

Amsterdam
School



for Social science Research

ASSR Working paper 03/06

August 2006

AUTOMATED LONGITUDINAL DATA COLLECTION ON THE WEB

Competition between Search Engines, 1993-2000

Ivar Vermeulen and Jeroen Bruggeman

Ivar Vermeulen is an assistant professor in Communication Science at the Free University of Amsterdam.

Jeroen Bruggeman is an assistant professor in Sociology at the University of Amsterdam.

ASSR WORKING PAPER SERIES

[HTTP://WWW.ASSR.NL/](http://www.assr.nl/)

EMAIL: ASSRWORKINGPAPERS@FMG.UVA.NL

ABSTRACT:

Competition is regarded as one of the key mechanisms of markets and of economic development in general. Measuring competition at a high level of precision over longer periods of time has always been difficult or extremely expensive. Nowadays the Web offers new opportunities, which we seized to investigate the entire population of 137 search engines from its inception till 2000, to monitor crowding and differentiation between specialists and generalists at the firm and population levels. We show that by employing web-based methods of data collection competition can be studied more systematically and more precisely than by traditional methods, at hardly any costs, and we provide a tentative explanation for the proliferation of specialist search engines. Our method of automated data collection is more generally applicable, though, and can be used for any study that needs large numbers of links between web pages.

We are grateful for comments received from Eelke Heemskerk, Gerben Korthouwer, and other ASSR members.

Please send correspondence to Jeroen Bruggeman, Department of Sociology and Anthropology, Universiteit van Amsterdam, Oudezijds Achterburgwal 185, 1012 DK Amsterdam, Netherlands, j.p.bruggeman@uva.nl. Ivar Vermeulen, Department of Communication Science, Vrije Universiteit Amsterdam, De Boelelaan, 1081 HV Amsterdam, Netherlands, ivar@science.uva.nl.

In both economics and sociology, competition is regarded as one of the key mechanisms of markets and of economic development in general (Smith 1776, Porter 1986). Measuring competition longitudinally at the micro level in order to predict dynamics at the macro level has always been a major challenge, which we pick up in this research note. In a case study, we apply automated web-retrieval techniques to trace competition in a population of organizations, in our case Internet search engines. Our method is general and can be applied to any study that relies on large numbers of links between web pages, which can also be used to indicate popularity, status, and complementarity of web sites, depending on the theoretical embedding.

Search engines have become the most important tools for retrieving information from the Web (Lawrence and Giles 1999), and they are the principals mediating information flows between web pages and Internet users worldwide¹. The economic value of this mediating position bridging a structural hole (Burt 1992) can be indicated by the value of current market leader Google, estimated at \$80 billion in 2005 (BBC 2005).

Recently, researchers have expressed concerns regarding the so-called gatekeeper role of search engines (Hinman 2004). It has been shown that the ranking algorithms search engines apply influence the accessibility and status of medical (Allen et al. 2002), political (Introna 2000), scientific (Bar-Ilan 2004), commercial (Vaughan and Thelwall 2004), and news resources (Dahlberg 2005). Moreover, increasing consultation of search engines by Internet users intensifies price competition in other industries (Brown and Goolsbee 2002). Considering these side effects of search engine use, some researchers regard the apparent market dominance of search engine Google as disturbing, because it results in a concentration of power within a single organization (Machill et al. 2004).

Although the hegemony of Google is currently undisputed, especially in the English-speaking parts of the world, there are also many search engine alternatives that Internet users can and do take recourse to. In Google's shadow, over a dozen of well-established competitors, such as Altavista and Alltheweb, provide services to

¹ For an overview, see: <http://www.zakon.org/robert/internet/timeline/>.

At <http://www.w3.org/History/19921103-hypertext/hypertext/WWW/Summary.html> the original summary of the World Wide Web project can still be found, as well as links to the first browser software. For the Web in general, Broder et al (2000), and Kleinberg and Lawrence (2001) describe its structure, Albert et al (1999) its diameter, and Huberman and Adamic (1999) its scale-free growth dynamics.

millions of Internet users a day, also (i) covering the entire Web and (ii) providing their services in English. These search engines, fulfilling (i) and (ii), we regard as *generalists*. Moreover, a great many other users, mostly those that do not have English as their first language, use over a hundred specialized search engines that cater to specific languages or geographical areas. Most members of this latter group of specialist search engines originated in the period from 1997 to 2000, an era of differentiation in the search engine industry. Then, over 80% of the new entrants specialized in finding pages from a specific non-English language, such as Portuguese, Russian, or Japanese, which often also implied geographical specialization. *Specialists* in our conceptualization are search engines that do not fulfil (i), (ii) or both, which in ecological theory corresponds to their having a more narrow niche than generalists (Hannan and Freeman 1977).

From the point of view of classical economics, this proliferation of specialized, and in terms of numbers of users, small search engines is puzzling. The search engine market, like most information distributor markets on the Internet, is characterized by significant economies of scale, both on the production side and on the demand side (Bakos and Brynjolfsson 2000). The first type of scale economies results from low marginal costs involved with extending a searchable database of web pages and providing access to this database to an increasing number of Internet users worldwide. The second type of scale economies is caused by search engines' becoming more valuable to advertisers if their market share increases. It is well documented that economies of scale favor large producers over small ones (e.g. Arthur 1996), and provide barriers to entry (e.g. Schmalensee and Willig 1989).

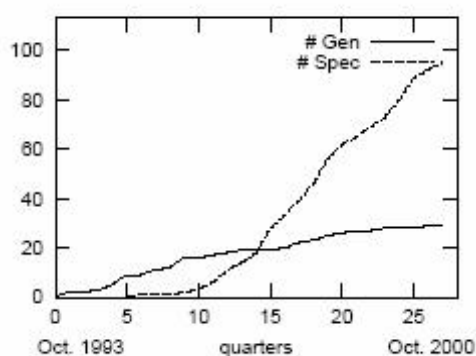
A possible answer to the question why specialists proliferate in the search engine market may be found in resource partitioning theory (Carroll 1985). According to this fragment of organizational ecology (Carroll and Hannan 2000), in mature and concentrated markets, a few large generalists having economies of scale advantages occupy the market center, leaving niche pockets in the market periphery where small specialists can enter and thrive. However, the search engine industry can't be regarded as a mature market yet, and concentration did not increase during our observation period (Vermeulen and Bruggeman 2004). Therefore the instantiation of resource partitioning theory seems not to be warranted.

Some observations do suggest that the specialist proliferation was fuelled by a redistribution of resources into a market center and a periphery, similar to the

process described by resource partitioning theory. In 1996, search engine HotBot was founded, which claimed to update its index every two weeks, and, consequently, to provide better and more up to date results than its competitors. HotBot was promptly reviewed as the “best search site on the Web” by a number of leading computer-related publications (PR Newswire 1998), which in turn persuaded established search engine entrepreneurs to imitate or improve upon HotBot. Their shifted focus of attention came at the expense of other, often international, activities of search engines. As an example of the latter, Altavista’s Latin-American search service AltaVista Magallanes extended its index, containing 300 million pages by the end of 1998, with only 300 pages from 1999 till 2001. These market failures due to increasing competition in the market center may have incited peripheral entrepreneurs to start up search engines, not so much as to engage in competition with market leaders, but to provide better services to users from their own countries of origin. Given these observations, we expect on the basis of resource partitioning theory (1) an increasing competition between generalists, and (2) that the increasing number of small specialists between 1997 and 2000 coincides with decreasing competitive pressure from large generalists on small specialist search engines.

Figure 1 shows the numbers of specialist and generalist search engines over the period of our study. During this early period of the market there were very few disbanding events. Niches grew, and the market grew even faster.

Figure 1 Numbers of generalist and specialist search engines, from October 1993 until October 2000.



In the ecological literature on organizations, competition is studied in terms of niches and niche overlaps (Hannan 2005). A niche is a set of resources on which an organization can survive. If two or more organizations compete for the same resources there is niche overlap, i.e., set intersection(s) consisting of consumers who regard products of multiple firms substitutable. In terms of niche overlap, substitutability of firm's products is seen from consumers (in general, resource providers') point of view. This is more accurate than measurements of substitutability on the basis of predetermined attributes like price and quality, since the latter may not be the actual attributes for consumers' decisions (e.g. Keller 1993). We measure niche, i.e., audience, overlap to determine competition between pairs of search engines first, and then aggregate to the (sub)population level. Because for a pair of overlapping niches of unequal size the smaller niche is always overlapped proportionally more than the larger niche, the competitive pressure associated with niche overlap is inherently asymmetric (McPherson 1983; Podolny et al 1996). Let sets I and J be the audiences of search engines i and j , respectively, and $i \neq j$. The competitive pressure, $c(i,j)$, received by search engine j by i , is defined accordingly, as the portion of J that is overlapped by I ,

$$c(i,j) = (I \cap J) / J \quad (1)$$

For each search engine j , the total competitive pressure, or crowding, received is the summation of these pair-wise relations of j with all other search engines i in the population,

$$C_{ij} = \sum_i c(i,j) \quad (2)$$

If one wishes, one may also look at a population as a network of organizations, where in this case arcs stand for competitive pressure of organizations on each other; then equation (2) is a weighted indegree measure of j . Following organizational ecology, we compare generalists with specialists, a distinction that applies to our population straightforwardly. To explore the dynamics of competition at the subpopulation level, we characterize it in terms of six aggregate measures: the mean competitive pressure received (a) by generalists; (b) by specialists; (c) by

generalists from generalists; (d) by generalists from specialists; (e) by specialists from generalists; and, (f) by specialists from specialists.

Our study covers the entire population of 137 Internet search engines in the seven years after its inception, from October 1993 to October 2000². Other studies focused on large generalists only (Gandal 2001; Lawrence and Giles 1998), or on a small subset thereof (Rindova and Kotha 2001); and, have a relatively small time window. To collect our data, we employed one of the objects of our research, the search engine Northernlight. This engine had a high coverage of the Web (estimated at 32.9%, ranking 3rd among its fellow search engines at the time of data collection; Hawking and Craswell 2001). Its search results were stable over time compared with other search engines; it could handle extensive queries with Boolean operators correctly; it yielded comprehensive results; and, one could reliably search the history of the Web, from Northernlight's own database. Although Northernlight originated in 1997, its index contains web pages dating back as far as 1993. We consider a page's "date" to denote the time it was last updated, not the time it was indexed by Northernlight. Parallel testing with other search engines made clear that Northernlight's results were not biased with respect to Northernlight-related queries. For 28 successive quarters we asked Northernlight for each of the 137 search engines in our study the number of Web pages that have a hyperlink to it³. This information we used to indicate for each search engine its niche (Maurer and Huberman 2000), assuming that a larger number of links indicates a larger audience, thus a larger niche. For niche overlap, we asked Northernlight for each pair of search engines the number of Web pages that have links to both search engines in the pair⁴. To get our data, our automated data-collector, specially made for this purpose, sent to Northernlight more than 250,000 queries over a two-week period in 2001, and stored its answers in our database. Shortly after our study,

² For (historical) search engine listings we consulted www.searchenginewatch.com, www.archive.org and www.searchengineguide.com, as well as several other sources.

³ In our measurement of audiences, we used a threshold of 10 links as a minimum for acknowledging the existence of a search engine audience, to sidestep noise.

⁴ We have no data on consumers visiting and comparing search engines (which together with infrastructural and labor market characteristics would indicate *fundamental* niches), and audience overlap as we measure it could indicate search engine complementarity instead (Baum and Korn 1996). If that were the case, though, using one search engine would for that user increase the value of other engines, like having shoes increases the value of shoestrings. Since time spent on using search engines is costly, though, we assume that users prefer to limit the number of search engines they use, and to find what they are looking for in one stroke. This claim is supported by the fact that search engines sharing their language have more niche overlap, although their search results are less complementary.

search engine Northernlight ceased its public services and focused on paid services only. Our data collection can be replicated by the use of currently available search engines such as Google APIs, even though it may take much longer to collect data⁵ and results may slightly differ.

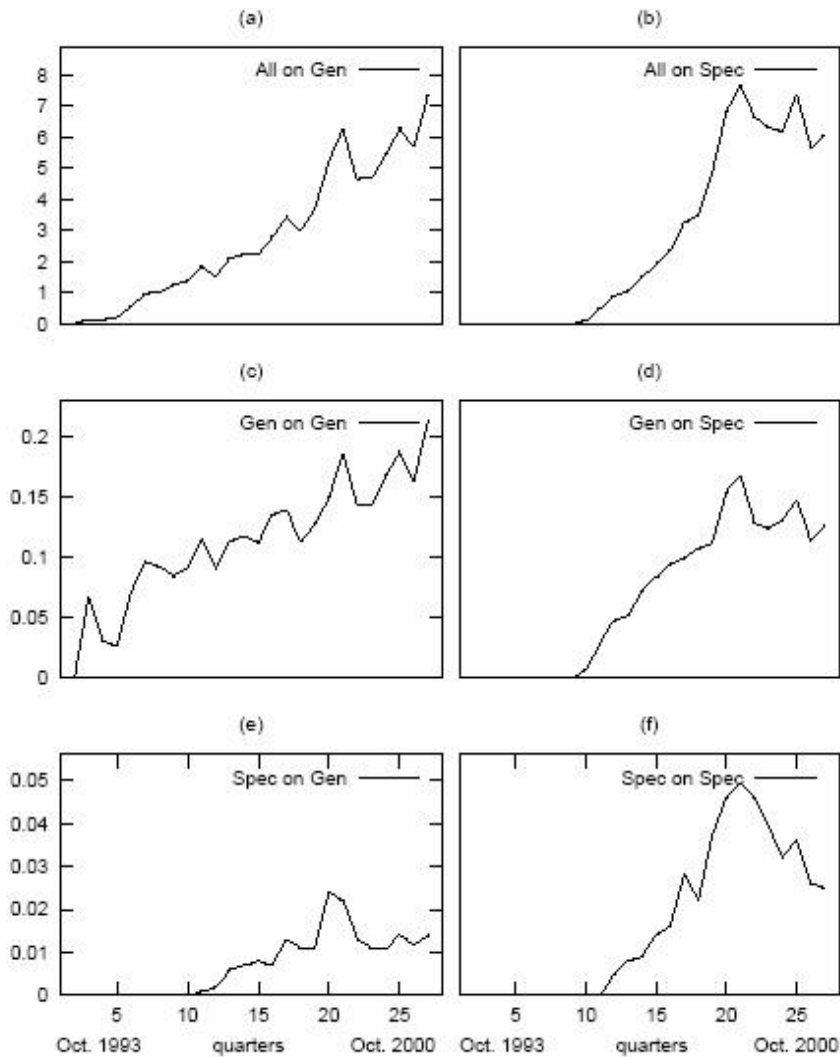
Figures 2a and 2b show that after a generic increase of competitive pressure, the mean pressure on specialist search engines decreased, whereas the mean pressure on the generalist search engines further increased. The decomposition of Figures 2a and 2b into Figures 2c to 2f shows that competition among generalists increased (2c), as we expected, whereas competition among specialists decreased after 1998 (2f). Competitive pressure by generalists on specialists, and vice versa, also decreased after 1998 (2d and 2e). The non-monotonic trajectory of competitive pressure may indicate a differentiation process (Borenstein and Netz. 1999): generalist search engines increased niche overlap with fellow generalists, while specialist search engines decreased niche overlap with generalist as well as with specialist competitors, as resource partitioning theory has it⁶. Our data do not permit us to draw conclusions about the causal relation between crowding and founding rates, among others because the growth of the specialist sub-population started roughly a year *before* competition started to decline. This may be explained by the presence of information asymmetry between Internet users and entrepreneurs. The latter group may have already noticed and acted upon opportunities created by the receding involvement of generalist search engines in peripheral activities, whereas common Internet users might lag behind. From an ecological perspective it is likely that decreasing competitive pressure on specialists has contributed both to their growing numbers and to their survival in the longer run⁷.

⁵ Google APIS allows registered members to post 1000 queries a day; a replication of our data set thus would take about 250 days.

⁶ High stress factors of multidimensional scaling prohibit presenting the outcomes of this method as results in the paper (Cox and Cox 1994).

⁷ It is interesting to note that organizations with large niches exert higher competition than those with small niches, but all niche sizes receive more or less the same competitive pressure, in fact the large niches receive slightly less.

Figure 2 Competitive pressure received by search engines, from October 1993 until October 2000



Notes: (a) The mean pressure received by generalists, (b) mean pressure received by specialists, (c) mean pressure exerted by generalists on generalists, (d) mean pressure by generalists on specialists, (e) mean pressure by specialists on generalists, (f) mean pressure by specialists on specialists. The maximal values, around quarter 20 in graphs (d) to (f), indicate the start of the differentiation of specialists.

A more general and possibly more important result of our study is that it shows that longitudinal data about an entire industry at the micro level can be reaped from the Web by using simple automated extraction tools. Our method can also be applied to studies of popularity, status, and complementarity of web sites, and the data obtained make possible to monitor changes at both micro and macro levels with higher efficiency and higher accuracy than if traditional survey methods were applied. This seems relevant not only for organizational and network scholars, but also for organizational strategists and policy makers.

References

- Albert, R., H. Jeong, A. and L. Barabási. 1999. The Diameter of the World Wide Web. *Nature* 401: 130-131.
- Allen, J.W. et al. 2002. The poor quality of information about laparoscopy on the World Wide Web as indexed by popular search engines. *Surgical Endoscopy* 16: 1, 170-172.
- Arthur, B. 1996. Increasing returns and the new world of business. *Harvard Business Review* 74: 100–109.
- Bakos, Y. and E. Brynjolfsson. 2000. Bundling and competition on the Internet. *Marketing Science* 19: 63–82.
- Bakos, Y. and E. Brynjolfsson 1999. Bundling Information Goods: Pricing, Profits, and Efficiency. *Management Science* 45:1613-1630
- Bar-Ilan, J. 2004. The use of web search engines in information science research. *Annual Review of Information Science and Technology* 33: 231–288.
- Baum, J.A.C. and H.J. Korn. 1996. Competitive dynamics of interfirm rivalry. *Academy of Management Journal* 39: 255-291.
- BBC 2005, June 8. \$80bn Google takes top media spot. Downloaded on May 15, 2006 from <http://news.bbc.co.uk/1/business/4072772.stm>
- Borenstein, S. and J. Netz. 1999. Why do all the flights leave at 8 am?: Competition and departure-time differentiation in airline markets. *International Journal of Industrial Organization* 17: 611–640.
- Broder, A., et al. 2000. Graph Structures in the Web. *Computer Networks* 33: 309-320.
- Burt, R.S. 1992. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, Mass

- Carroll, G.R. 1985. Concentration and specialization: Dynamics of niche width in populations of organizations. *American Journal of Sociology* 90: 1262-1283.
- Carroll, G.R. and M.T. Hannan 2000. *The Demography of Corporations and Industries*. Princeton University Press, Princeton, N.J.
- Cox, T.F. and M.A.A. Cox. 1994. *Multidimensional Scaling*. Chapman and Hall, London.
- Dahlberg, L. 2005. The Corporate Colonization of Online Attention and the Marginalization of Critical Communication? *Journal of Communication Inquiry* 29: 160-180.
- Gandal, N. 2001. The Dynamics of Competition in the Internet Search Engine Market. *International Journal of Industrial Organization* 19: 1103-1117.
- Hannan, M. T. and J. Freeman. 1977. The population ecology of organizations. *American Journal of Sociology* 82: 929-964.
- Hannan, M.T. 2005. Ecologies of Organizations: Diversity and Identity. *Journal of Economic Perspectives* 19: 51-70.
- Hawking, D. and N. Craswell. 2001. Measuring search engine quality. *Journal of Information Retrieval* 3: 33-59.
- Hinman, L.N. 2004. Esse est indicato in Google. Ethical and Political Issues in Search Engines. *International Review of Information Ethics* 3: 19-26.
- Introna, L.D. and Nissenbaum, H. 2000. Shaping the Web: why the politics of search engines matters. *The Information Society* 16: 169-185.
- Keller, K.L. 1993. Conceptualizing and Measuring Consumer-Based Brand Equity. *Journal of Marketing* 57: 1-22.

- Kleinberg, J. and S. Lawrence. 2001. The Structure of the Web. *Science* 294: 1849-1850.
- Lawrence, S. and C.L. Giles. 1999. Accessibility of information on the Web. *Nature* 400: 107-109.
- Lawrence, S. and C.L. Giles. 1998. Searching the World Wide Web. *Science* 280: 98-100.
- Maurer, S.M. and B.A. Huberman. 2000. Competitive Dynamics of Websites, www.hpl.hp.com/shl/
- McPherson, M. 1983. An ecology of affiliation. *American Sociological Review* 48: 519-532.
- Machill, M. et al. 2004. Navigating the Internet A Study of German-Language Search Engines. *European Journal of Communication* 19: 321-348.
- Podolny, J.M.; T.E. Stuart and M.T. Hannan. 1996. Networks, Knowledge, and Niches: Competition in the Worldwide Semiconductor Industry, 1984-1991. *American Journal of Sociology* 102: 659-689.
- Porter, M.E. 1986. *Competition in global industries*. Harvard Business School Press, Boston, Mass.
- PR Newswire 1998, March 4. Wired Digital Unveils HotBot 4.0, The Wired Search Center, With Enhanced Search Services and Smarter 'SuperSearch' Interface for Serious Web Searches. Downloaded on May 15, 2006 from <http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/3-4-98/428395&EDATE=>
- Rindova, V., and S. Kotha. 2001. Continuous Morphing. *Academy of Management Journal* 44: 1263-1280.

Schmalensee, R. and R. Willig (Eds.) 1989. *The Handbook of Industrial Organization*. Elsevier Science, The Netherlands.

Sonnenreich, W. and T. Macinta. 1998. *Web Developer.com Guide to Search Engines*. Wiley, New York.

Smith, A. 1776. *An inquiry into the nature and causes of the wealth of nations*. London: Strahan & Cadell.

Vaughan, L. and M. Thelwall. 2004. Search Engine Coverage Bias: Evidence and Possible Causes. *Information Processing & Management* 40: 693-707.

Vermeulen, I., J. Bruggeman. 2004. Competition and differentiation as an evolving network. *XXIV International Sunbelt Social Network Conference*, Portorož, Slovenia.