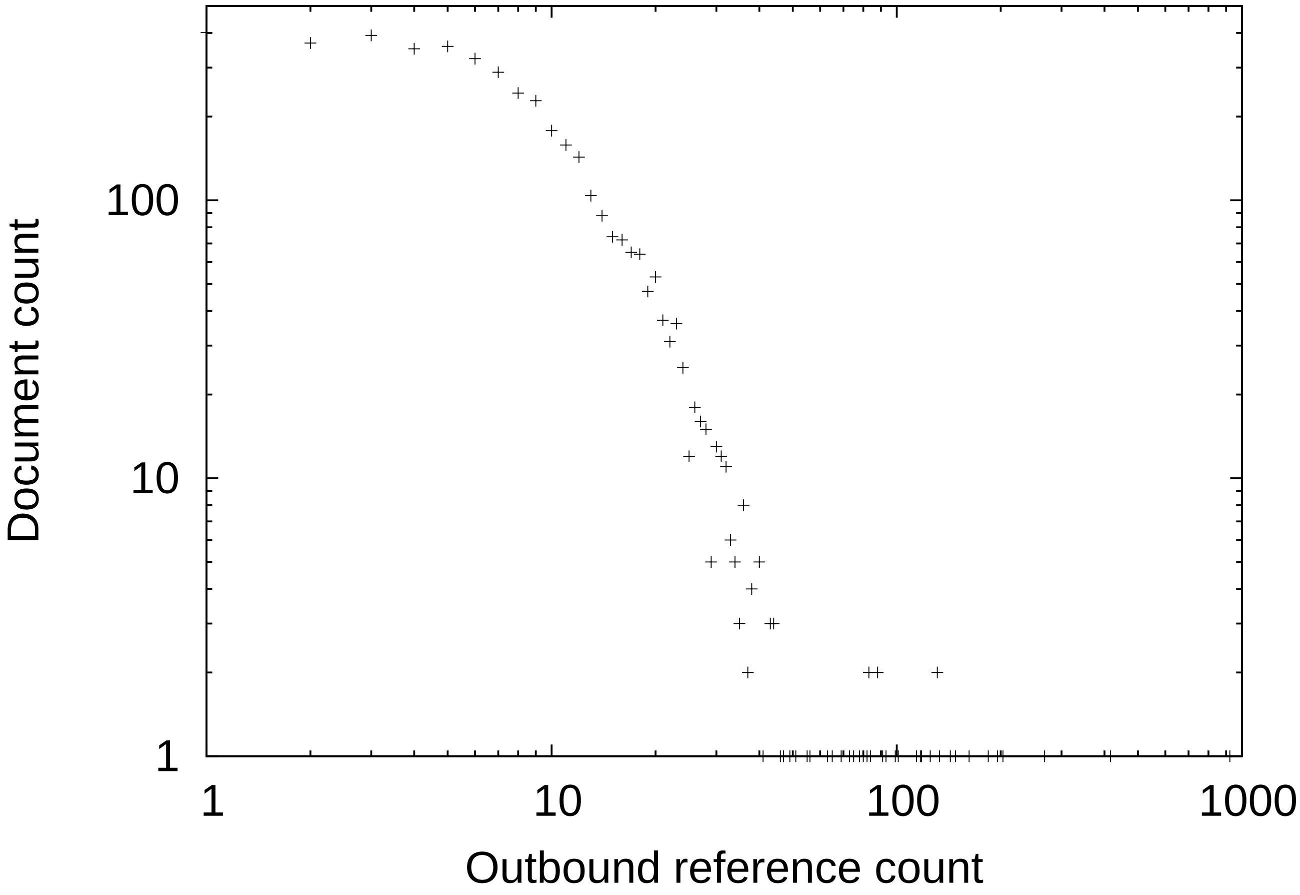

Figure 2: Out-degree distribution for RFCs

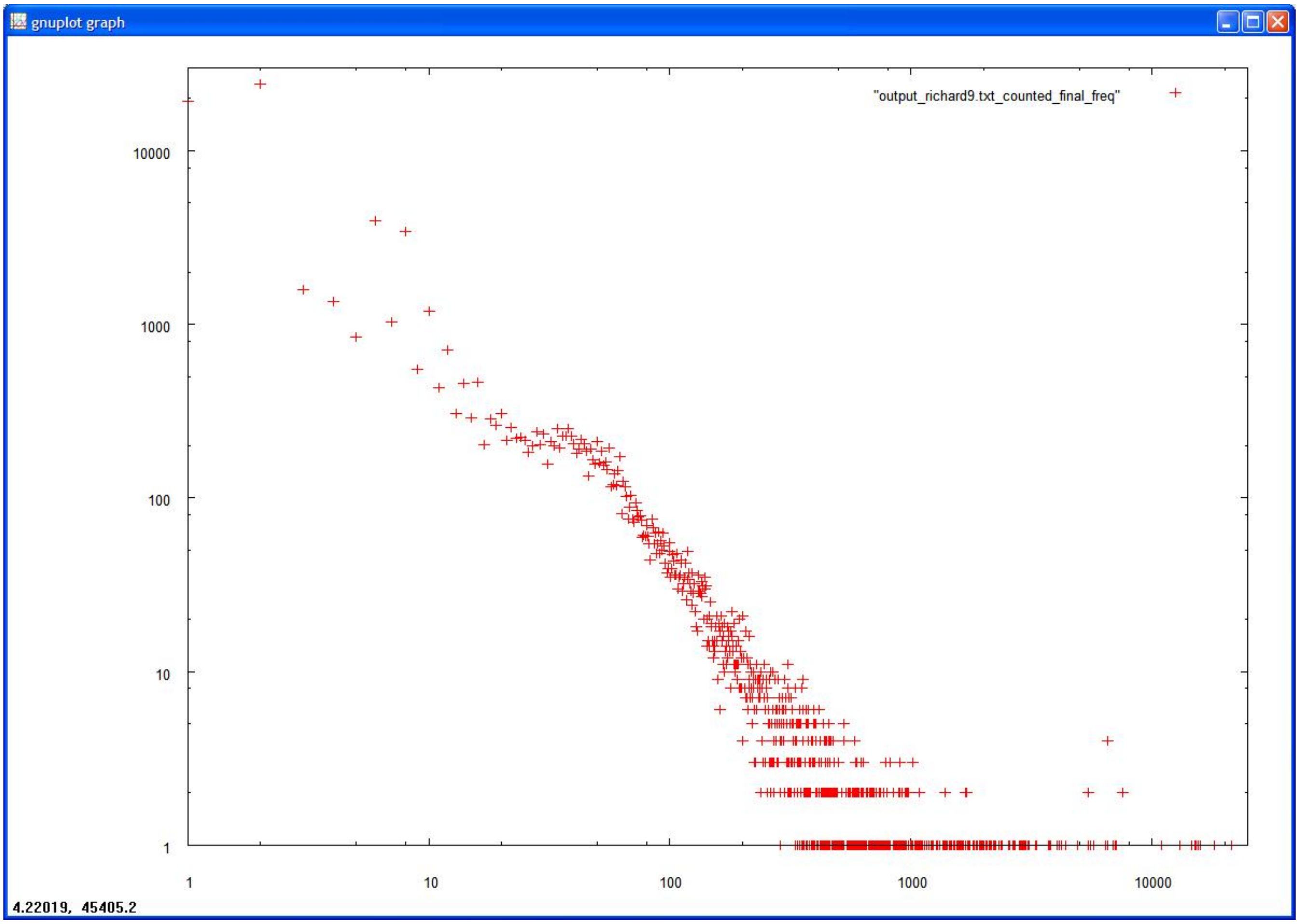

Figure 3: In-degree distribution for *www.cs.auckland.ac.nz*

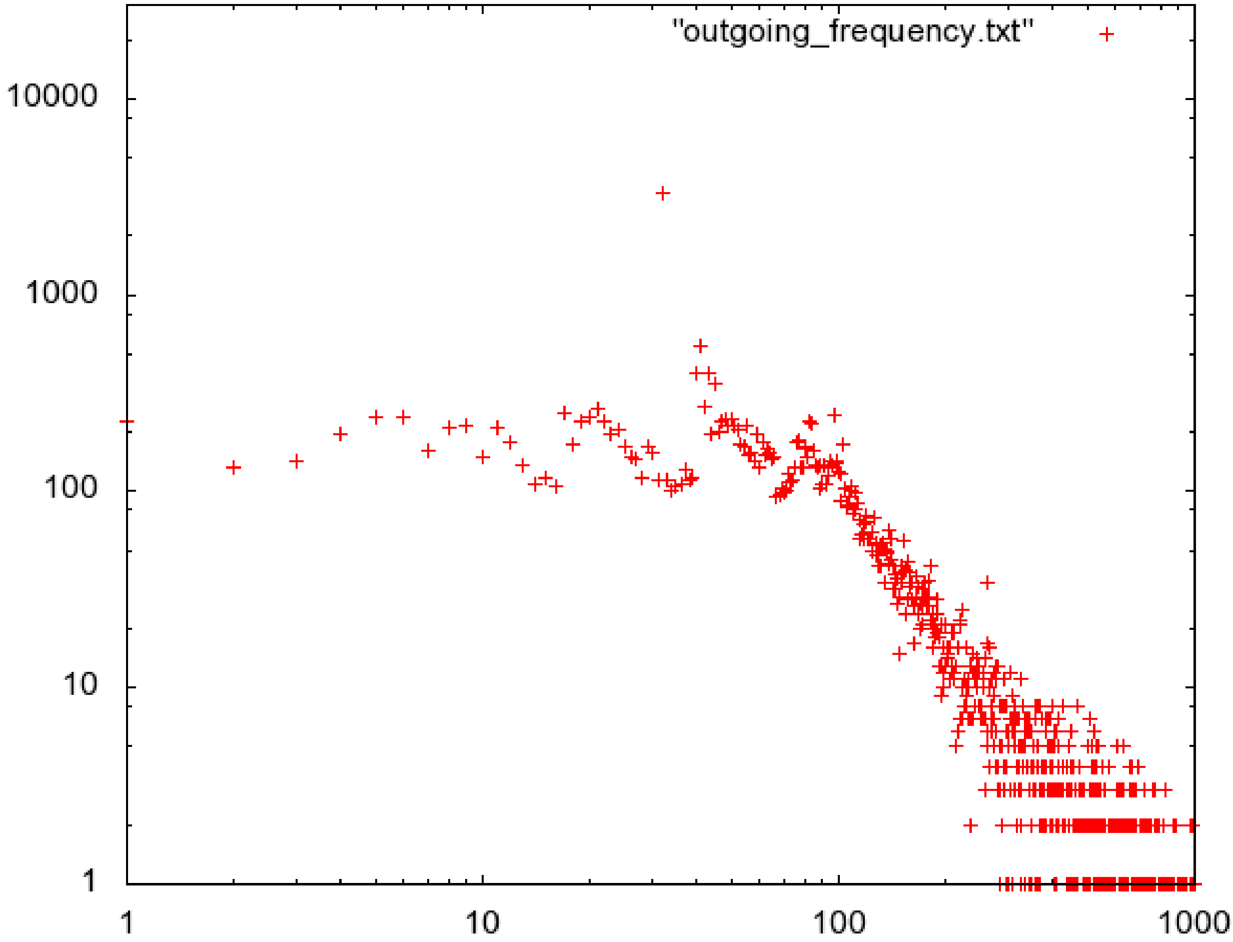

Figure 4: Out-degree distribution for ***www.cs.auckland.ac.nz***

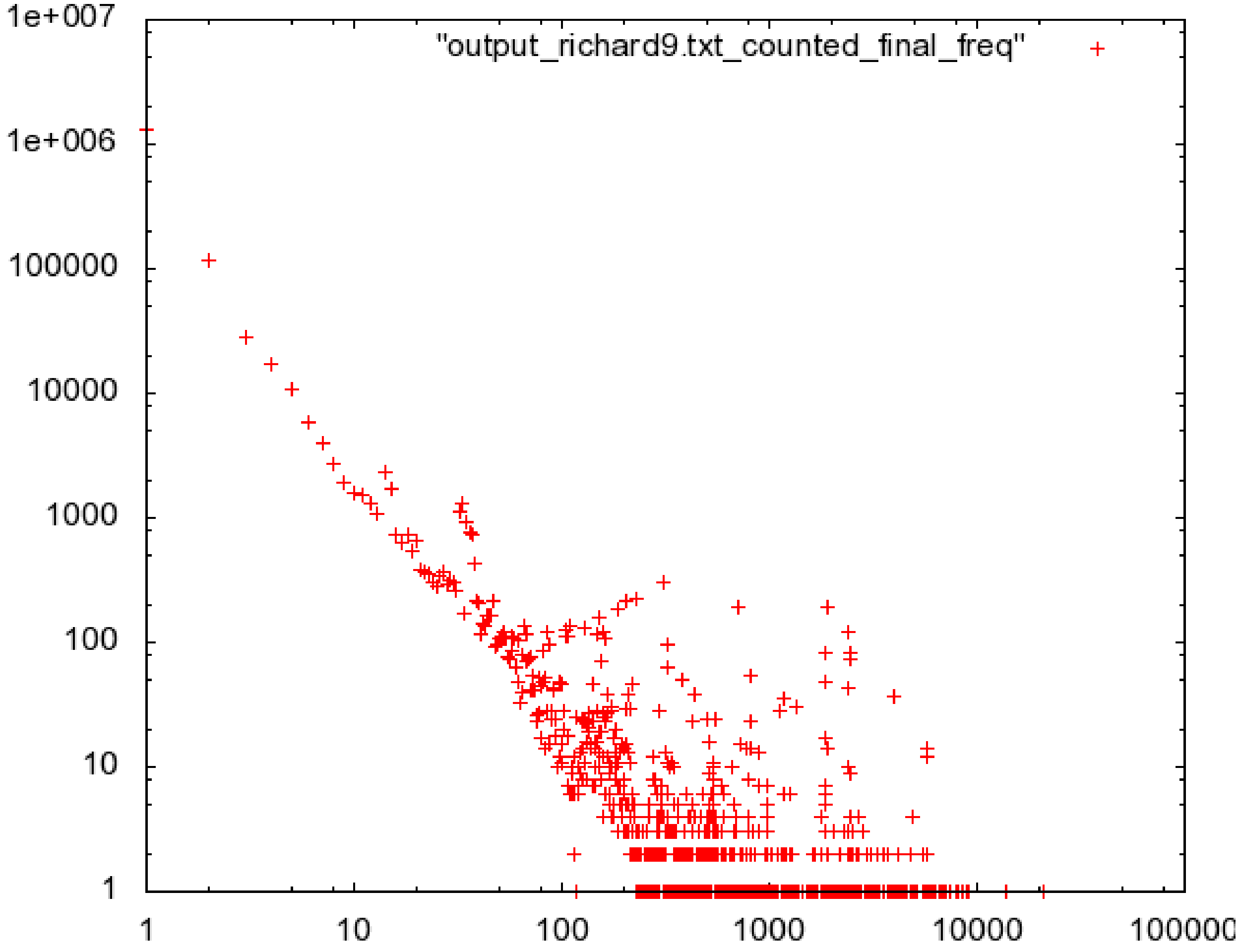

Figure 5: In-degree distribution for *www.auckland.ac.nz*

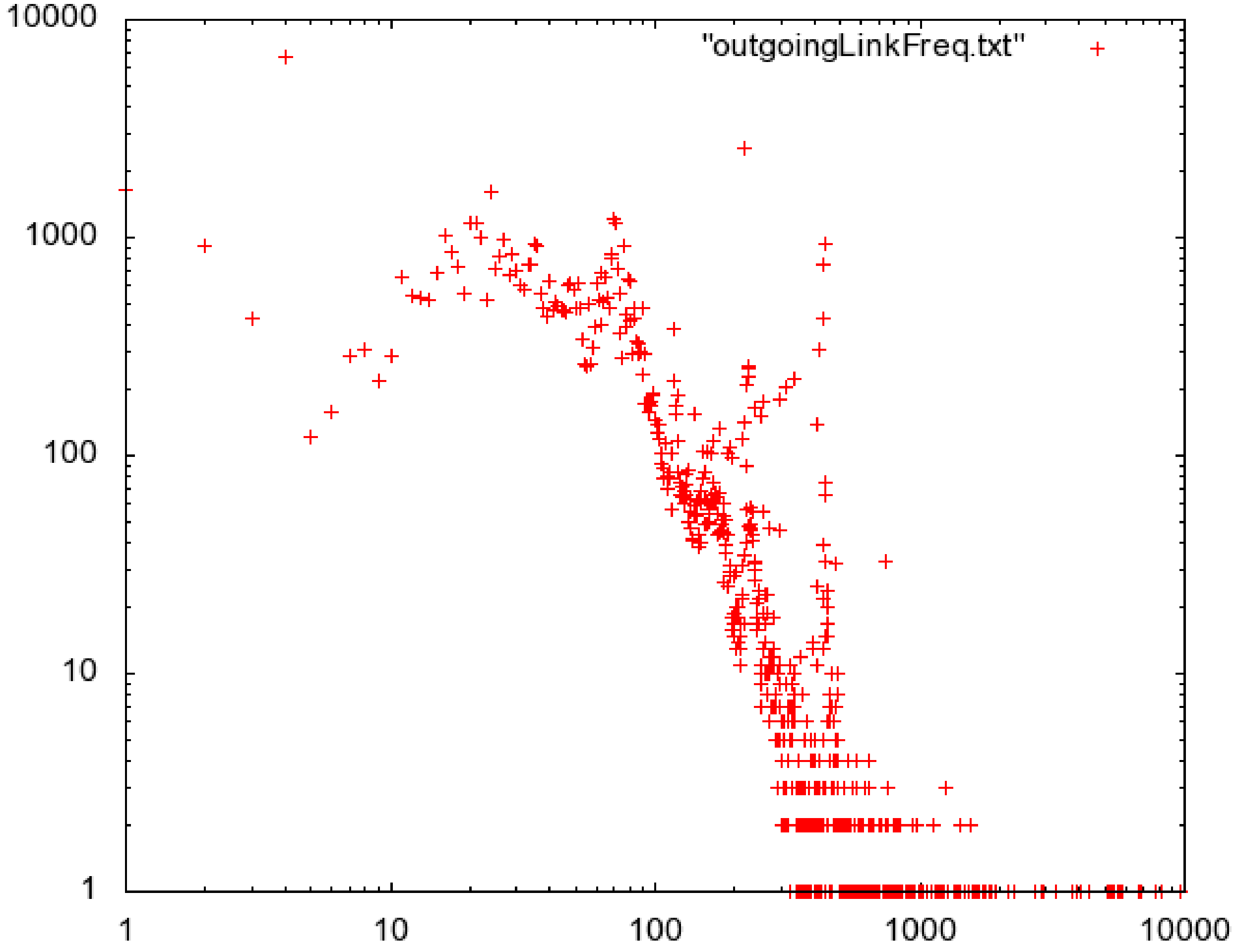

Figure 6: Out-degree distribution for *www.auckland.ac.nz*

5. ACKNOWLEDGMENTS

Bill Fenner kindly gave us direct access to the database he maintains of mutual dependencies among RFCs. This reduced the data collection job to two trivial SQL queries.

6. REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the World-Wide Web. *Nature*, 401:130–131, September 1999.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the World Wide Web. *Physica A*, 281:69–77, 2000.

[4] G. Baxter, M. R. Frean, J. Noble, M. Rickerby, H. Smith, M. Visser, H. Melton, and E. D. Tempero. Understanding the shape of Java software. In P. L. Tarr and W. R. Cook, editors, *OOPSLA*, pages 397–412. ACM, 2006.

[5] G. Caldarelli. *Scale-free Networks: Complex Webs in Nature and technology*. Oxford University Press, 2008.

[6] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, Jul 2001.

[7] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67(2):026112, 2003.

[8] RFC Editor. http://www.rfc-editor.org/.

[9] RFC Editor. 40 Years of RFCs. *Internet RFCs, ISSN 2070-1721*, RFC 5540, 2009.