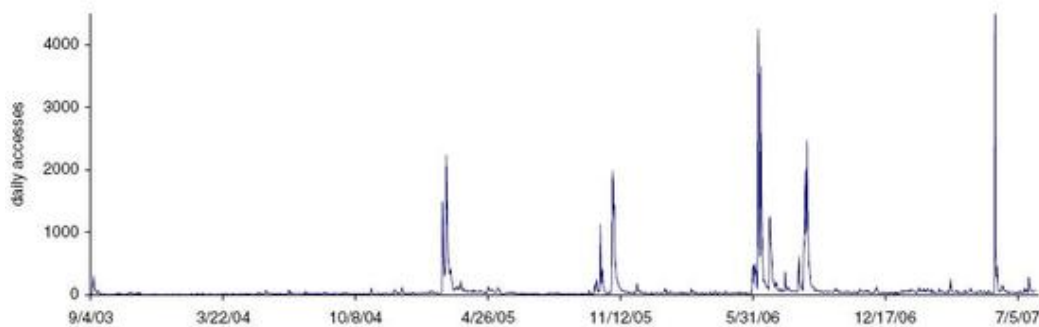


Scientists Construct Model of the World Wide Web



Traffic statistics for the Web page http://reverent.org/true_art_or_fake_art.html, where the long periods of low activity and short bursts of high activity are similar to other sites that the researchers analyzed. Credit: Simkin and Roychowdhury.

Although the Internet contains well over 100 million Web sites, two electrical engineers think they know what the traffic patterns of the entire Web look like.

Mikhail Simkin and Vwani Roychowdhury, electrical engineers from both the University of California Los Angeles and NetSeer.com, have constructed a model of the Web using the traffic statistics of just three Web pages: http://reverent.org/true_art_or_fake_art.html , http://reverent.org/sounds_like_faulkner.html , http://ecclesiastes911.net/disumbrated_art.html . (Traffic patterns from a dozen other Web pages the researchers studied were very similar.) Using several years of data from these three pages, the researchers show how the Internet overall reaches a self-organized critical (SOC) state with long-lasting traffic.

“One of the main implications of our findings is that traffic and [the corresponding] fame is a prolonged phenomenon instead of a one-time fling, and recurs in a spasmodic fashion,” Roychowdhury told *PhysOrg.com*.

Most of the time, traffic to any single Web page is relatively low and steady, where visitors come from search engines, Web directories, online encyclopedias, and other constant sources. But these long periods of low traffic are interrupted by bursts of heavy traffic that follow a power law, usually the effect of numerous blog entries linking to those pages.

The researchers use a branching model to describe the probability and extent of these bursts. Basically, there’s a certain probability that a viewer will post a blog entry with a link to that Web page, and then a certain number of viewers who will visit the Web page via the blogger’s link. The product of these two variables determines whether or not a Web page will reach the critical value of 1, which determines if the branch keeps growing or dead-ends.

“A system is in a critical state if a single movement in an individual constituent element leads, on the average, to the movement of precisely one other element in the system,” Roychowdhury explained.

If a system is in a super- or sub-critical state, movement of one element leads, on average, to the movement of either more or less than one other element, respectively. That means that a signal generated in a super-critical system should increase forever, while a signal in a sub-critical system eventually dies out.

“But in a critical system, something very interesting happens,” Roychowdhury said. “Almost all signal

cascades will die out, but some of them can last for a long time and can cover a large area. Clearly, sub- and super-critical systems are not that interesting unless we want a system that is either not that responsive or a system that explodes at the slightest provocation. Critical systems, on the other hand, allow for a responsive system to exist without it being blown apart. Many physical systems naturally gravitate towards a critical state, and this phenomenon is termed SOC.”

As the researchers explain, competition for viewers and links is a driving force of the Web, and this competition pushes the entire Web into an SOC state. Based on their data, the researchers determined the values for the two variables above for the “true art or fake art” site that closely produce its traffic patterns: they found that its link probability of 0.01 and referral number of 95 visitors per link results in a slightly sub-critical value of 0.95 for that particular Web page.

But since some Web pages are more interesting than others, some pages will achieve the critical value of 1 or even surpass it.

“To explain how the Web evolves into the SOC state, we need to use the concept of Darwinian fitness, which is a scientific measure of digital fangs and claws that help the Web page to fight for links with its competitors,” Simkin said. “The success in this competition depends not only on the Web page's own fitness, but also on the average fitness of other pages currently discussed in the blogosphere, with which our Web page competes.”

If this average is low, Simkin explained, then the fittest papers are super-critical. This means that, with time, they increase their share of the blogosphere. But in turn, this leads to the increase of the average fitness. The process continues until the fittest pages become exactly critical.

“One finding that is important for Webmasters is that our work disproves the so-called fifteen minutes of fame paradigm, according to which things can get popular soon after release and quickly become forgotten,” Simkin said. “One, of course, knows that this paradigm is manifestly wrong for immortal classics. However, our work shows it to be wrong not only for great creations, but for anything which is of any intrinsic (not created by advertisement) merit.”

The researchers found that the traffic to a Web page with fixed content can persist for at least several years.

“So one should not hurry to delete old Web pages,” Simkin said. “When there is enough – say a year – of access statistics, our model can be used to infer a page's fitness and predict the average volume and fluctuations of future traffic.”

Roychowdhury is a cofounder and Simkin a consultant for a start-up company called NetSeer.com that focuses on next-generation Internet advertising. The company utilizes similar physics-based modeling of the Web, though not the direct results of the present study.

The researchers add that the Web is just one of many complex systems that exhibit self-organized criticality, with other examples including evolutionary patterns, earthquakes, and citations in research papers. They suggest that their model could also be used to explain the spreading of cultural elements, like movies, books, and fashion styles.

More information: Simkin, M. V. and Roychowdhury, V. P. “A theory of web traffic.” *Europhysics Letters*, 82 (2008) 28006.

Copyright 2008 PhysOrg.com.

All rights reserved. This material may not be published, broadcast, rewritten or redistributed in whole or part without the express written permission of PhysOrg.com.

This document is subject to copyright. Apart from any fair dealing for the purpose of private study, research, no part may be reproduced without the written permission. The content is provided for information purposes only.