DOI:10.1145/1378704.1378720

Why Wikipedia's remarkable growth is sustainable.

BY DIOMIDIS SPINELLIS AND PANAGIOTIS LOURIDAS

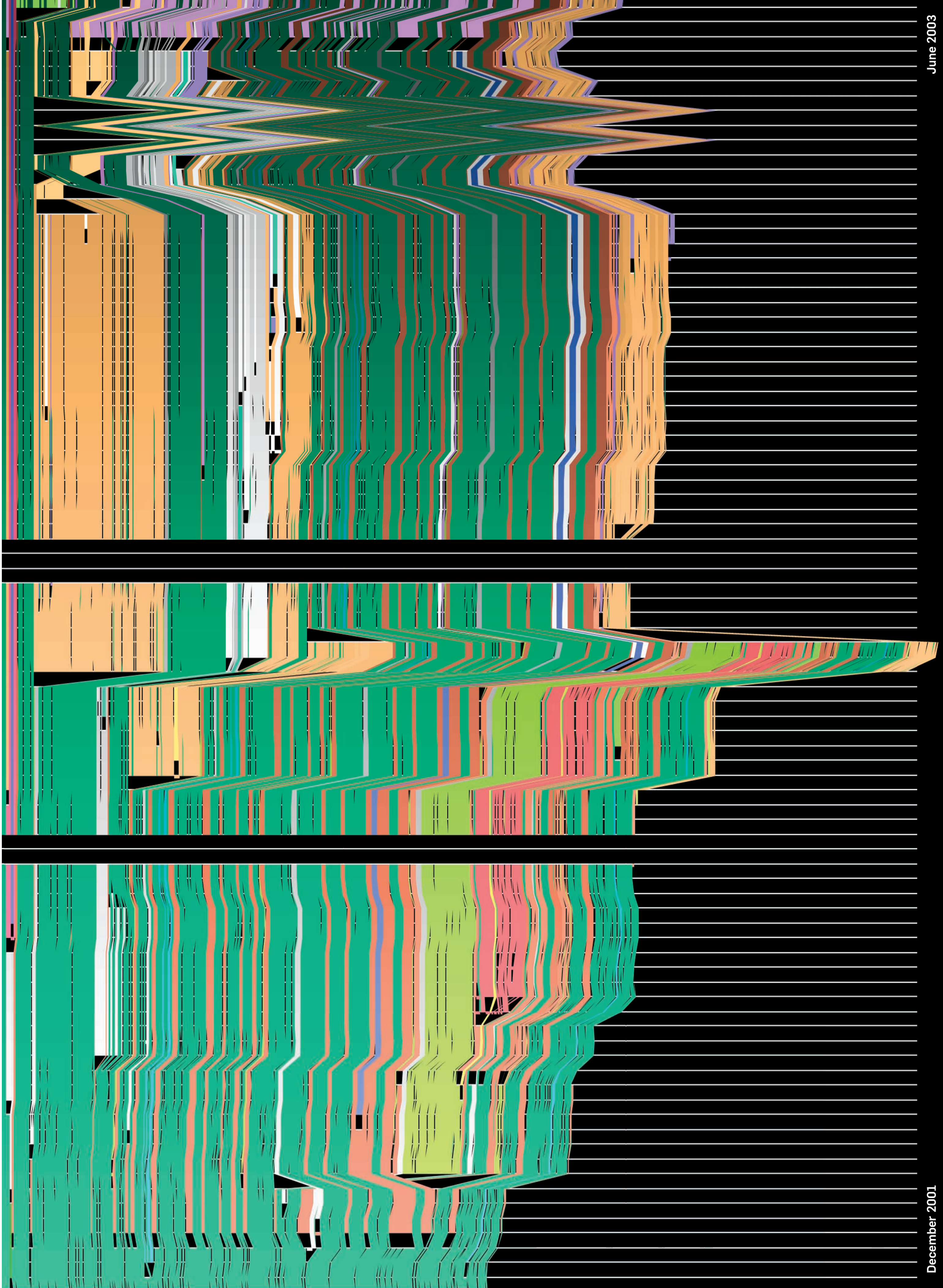
The Collaborative Organization of Knowledge

WIKIPEDIA (WWW.WIKIPEDIA.ORG) is a freely available online encyclopedia anyone can edit, contributing changes, as well as articles.[10] With more than a million entries, hundreds of thousands of contributors, and tens of millions of fully recorded article revisions, Wikipedia's freely available database has also made it possible to study how human knowledge is recorded and organized through an open collaborative process. Although citation analysis[6] can establish how new research builds on existing publications, the fully recorded evolutionary development of Wikipedia's structure has allowed us to examine how existing articles foster development of new entries and links. Motivation for our longitudinal study of Wikipedia evolution followed from our observation that even though Wikipedia's scope is increasing, its coverage is apparently not deteriorating. To study the process of Wikipedia growth we downloaded the February 2006 snapshot of all recorded changes and examined how entries are created and linked. Inspecting the timestamps on individual entry definitions and references, we found that links to nonexistent articles often precede creation of new articles. Also, tracking the evolution of article links allowed us to empirically validate Barabási's hypothesis on the formation of scale-free graphs through incremental growth and preferential attachment.[1] Our findings paint a picture of sustainable growth, suggesting that Wikipedia's development process delivers coverage of more and more subjects.

The phenomenal growth of Wikipedia is attributable to a mixture of technologies and a process of open participation. The key technology behind Wikipedia is that of a Wiki—online lightweight Web-based collaboration.[4] Wikipedia content appears online as static HTML pages, though each such page includes an edit button anyone can use to modify its content; editing most articles requires no prior authorization or arrangement. The system maintains the complete edit history of each page and supports a "watchlist" mechanism that alerts registered users when a page they are interested in changes.

The page history and watchlist facilities promote low-overhead collaboration and identification of and response to instances of article vandalism. We found that 4% of article revisions were tagged in their descriptive comment as "reverts"—the typical response to vandalism. They occurred an average of 13 hours after their preceding change. Looking for articles with at least one revert comment, we found that 11% of Wikipedia's articles had been vandalized at least once. (The entry for George W. Bush had the most revisions and reverts: of its 28,000 revisions one-third were reverts and, conceivably, another third vandalism.) Articles prone to vandalism can be administratively locked against revisions, a step rarely taken; in our study only 0.13% of the articles (2,441 entries) were locked.

STUDYING COOPERATION AND CONFLICT BETWEEN AUTHORS WITH HISTORY FLOW VISUALIZATIONS. FERNANDA B. VIÉGAS, MARTIN WATTENBERG, AND KUSHAL DAVE. PROCEEDINGS OF SIGCHI 2004. HTTP://ALUMNI.MEDIA.MIT.EDU/~FVIEGAS/PAPERS/HISTORY_FLOW.PDF


June 2003
December 2001

When edited, an entry's content doesn't use the Web's relatively complex, error-prone HTML syntax but rather a simplified text annotation scheme called wiki markup, or wikitext. Creating a link from one entry to another is as simple as enclosing the other entry's identifying name in double square brackets. Markup tags can also group together related articles into categories (such as "Nobel laureates in physics," "liberal democracies," and "bowed instruments"). One use of a category tag is to mark entries as stubs, indicating to readers and future contributors that a particular entry is incomplete and requires expansion. In the snapshot we studied, about 20% of the entries were marked as stubs. For a better idea of Wikipedia's process and technology, access an entry in your own specialty and contribute an improvement.

Existing research on Wikipedia employs descriptive, analytic, and empirical methodologies. A series of measurements has been published that identifies power laws in terms of number of distinct authors per article, articles edited per author, and ingoing, outgoing, and broken links.[13] On the analysis front, notable work has used simulation models to demonstrate preferential attachment,[3] visualization techniques to identify cooperation and conflict among authors,[12] social-activity theories to understand participation,[2] and small-worlds network analysis to locate genre-specific characteristics in linking.[8] Finally, given the anarchic nature of Wikipedia development, it is not surprising that some studies have also critically examined the quality of Wikipedia's articles.[7,11] The work we describe here focuses on the dynamics of Wikipedia growth, examining the relationship between existing and pending articles, the addition of new articles as a response to references to them, and the building of a scale-free network of articles and references.

We hypothesize that the addition of new Wikipedia articles is not a purely random process following the whims of its contributors but that references to nonexistent articles trigger the eventual creation of a corresponding article.

Methods

The complete content of the Wikipedia database is available online in the form of compressed XML documents containing separate revisions of every entry, together with metadata (such as the revision's timestamp, contributor, and modification comment). We processed the February 2006 complete dump of the English-language Wikipedia, a 485GB XML document. (In June 2008, we looked to rerun the study with more recent data, but complete dumps were no longer available.) The text of each entry was internally represented through the wiki-specific annotation format; we used regular expressions and explicit state transitions in a flex-generated analyzer for parsing both the XML document structure and the annotated text. From the database's entries we skipped all entries residing in alternative namespaces (such as "talk" pages containing discussions about specific articles, user pages, and category pages). In total, we processed 28.2 million revisions on 1.9 million pages.

For each Wikipedia entry we maintained a record containing the contributor identifiers and timestamps for the entry's definition and for its first reference, the number of efferent (outgoing) article references (unique references to other Wikipedia articles in the current version of the entry), the number of unique contributors, the number of revisions, a vector containing the number of the entry's afferent (incoming) references from other Wikipedia articles for each month, and a corresponding vector of Boolean values identifying the months during which the entry was marked as a stub. (The source code for the tools we used and the raw results we obtained are at www.dmst.aueb.gr/dds/sw/wikipedia.)

Growth and Unresolved References

We were motivated to do this research when one of us (Spinellis), in the course of writing a new Wikipedia entry, observed that the article ended up containing numerous links to other nonexistent articles. This observation led us to the "inflationary hypothesis" of Wikipedia growth, that is, that the number of links to nonexistent articles increases at a rate greater than the rate new articles are entered into Wikipedia; therefore Wikipedia utility decreases over time as its coverage deteriorates by having more and more references to concepts that lack a corresponding article. An alternative—the "deflationary hypothesis"—involves links to nonexistent articles increasing at a rate less than the rate of the addition of new articles. Under this hypothesis we are able to project a point in the future when the Wikipedia engine of growth (discussed in the next section) will stall.

It turns out that the reality of Wikipedia development is located comfortably between the two extremes of nonexistent link inflation and deflation. Figure 1 outlines the ratio between incomplete and complete articles from 2001 to 2006. Incomplete articles either don't exist in Wikipedia or exist but are marked as stubs. Although many stub articles contain useful information (often a link to an authoritative page with more detail), some pages also require additional work to be helpful but are not marked as stubs. For the purposes of our study we assume that the two effects cancel each other out.

The covered ratio from 2003 to 2006 seems stable, with about 1.8 missing or stub articles for every complete Wikipedia article. During the same time the number of articles surged from 140,000 to 1.4 million entries, showing that the apparently chaotic Wikipedia development process delivers growth at a sustainable rate.

References Lead to Definitions

Wikipedia's topic coverage has been criticized as too reflective of and limited to the interests of its young, tech-savvy contributors, covering technology and current affairs disproportionably more than, say, world history or the arts.[5] We hypothesize that the addition of new Wikipedia articles is not a purely random process following the whims of its contributors but that references to nonexistent articles trigger the eventual creation of a corresponding article. Although it is difficult to claim that this process guarantees even and unbiased coverage of topics (adding links is also a subjective process), such a mechanism could eventually force some kind of balance in Wikipedia coverage.

The empirical findings outlined in Figure 2 support our hypothesis concerning the drive behind the addition of new articles. In particular, a reference to a nonexistent entry appears to be positively correlated with the addition of an article for it. Figure 2a tallies the number of articles with a given time difference between an entry's first reference and its subsequent definition. Most articles by far seem to be

Figure 1: Coverage of Wikipedia articles.

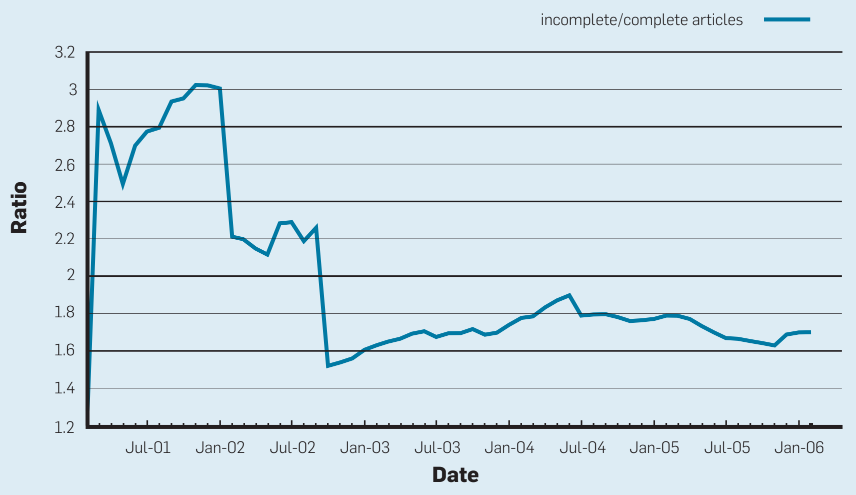


Figure 2a: References to an entry typically precede the entry's definition; number of entries with a given difference between the time of the first reference to the entry and the addition of its definition.

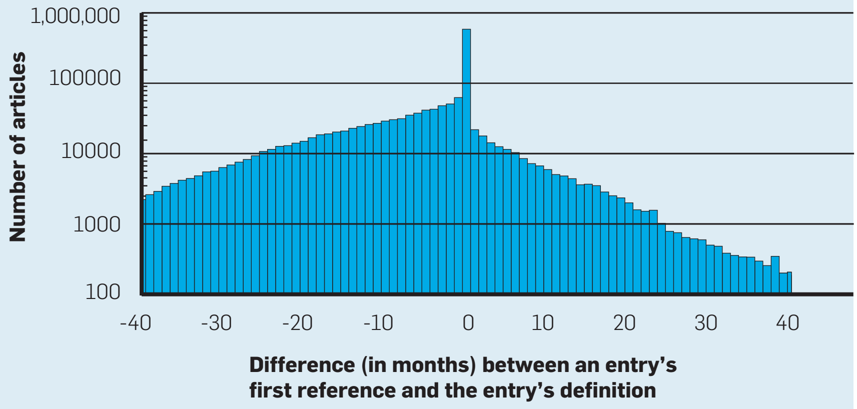
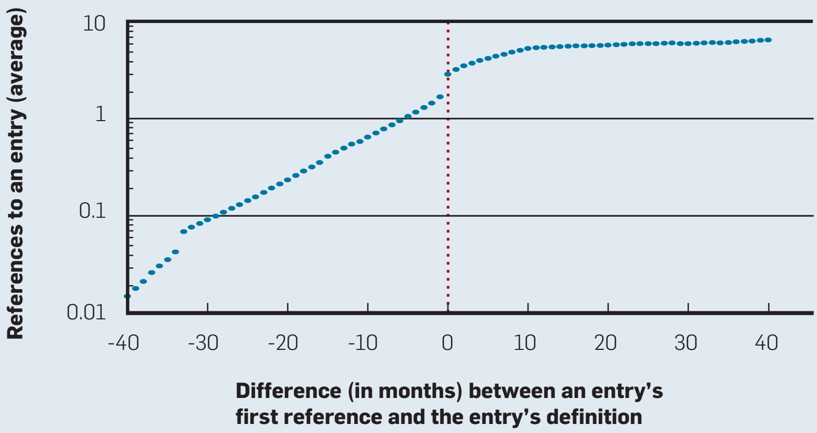


Figure 2b: References to an entry typically precede the entry's definition; number of references to an entry at the time of its definition.



created in the month of their first reference. Interestingly, the reference and subsequent definition of an article in Wikipedia appear to be a collaborative phenomenon; from the 1.7 million entries for which both the contributor entering the first reference and the contributor entering the first definition are known, that contributor is the same for only 47,000, or 3%, of entries.

Similarly, the mean number of first references to entries (see Figure 2b) rises exponentially until the referenced entry becomes an article. (For calculating the mean we offset each entry's time of definition and time points in which it was referenced to center them at time 0.) The point in time when the referenced entry becomes an article marks an inflection point; from then on the number of references to a defined article rises only linearly (on average).

Building a Scale-Free Network

We established that entries are added to Wikipedia as a response to references to them, but what process adds references and entries? Several models have been proposed to explain the appearance of scale-free networks like the one formed by Wikipedia's entries and references. The models can be divided into two groups:[9] treating power laws as the result of an optimization process; and treating power laws as the result of a growth model, the most popular of which is Barabási's preferential attachment model.[1] In-vitro model simulations verify that the proposed growth models do indeed lead to scale-free graphs. Having the complete record of Wikipedia history allows us to examine in-vivo whether a particular model is indeed being followed.

Barabási's model of the formation of scale-free networks starts with a small number (m_0) of vertices. Every subsequent time step involves the addition of a new vertex, with $m \leq m_0$ edges linking it to m different vertices already in the system. The probability P that a new vertex will be connected to vertex i is $P(k_i) = k_i/\Sigma_j k_j$, where k_i is the vertex's connectivity at that step.

The situation in Wikipedia is more complex, as the number of vertices and edges added in a time step is not constant and new edges are added between existing vertices as well. We therefore consider a model where at each time step t a month, a variable number of entries and r_t references are added. The references are distributed among all entries following a probability $P(k_{i,t}) = k_{i,t}/\Sigma_{j,t}k_{j,t}$, with the sums and the connectivities calculated at the start of t. The expected number of references added to entry i at month t is then $\{k_{i,t}\} = r_t P(k_{i,t})$. We find a close match between the expected and the actual numbers in our data. Figure 3a is a quantile-quantile plot of the expected and the actual numbers at the 1,000-quantiles; Figure 3b outlines the frequency distributions of the number of articles (expected vs. actual) gaining a number of references in a month. The two data sets have a Pearson's product-moment correlation of 0.97, with the 95% confidence interval being (0.9667, 0.9740). If na_x is the number of articles that gained x > 30 (to focus on the tails) references in a month and na'_x is the expected number of such articles, we have na_x 1.11na'_x(p-value < 0.001).

It has never been possible to examine the emergence of scaling in other

Figure 3a: Expected and actual number of references added each month to an entry; quantile-quantile plot of the expected and actual number of references added each month to each article.

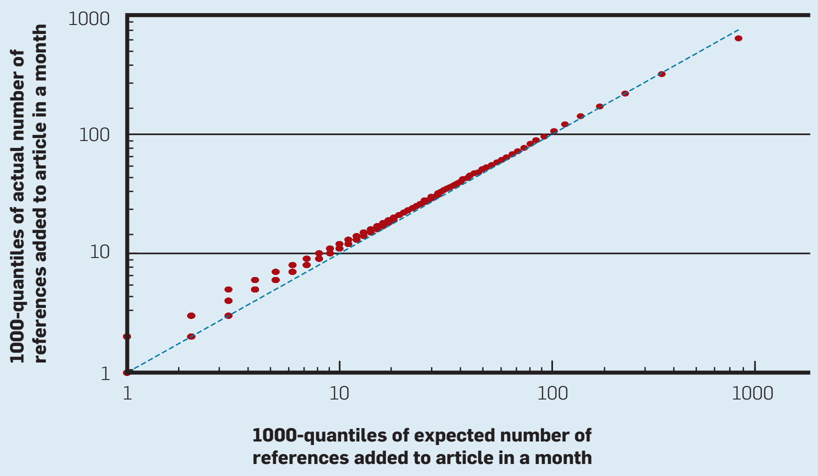
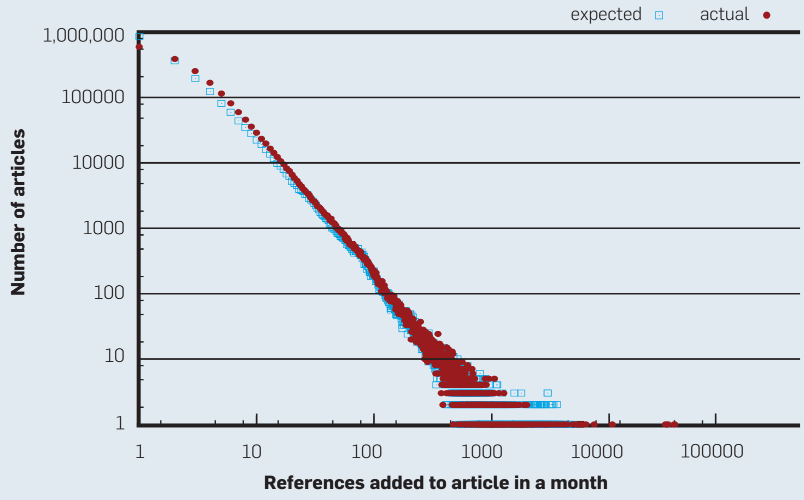


Figure 3b: Expected and actual number of references added each month to an entry; frequency distributions of the expected and actual number of references added each month to each article.



big real-world networks like the Web, as there is no full record of their evolution. Wikipedia now allows us to witness, and validate, preferential attachment at work on its graph.

Conclusion

The usefulness of an online encyclopedia depends on multiple factors, including breadth and depth of coverage, organization and retrieval interface, and trustworthiness of content. In Wikipedia more depth eventually translates into breadth, because the Wikipedia style guidelines recommend the splitting of overly long articles. The evolution of articles and links in Wikipedia allows us to model the system's growth. Our finding that the ratio of incomplete vs. complete articles remains constant yields a picture of sustainable coverage and growth. An increasing ratio would result in thinner coverage and diminishing utility and a decreasing ratio of incomplete vs. complete articles to eventual stagnation of Wikipedia growth.

The idea of growth triggered by undefined references is supported by our second finding—that most new articles are created shortly after a corresponding reference to them is entered into the system. We also found that new articles are typically written by different authors from the ones behind the references to them. Therefore, the scalability of the endeavor is limited not by the capacity of individual contributors but by the total size of the contributor pool.

Wikipedia's incremental-growth model, apart from providing an in-vivo validation of Barabási's scale-free network-development theory, suggests that the processes we have discovered may continue to shape Wikipedia in the future. Wikipedia growth could be limited by invisible subjective boundaries related to the interests of its contributors. Our growth model suggests how these boundaries might be bridged. Consider that references to nonexistent entries prompt creation of these entries and assume that all human knowledge forms a fully connected network. Wikipedia's coverage will broaden through a breadth-first graph traversal or flood-filling process, albeit over an uneven time progression.

It turns out that the reality of Wikipedia's development is located comfortably between the two extremes of nonexistent link inflation and deflation.

How far might the Wikipedia process carry us? In Jorge Luis Borges's 1946 short story "On Exactitude in Science," the wise men of the empire undertake to create a complete map of the empire; upon finishing, they realize the map was so big it coincided with the empire itself.

References

1. Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science 286*, 5439 (Oct. 15 1999), 509–512.
2. Bryant, S., Forte, A., and Bruckman, A. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (Sanibel Island, FL, Nov. 6–9). ACM Press, New York, 2005, 1–10.
3. Capocci, A. Servedio, V., Colaiori, F., Buriol, L., Donato, D., Leonardi, S., and Caldarelli, G. Preferential attachment in the growth of social networks: The case of Wikipedia. *Physical Review E*, 74, 036116 (2006).
4. Cunningham, W. and Leuf, B. *The Wiki Way: Quick Collaboration on the Web.* Addison-Wesley, Boston, MA, 2001.
5. Denning, P., Horning, J., Parnas, D., and Weinstein, L. Wikipedia risks. *Commun. ACM 48*, 12 (Dec. 2005), 152–152.
6. Garfield, E. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities.* John Wiley & Sons, Inc., New York, 1979.
7. Giles, J. Internet encyclopaedias go head to head. *Nature 438*, 7070 (Dec. 15, 2005), 900–901.
8. Mehler, A. Text linkage in the wiki medium: A comparative study. In *Proceedings of the EACL 2006 Workshop on New Text: Wikis and Blogs and Other Dynamic Text Sources* (Trento, Italy, Apr. 6, 2006), 1–8.
9. Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics 1*, 2 (2003), 226–251.
10. Remy, M. Wikipedia: The free encyclopedia. *Online Information Review 26*, 6 (2002), 434.
11. Stvilia, B., Twidale, M., Smith, L., and Gasser, L. Assessing information quality of a community–based encyclopedia. In *Proceedings of the International Conference on Information Quality* (Cambridge, MA, Nov. 4–6, 2005), 442–454.
12. Viégas, F., Wattenberg, M., and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria, Apr. 24–29). ACM Press, New York, 2004, 575–582.
13. Voß, J. Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics* (Stockholm, July 24–28, 2005), 221–231.

This work is partially funded by the European Commission's Sixth Framework Programme under contract IST-2005-033331 "Software Quality Observatory for Open Source Software (SQO-OSS)."

Diomidis Spinellis (dds@aueb.gr) is an associate professor of information system technologies in the Department of Management Science and Technology at the Athens University of Economics and Business, Athens, Greece.

Panagiotis Louridas (louridas@acm.org) is a software engineer in the Greek Research and Technology Network and a researcher at the Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece.