The Social Factor of Open Science

Tobias Fries

Abstract Increasing visibility in the Internet is a key success factor for all stakeholders in the online world. Sky rocketing online marketing spending of companies as well as increasing personal resources in systematic "self-marketing" of private people are a consequence of this. Similar holds true for the science and knowledge creation world—here, visibility is also a key success factor and we are currently witnessing the systematic exploitation of online marketing channels by scientists and research institutes. A theoretical base for this novel interest in science marketing is herein provided by transferring concepts from the non-science online marketing world to the special situation of science marketing. The article gives hints towards most promising, practical approaches. The theoretical base is derived from considerations in the field of scale-free networks in which quality is not necessarily a predominant success factor, but the connectivity.

Introduction

New aspects of Web 2.0, together with those that are already familiar, are about to completely revolutionize the world of academic publishing. The gradual removal of access barriers to publications—like logins or unavailability in libraries—will increase the transparency and accordingly the use of Web 2.0 elements. We can envisage evaluation and suggestion systems for literature based on such factors as relevance and reputation along similar lines to search engines. Conversely, while it is conceivable that networking systems and search engine technology will consequently become increasingly prone to manipulation, this will at the same time be preventable up to a certain point.

T. Fries (✉)
Witten/Herdecke University, Witten, Germany
e-mail: Tobias.Fries@gmx.de

The Transition from Science to Open Science

Admittedly some way behind the consumer sector, the field of science is now beginning to grow accustomed to the Internet. Certain ingrained dogmas that previously stood in the way, like a general sense of apprehension on the grounds of it being "unscientific" are being cast aside—slowly but surely—to reveal the innovative concepts, even though this is proving to be a slightly hesitant process. Just how hesitant becomes clear when we take a look at the use of Wikipedia, by way of an example. Wikipedia is admittedly not a primary source of information and calls for considerable caution when quoting, but this doesn't make it apply any less to the seemingly objective conventional science. This is not meant as a disparaging remark or an insult to the significance of scientific relevance but merely serves to point out comparable dangers. We cannot judge the objectivity of unofficial editors who contribute to the Social Internet on a voluntary basis any more than we can assess the source of funds used to sponsor an academic study. It is in any event wise to exercise basic level of caution when employing it for any objective purpose.

Anyone—including scientists—who acquainted himself/herself with the new media early on is already at a great advantage, even now. The earliest pioneers were able to send the treatises to and from much more frequently via email than with the traditional postal system, and this in turn considerably shortened editing cycles, and whoever dared to blow caution to the wind and post his/her text on the Internet, despite any fears of data theft, was rewarded with tangibly higher citation rates. The reasons for this are intuitively plausible to anyone who has ever carried out research work him- or herself: we only quote what we find. No matter how brilliant an unavailable text may be, if it is unknown, nobody will cite it. This conclusion can be drawn by taking into account the work of de Solla Price (1976), who analyzed Cumulative Advantage Processes on the example of paper citations. By showing that the Science Citation Index only used to consider 1573 sources out of 26000 journals of interest, he calculated a possible total reach of 72 % by having access to only the first 6 % of journals. He also found out that a longer presence in the archives increased citation rates as a result of higher potential availability. To transform de Solla Prices findings into other words, it is not the content alone that leads to its subsequent utilization in academic circles but also the extent of its reach. This correlation can also be expressed in the form of an citation conversion equation, by dividing the number of quotes (k) by the number of times a text is read (n):

$$k_{Cit} / \ n_{Read} = \ C \qquad (1)$$

Factor C is used here to denote the citation conversion rate, a coefficient already familiar from conventional web analysis. The conversion rate refers to the ratio between the number of orders placed with online shops and the overall amount of web traffic or the number of baskets/shopping carts filled without proceeding to the check-out. It serves as an indication of the quality of the website which, by deduction, can also provide a straightforward analogy to the quality of the academic publication. Although the concept of the quality of academic publications

has been around for decades, a reliable evaluation has yet to be realized, because no-one so far has managed to track the frequency with which they are read. State-of-the-art performance assessment of academic articles is largely restricted to the quantitative number of citations and, very occasionally, the publication location and the circulation of the journals and books can, with considerable limitations, provide some clues as to the distribution. Acceptance for publication is, however, much more subjective than academia would like to admit. Publication frequently depends on personal contacts, political considerations or the scientific expertise of small editing committees, who may not necessarily fully recognize the significance of a treatise that is indeed new from an academic point of view.

By contrast, academics who post their publications on the Internet in an open, search-friendly format, stand a better chance of being quoted which, given the specific rationale of de Solla Price findings, consequently creates a self-perpetuating effect with every additional citation. Anyone who can be accessed is absolutely bound to be read more frequently, and even where the quality is inferior, may well be quoted less often (in relative terms) but more frequently from the point of view of the absolutely greater number of readers. Since the absolute citation rate in the current status quo is one of the key indicators of quality, a citation volume derived from the high statistical figures leads to a real perception of quality, possibly even when there are better articles on the same topic.

In their own interests, anyone who has grasped this concept is hardly likely to hide the treatises they have written behind the log-ins of a personal homepage, for which there may be a charge, or the internal domains of publishing companies' websites. Academic publications are not a mass product, and anyone who wants to earn money on the basis of the print run would be better off with books dealing with sexuality, entertainment or automotive topics, as these subjects regularly attain a publication run of a million copies or more. In so far as academics are ever involved in their field of interest for money, this is earned indirectly through lectures, consultation fees or application products derived from scientific research, to which the publication itself only contributes the legitimizing reputation. Browsing the Internet with the help of Google Scholar, Google Books or any other search engine while restricting one's search to documents with the ending.pdf will nowadays turn up a large number of academic publications whose full text is extremely specific and its perusal accordingly highly productive for one's own publications. Some publishers and authors even go as far as placing long passages of their books verbatim on online book portals like Amazon, specifically for search purposes, which a good many academics employ for the inclusion of such books that might not otherwise have been selected for citation. There is undoubtedly a conflict of goals and it is proving to be a problem for quite a number of researchers today: the more pages of a book or publication are read in a public domain, the fewer copies are going to be sold. If we regard the turnover achieved with the product and the number of book sales as a yardstick for measuring success, then this viewpoint is justified. If we go one step further, however, to the level of reach or prominence they gain, the number of sales is of absolutely secondary importance.

Barabási/Albert provide a next level of scientific explanation for this correlation with their analysis of networks, which can be regarded as the indisputable standard in network research—at least in terms of the number of recorded citations (cf. Barabási and Albert 1999). At the time of writing this article, Barabási had been quoted by more than 12,000 other scientists, according to Google Scholar. The content of the article deals with the clustering behavior of nodes in scale-free networks. This designation refers to graphs displaying structures that resemble their own but on a different scale: in other words, their structures look similar when enlarged or decreased in size. Another feature of scale-free networks is an exponential function in the number of clusters. Conventional random networks typically display a bell-shaped curve in their distribution function, according to which most nodes have a similar number of links and, in the boundary areas of the bell-shaped curve, tend to be those that deviate from this mean (cf. Erdős and Rényi 1960). Examples of conventional random networks include transport or electricity networks, as depicted in Figs. 1 and 2 for the European high-speed railway network.

During their investigation, Barabási and Albert (1999) looked at Internet links and initially assumed a random network of the kind introduced by Erdös and Rényi (1960). They were, however, surprised to discover an exponential distribution function rather than a bell-shaped curve. This contained a small number of websites (nodes) which were linked to an extremely large number of other
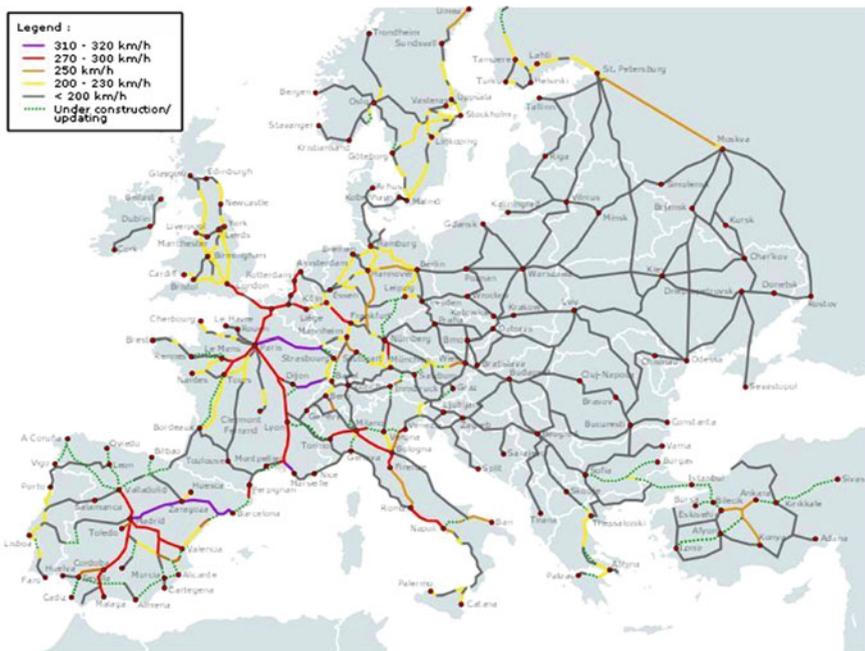

Fig. 1 The high-speed railway network in Europe. It represents a random network as proposed by Erdős and Rényi. (*Source and copyright of the diagram* Akwa and Bernese media and BIL under a creative commons licence)
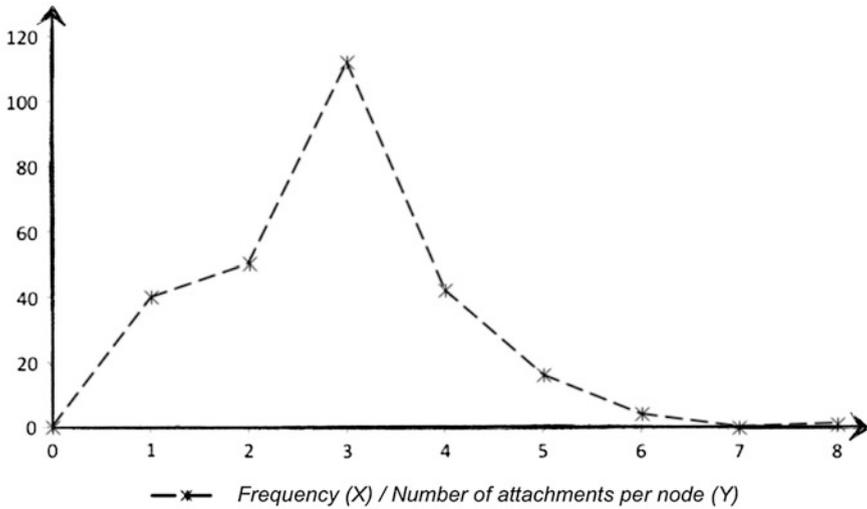
Fig. 2 The frequency distribution of the links between the nodes of the European high-speed railway network. The density function superimposed on the *dots* illustrates the underlying Poisson distribution

websites and an extremely large number of websites to which just a very few other sites pointed. Taking their research further, they came across a probability mass function in other scale-free networks, with the help of which it was easier to find well-attached nodes, so they were linked up more often. They called this phenomenon "Preferential Attachment" and succeeded in proving that clustering dynamics of this kind give rise to ever-increasing imbalances in the network system over a longer period of time, so that well-attached nodes accumulate further links whereas less well-attached nodes attract even fewer new links. Barabási later coined the very apt designation "The Rich get Richer Phenomenon" to describe this effect (cf. Barabási and Bonabeau 2003, p. 64f).

Applying this investigation to science, it confirms the hypothesis proposed at the beginning of this treatise that reach, rather than quality, can lead to truth—particularly in the case of low visibility and accordingly less likelihood of the better essay being linked. Both de Solla Price and Barabasi/Albert identified exponential distribution functions in scientific citation, one called the phenomenon "Cumulative Advantage Processes", the other one "Preferential Attachment".

There are nevertheless far-reaching discussions currently in progress in academic circles regarding free access to publications, intellectual property rights and the supposed protection of scientific independence. The conflicting goals of scientists and publishers might be interpreted as the reason behind these discussions. Scientific advancement does not necessarily have to be the main priority of the publishing companies, seeing as they are financing an infrastructure and pay its owners dividends. They themselves do not have a share in the indirect income raised through the reputation of the scientist concerned, so the publishers' main interest must lie in the

direct generation of revenue, while intellectual property rights and limited access safeguard their influence. By contrast, if we take a closer look, the researcher has an interest in barrier-free accessibility and is therefore trying to break free of the publishers' sphere of influence, although they give him the benefit of the Open Science network. The fact that discussions of this nature are already in progress could be construed as friction due to a change of paradigms, because science could become more objective, more transparent and consequently more democratic, offering impartial benefits to researchers and research outcomes alike, if it were organized properly and supported by the rigorous use of certain aspects of the social web. To ensure that publishers don't come away empty-handed from a turn-around of this kind, they would be well advised to grasp the mechanisms behind this trend—the sooner, the better—and help to shape the framework in which science will be operating in the future through timely investments in the infrastructure for this Science Web 2.0. The fact that those who get noticed early on benefit from long-term reach applies to publishers as well, because a publishing company's reputation—and accordingly being published within its infrastructure—already attracts the better scientists; an innovative form of publication will make no difference to the fundamental research results achieved by de Solla Price and Barabási/Albert. The early bird catches the worm, regardless of whether that bird is a publisher or a researcher.

The Importance of Network Science for Research Relevance

There are only a few research results that can be described as "objectively good". It is usually those publications about which enough other people speak favorably in terms of that particular topic that are perceived as being "good". The few contributions to research that are objectively and indisputably good receive favorable feedback anyway, due to the excellent quality of their work, provided they attract a sufficiently wide audience. More often than not, however, work that is merely mediocre receives positive feedback because its stands out from the rest or because the author already enjoys a good reputation. The most obvious forms of favorable feedback in academic circles are citations and references. And the higher the number, the greater the likelihood of being quoted again in the future, since it is easier to find a frequently cited article than one that has been cited less often. So the "Rich get Richer" phenomenon applies to researchers, too. On a practical level, this explains why it is much easier for a professor who has been publishing articles for many years to get his next paper published in an acclaimed journal than a young person, even though he might be the more brilliant of the two, who is trying to make his debut with an outstanding first work. At the same time it becomes clear why it may be of an advantage for the brilliant young individual to ask the professor for his support in publishing his first article, as this would allow the young author to establish his first links via the science network while the older research might profit a little from the spillover effect of the excellent debut article.

Seen as a whole, this section serves to explain why publishing scientists have an interest in being quoted as frequently as possible. It is equally conceivable that there are sufficient reasons for manipulating citations, always with the aim of being cited more often by others. A number of scientists managed to solve this problem in the past by means of citation cartels, and in case this is controversial, let it be said that it definitely applies to some operators of websites at least, because it constituted a problem for search machines for a long time.

The following section addresses feasible aspects of the Social Internet for Open Science, always against the backdrop of possible manipulation and the ensuing consequences in practice. This list, like the whole book, is an incomplete preview of a topic that is still under development but will revolutionize science.

Aspects of the Social Web in Open Science

Users particularly welcome innovations when they hold the promise of an advantage of some kind. Individual features embedded in otherwise linear platforms will be just as unsuccessful as those that regularly fail nowadays in consumer applications. The intelligent consolidation of information to create an outcome that is of broad use to all stakeholders involved will assert itself unless an inferior, but high-reach solution achieves exclusive prominence in the eyes of the users. For this reason, the aspects set out below can only realize a fraction of their collective, self-multiplying effect. They are nevertheless being included individually for the sake of maintaining the linear structure of the text.

The Basic Principles of Search Engines: Relevance and Reputation

The purpose of search engines is to present results of relevant Internet searches, arranged according to their reputation. Relevance is assessed by subjecting the available contents and the contents of other referring websites to a special quantification process, while the reputation is determined from a wide range of other aspects, beginning with the number of referring websites and including users' appraisals and the surfing patterns of people visiting the websites. Due to the widespread manipulation of search results, search engines have long since moved away from metrics based solely on the number of inbound links—a practice which, despite similar manipulation incentives, has not yet caught on in the field of science, where the number of citations still prevails as the standard benchmark.

The operators of Open Science platforms accordingly have a similar responsibility to that of search engines. In present-day research, scientists already have to select from the literature pertaining to their specific sphere of research. Purely determining the topical relevance is the simplest task; assessing the technical relevance is much more difficult. In cases of uncertainty or other causes of hesitation, the

researcher will also take the reputation of a scientist into account when contemplating who to quote. Each and every piece of research is restricted by the subjective field of vision of that which the scientist finds. Publications that escape his or her notice due to a language or access barrier will not be included in the shortlist.

The challenge in setting up Open Science is to achieve as comprehensive a selection of scientific texts as possible with the lowest possible access barrier, to enable publishing scientists to obtain the desired level of relevance and reputation with the help of publication platforms. Merely widening the field of vision for researchers is the easiest task: increasing objectivity and transparency in the assessment of reputation will prove to be a challenge. Bearing the evolution of search engines in mind, we anticipate a highly dynamic advancement and permanent alignment of the algorithms employed.

Identification of the Protagonists

The fact that there are sometimes two or more scientists with the same name can be misleading, but this is not so serious that it renders research impossible, although common names do make it more difficult to identify which researcher is meant, especially when they are both active in the same field. With the Open Science Web approach, every scientist can be allocated an unambiguous profile, complete with photo, a brief CV, main research focus and, in particular, a specific ID number. Existing Science Communities like ResearchGate,[1] which already provide a representative picture, seem to be particularly suitable. Profiles with an open platform identification number enable an integrated use of Open Science features, as introduced below. The open-platform architecture is of particular importance. A platform such as ResearchGate, for instance, will gain a strategic lead as far as reach is concerned, along similar principles to those presented above, if it makes its own researcher ID available on the Internet for other academic purposes free of charge and without any barriers. The permanent core reference to this platform will lead every researcher back to the original community—a long-term benefit in terms of influence and reputation cannot actually be foreseen at the moment, due to the growing number of members, but it is highly probable when we consider the dynamic evolution of scale-free networks.

Ascertaining a Quality Factor from Likes, Dislikes, Assessments and Comments

One of the most straightforward uses of the Social Web 2.0 for Open Science is the ability to transfer positive and negative ratings and comments. These features

---

[1] www.researchgate.com.

are not new by any means, the most prominent among them being those employed by Facebook, but even there they were not new, owing to their simplicity. Blogs, online book marketplaces and bidding platforms recognized the principle of assessment for boosting one's reputation at a much earlier stage and used it to their own advantage. It can basically be divided into simple expressions of approval or disapproval (like or dislike), an interesting aspect being that Facebook only allows the affirmative "like" vote which, judging by the demographic structure of its users, may be of inestimable value in protecting the psychological development of school-children/minors, seeing as countless cases of cybermobbing have been heard of even when the voting is limited to favorable "thumbs up" ratings. Although science ought to be regarded as objective and rational, it would be wrong to underestimate the interests that lie behind research results, and which might play a role in influencing the assessment of publications beyond the limits of objectiveness. Only a process of experimentation and subsequent evaluation can determine whether the accumulation of negative votes leads to an objective improvement in assessment or encourages the intentional underrating of undesirable research results. In the interests of transparency, however, it would probably make sense to show features of this kind with a clear, publicly visible reference to the originator. In this way, likes, dislikes, assessments and comments would reflect straight back on the reputation of the person passing the criticism and would consequently be better thought-out than anonymous comments. This contrasts starkly with the fear of uncomfortable, but justified truths which are more easily expressed anonymously. It might be possible to experiment with both forms in order to ascertain a quantified quality factor that would also be taken into consideration in evaluating the reputation of an article or researcher.

Crowd Editing

The possibility of crowd editing is a completely new, feasible feature in the Open Science web. Strictly speaking, it amounts to the steadfast further development of joint publications. While the ideal number of academics working on a treatise is limited to two, three or occasionally four scientists, crowd editing opens up a publication on a general level. As introduced earlier on in this book (see chapter Dynamic Publication Formats and Collaborative Authoring), anyone reading the article can contribute voluntarily to it provided he/she has something relevant to add. Old versions can remain stored in archives, as is the case with Wikipedia articles, and subject-related changes can be either approved or rejected by a group of editors before an official new version of the article is published. It is conceivable that the relevant subversion could be cited—not really a new procedure—but the principle of crowd editing might increase the frequency of amendments.

Suggestion Systems for Articles During the Writing Process

The potentiality of suggestion systems is, however, really new. Whereas authors today actively look for literary sources in conventional and digital libraries, innovative technologies enable smart suggestion systems. The insertion of context-based Internet advertising is a long-established practice, whilst its academic counterpart is still in its infancy. Only Google, in its capacity as trailblazer of search-engine technology, already proposes search-related topics and authors, thus paving the way for the intelligent linking of academics and their publications.

It starts to become exciting when suggestions for potentially interesting, subject-related articles are put forward during the actual writing process. This might to a certain extent release researchers from the somewhat less intellectual task of merely compiling information while simultaneously providing them with additional sources, which they might not have found so easily on their own, since they are only indirectly linked to the topic in question via another association, for instance. Special attention should be paid, when developing the relevant technologies, however, to the selection algorithm, which harbors the risk of tempting the researcher into a convenience trap. The mental blanking out of other sources might represent one aspect of a trap of this kind—a phenomenon that is likewise rooted in the network theory. In this case, the sources that attract most attention are those that are closest to the interests of the researcher in question and are already most visible (cf. Barabási and Albert 1999). The predefined ranking of pop-up results is another hazard. There are countless analyses of the recorded click rate for search results using the Google search engine. Various analysis in Google Analytics reports conducted over several years have repeatedly provided a similar picture—about 80 % of all clicks landed on the first five search results that appeared on the screen, 18 % on the remaining ones on the first page and only 2 % on the second page. This data has been retrieved by comparing search statistics with click statistics. Due to their previous experience and working routines, one can assume that academics conduct their research more thoroughly than general consumers. Nevertheless, such attributes as convenience and circumstances like being in a hurry are only human and also apply to a certain extent to researchers, which bodes quite well for the first secondary sources in the list, at least.

Against this backdrop it emerges what a high priority status the algorithm will have with regard to the presentation of suitable secondary literature. Due to the great resemblance in structure, we assume that this feature will operate along much the same lines as search engines, so it is likely to face similar challenges and problems. We will revert to this topic further down, in the section dealing with the presentation of results.

Once these technical problems have been solved satisfactorily, we can envisage a completely new form of academic writing, along the lines of the example outlined briefly below:

Example of Academic Writing

A researcher has an idea for an article and already possesses some previous knowledge of the subject-matter, which allows him to put his idea into words straight away. So, using a web application designed specifically for academic writing, he begins to type his idea into the space provided. Since he is logged in, the platform is not only able to create direct references to his previous work and topics processed on the platform but can also read his current input and compare it with texts contributed by other scientists. While he is writing, the researcher can now view context-related excerpts on the screen next to his own text, which might be of interest for the passage he is writing. Other, more general articles dealing with the subject concerned, which might be of relevance to this treatise, appear elsewhere. Based on the topics and contributions evaluated on the platform, the researcher in this particular example also receives suggestions as to which other scientists he should contact for the purpose of exchanging information and views.

This case illustrates a scenario of higher transparency on several levels. Besides those relating to texts, the researcher also receives suggestions relating to people who might prove to be an interesting point of contact. This might conceivably be extended to announcements for specialist conferences, other relevant events or items that match the theme.

Parallels to Search Engines

Expressed in simplified terms, search engines consist of a crawler and an indexer. The crawler visits websites, records the contents and proceeds to the next website via the links provided. There is also a so-called scheduler designed to determine the order in which the next sites will be visited by arranging the links according to priority. The indexer orders the recorded contents and allots them priority for displaying in the lists of results. The precise technical principle is of secondary importance in this context, but a general outline of the analogical derivation process is no doubt useful. To sum up, a search engine arranges results according to their relevance and reputation. It is this principle that will be crucial for a suggestion system in Open Science, too; other technical features that will also be essential for a suggestion system include a crawler, an indexer and a scheduler, thus displaying numerous parallels between search engines and social science in the Open Web.

The background is that, even where the content is of equal relevance, one or more additional coefficients are needed to determine which results appear at the top of the list. These might be such dimensions as frequency of citation, the number of favorable comments and maybe even comments posted by other highly rated scientists. These other dimensions may be varied and, in the interests of maintaining a high standard of output, subject to dynamic change. This is due to the high probability of leading, and implicitly more relevant, search results being used more often for quoting, so scientists strive to optimize their own input.

Similar Problems to Those of Search Engines

The history of search engines is dotted with attempts to influence this process—initially through the frequent repetition of keywords taken from the body of the text. Since this was easy for the author himself to manipulate, the quality of an assessment based primarily on this factor was fairly meaningless. For this reason, external criteria such as the number of links from other websites were added, but they were also easily influenced by means of self-developed networks. We have observed a kind of cat and mouse game between search engines and so-called search engine optimizers over the past 15 years. These SEOs began by inserting a large number of keywords on their websites, which led to the search engines introducing a kind of maximum quota. Everything over and above that quota was classified as spam and greater importance was ascribed to the number of incoming links. So the SEOs began devising their own website structures that pointed to the target sites to be optimized. Search engines consequently began to evaluate the number of different IP numbers as well, so the SEOs retaliated by setting up different servers, whose sites highlighted the target sites in the shape of a star or a circle. And we could add many more examples to his list. Similar developments are to be expected in the scientific sphere, particularly as the setting up of citation networks is nothing unusual even in traditional academia. What does need to be solved is the problem of avoiding cartels of this kind and it is essential that we learn as much as possible from past experience with search engine optimization.

Similar Solutions to Those of Search Engines

Solutions in the field of search engine technology are increasingly permeating the domain of network science. Analyzing typical and atypical linkages has now advanced so far that it can determine with reasonable probability whether a more or less naturally evolved linking network is behind a certain website or whether there are numerous links bred on search engine optimizers' own farms. The solution is not yet complete but the number of very crude manipulations has receded noticeably during the past few years, as Google and other search engines were evaluating search engine positions for those detected. Similar occurrences are to be anticipated in the academic sphere of Open Science. In such areas where network references are unmistakably concentrated in denser clusters than the extent of the subject-matter would normally justify, an algorithm will be employed to reduce the reputation factor to a natural size. Search engines meanwhile go one step further and remove excessively optimized sites completely from the index, a move that can only be reversed by dismantling the linkage cartel or stopping the manipulations. Whilst the hitherto anonymously functioning search engines are only just beginning to identify users in the registered domains and to incorporate their search and surf patterns in the reputation assessment process, this has been

common practice in the publication of scientific treatises on the social web right from the start due to the clear authentication system described above. This has the added advantage of being able to include commenting and rating behavior, and possibly even the amount of time spent on a page of a treatise, in the reputation assessment of an article. It is not possible to forecast the entire range of potential manipulations as yet, and a certain amount of reciprocal technological upgrading is also to be anticipated in academic circles—in the interests of unbiased, relevant results on the one hand and motivated by a desire for upfront placements, which hold the promise of additional citations, on the other.

References

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509–512.

Barabási, A.-L., & Bonabeau, E. (2003). Scale free networks. *Scientific American* 60–69.

Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5,* 17–61.

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5–6), 292–306