Social Dilemmas and Internet Congestion

Bernardo A. Huberman and Rajan M. Lukose

Dynamics of Computation Group
Xerox Palo Alto Research Center
Palo Alto, CA 94304

Abstract

Since the Internet is a public good and its numerous users are not charged in proportion to their use, it appears rational for individuals to consume bandwidth greedily while thinking that their actions have little effect on the overall performance of the Internet. Because every individual can reason this way, the whole Internet's performance can degrade considerably, making everyone worse off. An analysis of the congestions created by such dilemmas predicts that they are intermittent in nature with definite statistical properties, leading to short lived spikes in congestion. Internet latencies were measured over a wide range of conditions and locations and were found to confirm these predictions, thus providing a possible microscopic mechanism for the observed intermittent congestions of the Internet.

The explosive growth of the Internet has led to the emergence of a computational network that covers the globe, with a staggering number of users connected at any given time to each other, browsing World Wide Web pages, sending messages, transferring files and searching for information on a range of topics that defy classification. At the time of this writing it is estimated that there are 80 million WWW pages in existence, with only half of them indexed at any given time by the best search engines [1, 2].

While this phenomenal growth has created exciting opportunities for information access, economic transformation, and interactivity on a global scale, it has also revealed that the Internet is not immune to the problems intrinsic to any public good, as congestions sometimes threaten to render it useless. Like consumers of a natural resource or drivers during rush hour, Internet users, and in particular "surfers", are faced with a social dilemma of the type exemplified by the well-known tragedy of the commons [3]. Since Internet surfers are not charged in proportion to their use, it often appears rational for them to consume bandwidth greedily while thinking that their actions have little effect on the overall performance of the network. Because every individual can reason this way, the whole Internet's performance can degrade considerably, making everyone worse off. But as users experience a congested network they reduce or even desist in their use, freeing up bandwidth that can once again be consumed, and so on.

Given this scenario and the fact that Internet surfing is completely uncoordinated, one might then expect that over time, the resulting latencies would fluctuate around some mean value determined by the capacity and protocols of the network. A study of these latencies within the context of standard models of social dilemmas, however, leads to surprising results. While most of the time the available bandwidth is not completely taken up by greedy behavior, we show that uncertainties about the state of the network lead to sudden spikes of congestion, sometimes called Internet "storms" [4], which subside when slowdowns become intolerable. Moreover, we demonstrate that these intermittent spikes in congestion possess non-Gaussian statistical properties. Experiments in which we measured Internet latencies over a wide range of conditions and locations yield data that corroborate these predictions. Thus the unfolding of a social dilemma provides a possible microscopic mechanism for the observed intermittent congestion of the Internet.

In the simplest case of a social dilemma [5–7], users try to maximize their utility, in this case proportional to the speed with which they can access remote information. They are considered cooperators if they restrain their use, or defectors if they consume bandwidth greedily. Collective benefits increase linearly in the number of cooperators, at a rate b per cooperating member [8]. In the context of the Internet, b is proportional to the speed (or inverse latency) with which remote sites can be accessed, a quantity that exhibits fluctuations over many time scales. Each contributing individual bears a personal cost, c, that reflects the penalty paid for delaying access to the information needed. Let k_i denote whether member i is cooperating ($k_i = 1$) or defecting ($k_i = 0$). Then the utility at time t for member i is given by

$$U_i = bf - ck_i \tag{1}$$

where the fraction of users cooperating, $f = n/N$, is the number cooperating divided by the size of the user community.

When all members contribute successfully, each receives net benefits $b - c$, independent of the group size. The production of the collective good represented by the bandwidth of the Internet becomes a social dilemma when $b > c > b/N$. In this regime, although the good of all is maximized when everyone restrains from overuse $(b - c > 0)$, the dominant strategy in a one-time session using the Internet is to defect since the marginal benefit of personal restraint is less than the personal cost $(b/n < c)$.

This logic changes when the interaction is ongoing since expectations about the future will also influence an individual's decision to use up bandwidth. To study the dynamics of collective action we use an equation which governs the transitions between cooperation and defection in which the average fraction of users who cooperate, f, is determined by,

$$\frac{df}{dt} = -\alpha[f - \rho(f)] \tag{2}$$

where α is the average Poisson rate at which users re-evaluate their preferences. The function $\rho(f)$ is the average probability that defecting users will switch to a cooperative strategy [9], and so contains the effect of users' expectations on their immediate decisions. It has been shown that for collective action problems in which many agents participate, this averaged mean-field equation is very accurate [10], and is therefore appropriate for a large system such as the Internet.

Thus, in this conception, users monitor their level of utility, and, if latencies are noticeable, decide to switch their behavior between effective cooperation and defection by considering the cost of postponing their current activity as well as their expectations of the future state of the network. Specifically, users who experience congestion tend to believe that others do as well, and expect that if they try again later, the network is likely to be less congested. Similarly, users who experience little congestion will tend to take advantage of it and consume as much bandwidth as they require. Thus, since users believe that other users are similar to them, they expect that if the network is heavily congested now, latencies will be lower later. Cooperation will be followed by cooperation and defection by defection on the part of other users. In addition, there is always uncertainty in the users' estimate of the current state of the Internet, which is parameterized by a Gaussian deviate variance σ. Taking account of these factors, $\rho(f) = \frac{1}{2}\left[1 + erf\left(\frac{(1-f)-f_{crit}}{\sigma\sqrt{2}}\right)\right]$, where *erf* denotes the error function, $f_{crit} = \frac{1}{H\alpha}\left(\frac{Nc-b}{b-c}\right)$, and H is the rate at which users discount future utility [11].

Eq. 2 determines the average dynamics of the system when there is a diversity of users preferences [10]. Its fixed points, obtained by setting the time derivative to zero, correspond to equilibrium situations in which most users are defecting or cooperating. Since the Internet does in fact work most of the time, we can consider its equilibrium state to be effectively one of cooperation, even though most individuals do not explicitly conceive of their behavior in those terms. However, even in equilibrium occasional fluctuations take place whereby individuals might switch over short periods of time from non-use or restrained use to heavy use, and vice versa. These switches are due to different levels of need and impatience in obtaining information, to the uncertainty that individuals have about the congestion of the network, and to random changes in the very environment of the Internet. In particular, the random environment can be thought of as a Gaussian white noise added to a mean performance benefit, so that $b = \overline{b} + \beta(t)$, where $< \beta(t) >= 0$ and $< \beta(t)\beta(t') >= 2\sigma^{'}\delta(t-t')$ with $\sigma^{'}$ such that the changes are small compared to the mean.

Each of these uncertainties can cause an individual to mistakenly perceive the total amount of Internet use to be different from its actual value. Because of this misperception, the user might act against the equilibrium condition, causing the system to move away from the fixed point. The more uncertainty there is in the system, the more likely it is that there will be fluctuations away from the equilibrium state.

When fluctuations in the levels of restraint take place, the group recovers in time according to an equation that governs the dynamics of fluctuations away from the equilibrium state. Denoting a small deviation from equilibrium by δ, the fraction of individuals restraining at any given time will be given by $f(t) = f_o - \delta(t)$ with $\delta << f_o$, where, in this case, we consider a cooperative equilibrium for which $f_o \approx 1$. Replacing this expression in Eq. 2 and using the stochastic variable $b(t)$ in the expression for f_{crit}, we can perform a linear stability analysis around the fixed point to derive an equation for the dynamics of fluctuations around the cooperative equilibrium. One obtains

$$\frac{d\delta}{dt} = [-A + B(t)]\delta \tag{3}$$

where the constant $A = \alpha\left[1 + \frac{1}{\sigma\sqrt{2\pi}}\left(1 - \frac{(1-f_0)^2}{2\sigma^2}\right) - \frac{(1-f_0)\overline{f_{crit}}}{\sigma^3\sqrt{2\pi}}\right]$ and the white noise $B(t) = -\frac{(1-f_0)}{\sigma^3\sqrt{2\pi}H(b-c)}\beta(t)$ has zero mean and standard deviation $\overline{\sigma} = \frac{(1-f_0)}{\sigma^3\sqrt{2\pi}H(b-c)}\sigma'$.

With these properties of the random function $B(t)$, Eq. 3 is a stochastic differential equation that can be analytically solved using the Stratonovich formalism [12]. The solution is the stochastic process,

$$\delta(t) = \delta(0)e^{-At+w_t} \tag{4}$$

where $\delta(0)$ is the initial value of the fluctuation and w_t is a Wiener process $B(t) = \frac{dw_t}{dt}$, such that $< w_t >= 0$ and $< w_t^2 >= \overline{\sigma}t$. This equation shows that typical fluctuations away from overall restraining relax back exponentially fast to the equilibrium point, with a time constant, $1/A$ that, as can be seen, is a complicated function of the user's evaluation rate and the network parameters.

On the other hand, the *m*th moments of δ, given by $< [\delta(t)]^m >= [\delta(0)]^m < e^{(-At+w_t)m} >$ show anomalous behavior. To see this, notice that since $< e^{mw_t} >= e^{m^2\overline{\sigma}t}$, we get

$$< [\delta(t)]^m >= [\delta(0)]^m e^{m(m\overline{\sigma}-A)t} \tag{5}$$

As this clearly shows, the higher moments can actually grow in time, with increasing order, at a rate given by $m(m\overline{\sigma} - A)$. Remarkably, the average number of defecting individuals, given by the first moment of $\delta(t)$, *increases* in time when the variance in the background speed makes the exponent in Eq. 5 positive (i.e. , $\overline{\sigma} > A$) while almost all realizations of the fluctuations relax exponentially fast to zero, as indicated by Eq. 4[1].

In order to understand this paradox, it is useful to remember that the higher moments of $\delta(t)$ are related to the probability of very unlikely events. The fact that the moments grow in time indicates that the behavior of the system is dominated by occasional large bursts in which the number of individuals suddenly found not to be restraining is very large, leading to spikes or "storms" in the level of Internet congestion.

This prediction is borne out in simulations of the full dynamics of this model of collective action with fluctuating benefits. The representative simulation shown in Fig. 1, which was performed without any averaging or linearization, corroborates these analytical results and clearly shows the spikes in the number of defectors.

In order to understand the nature of these sudden bursts, we need to compute their quasi-stationary distribution from their evolution over relatively long time periods. Dividing both sides of Eq. 3 by $\delta(t)$ and integrating over a time interval T gives, up to a constant, $\ln \frac{\delta(t)}{\delta(0)} = \sum_0^T B_n$, where the integral of $B(t)$ over time has been replaced by a sum over discrete time steps, at each of which the random integrand acquires the value B_n.

[1] This paradoxical behavior is reminiscent of the analysis of ecological extinctions in random environments [13].

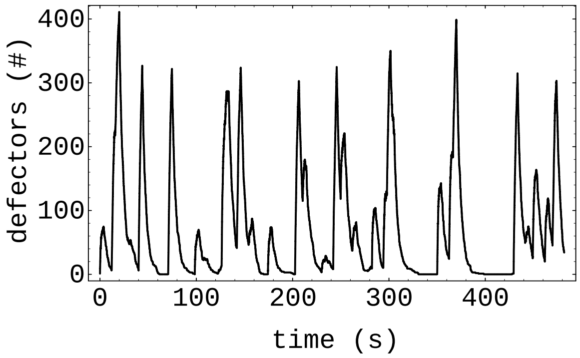

Fig. 1. The number of defectors as a function time for the following parameter values: $N = 510$, $\bar{b} = 480$, $\sigma' = 20$, $c = 1$, $p = 0.6$, $H = 1$, and $\alpha = 0.25$. The benefits had Gaussian fluctuations on a time scale longer than the average individual re-evaluation rate α.

Since the values of B_n are normally distributed with zero mean and variance $\overline{\sigma}^2$, the sum of the right hand side can be evaluated by invoking the Central Limit Theorem. Therefore, the logarithm of $\delta(t)$ is normally distributed with zero mean and variance $\overline{\sigma}^2$. In other words, $\delta(t)$ is distributed lognormally, i.e. according to a distribution whose density function is $P(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x)^2}{2\sigma^2}\right]$ with the mean value of x given by $\mu = e^{\sigma^2}$, and its variance by $\mu^2\left(e^{\sigma^2} - 1\right)$.

To verify the existence of the predicted congestion spikes in the Internet we performed a series of experiments in which we measured round-trip times for Internet Control Message Protocol (ICMP) ping packets. The time it takes for the ping packet to make its round-trip is a measure of the congestion of the network. Typically, these packets were forwarded through about 30 routers, which act as sources of speed fluctuations, thus justifying the assumption of delta-correlated noise in b. Over the course of the measurements, use of the *traceroute* command showed that the pings typically followed the same sequence of routers.

Our measurements of round-trip times for ping packets sent every minute over the course of a day demonstrated that there are daily periodicities in the level of congestion around which higher frequency fluctuations occur. Since these daily periodicities are an exogenous source of congestion, we studied instead the latencies of pings sent at a higher frequency over shorter periods of time. During these experiments the stochastic dynamics of the congestion could then be considered as quasi-stationary.

Fig. 2 shows a time series of nearly $10,000$ pings sent from a public workstation at Stanford University in California to a WWW server in the United Kingdom over the course of about 45 minutes. Since the trans-Atlantic path taken by the packets is one of the most congested in the world, and since the time difference between the US and the UK is large, this data provides a test of the theory which is not complicated by too much

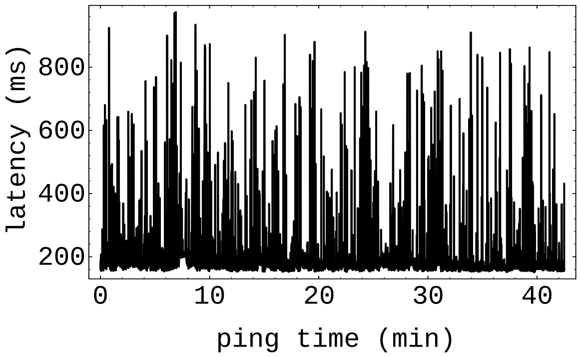

Fig. 2. Time series of round-trip times for 56 byte ICMP ping packets sent from **elaine.leland.stanford.edu** to **www.cranfield.ac.uk**. The pings took place starting at about 7:00 P.M. PST on January 28, 1997 and took about 45 minutes to complete. Of the $10,000$ packets sent, 179 of them timed-out and were removed from the series. The mean latency of the sample series is $189ms$, and each ping was sent $25ms$ after the return of the previous one. Thus the data is a high frequency measure of the congestion.

variation in the background due to exogenous factors. The series clearly exhibits latency spikes, or storms, superimposed on a noisy background. The spikes are bounded in size due to the fact that the Internet routers drop packets when not transmitted over long times. This kind of behavior was observed in many other such experiments, involving measured latencies between remote sites in the U. S. as well as within the San Francisco Bay Area.

Since the time series of pings was measured at unequally-spaced intervals, their autocorrelation function was calculated by first calculating the Lomb periodogram power spectrum and then taking the inverse Fourier transform [14]. The autocorrelation function for the pings, which we calculated by taking the inverse Fourier transform of the Lomb periodogram for the observed $9,821$ latencies, showed that the ping latencies are uncorrelated in time. This supports the assumption of stationarity, since periodicity in the congestion over the course of the measurements would be revealed as structure in this function.

Fig. 3 shows the histogram of ping times along with a superimposed lognormal distribution with the same mean and variance as the data. As can be seen, a lognormal distribution appears to be a good fit, although the statistical significance is not high. This is to be expected since the network's built-in bounds on extreme latencies prevent the realization of any long tail in the distribution, making a quantitative statistical comparison difficult.

These results support the interpretation of large congestion peaks or storms as due to the aggregate behavior of individuals confronted with a social dilemma. Given their generality they may also hold for urban and highway traffic. Concerning the statistics of the congestions, they show that care must be exercised when speaking about the average

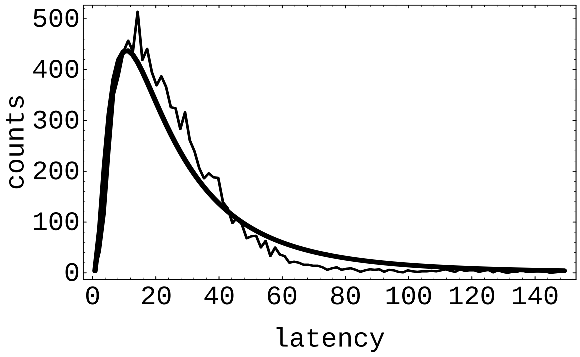

Fig. 3. Histogram of sampled ping latencies (thin line) and a lognormal distribution with the sample mean and variance (thick line). The baseline for the ping latencies has been subtracted, so that real latency values in ms can be obtained by adding the constant value 150.

performance of the Internet, for since the distribution of latencies has a large skewness, the mode, which corresponds to typical behavior, does not equal the mean. Thus data such as the Internet Weather Report provided by Matrix Information and Directory Services (MIDS) [4], which is averaged over thousands of sites, can be misleading in terms of the large congestions they imply. Rather, a typical user of the Internet would encounter latencies shorter than those exhibited in the MIDS data.

The existence of a well defined statistical distribution of Internet latencies can also help in the design of restart algorithms for accessing remote sites of the World Wide Web during times of high congestion. We showed earlier that the introduction of the notion of risk into the execution of computer programs leads to the construction of computational portfolios that can be used to increase the speed and reliability with which Web pages can be accessed [15]. Thus, knowledge of the underlying distribution allows for speedy construction of such portfolios without having to sample data over very long times.

Another aspect of our experiments worth mentioning is that the measured autocorrelations do not exhibit the long time tails that have been found in studies of Ethernet packet flow rates [16]. This difference can be attributed to the fact that those autocorrelations seem to correspond to background levels of traffic, which have different statistical properties than the higher frequency congestion related spikes that we studied.

Finally, we point out that congestion problems as presently experienced by users of the Internet would disappear if individuals were charged in proportion to their bandwidth consumption. Such schemes, well known in the context of highway and urban traffic [17, 18], are also an active area of research in the Internet community [19–21, 2] and hold the promise of relieving congestion. In the meantime, the Internet in its present state offers a unique environment for studying the social dynamics underlying network externalities [22].

References

[1] Special issue: The internet. *Scientific American*, 276(3), March 1997.

[2] Ellen Germain. Fast lanes on the internet. *Science*, 273:585–588, 1996. Computers '96: News.

[3] Garrett Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

[4] Internet Weather Report, http://www.mids.org/.

[5] Mancur Olson. *The Logic of Collective Action*. Harvard University Press, Cambridge, 1965.

[6] Thomas C. Schelling. *Micromotives and Macrobehavior*. W. W. Norton and Company, Inc., United States, 1978.

[7] Natalie S. Glance and Bernardo A. Huberman. Dynamics of social dilemmas. *Scientific American*, pages 76–81, March 1994.

[8] Jonathan Bendor and Dilip Mookherjee. Institutional stucture and the logic of ongoing collective action. *American Political Science Review*, 81(1):129–154, 1987.

[9] Bernardo A. Huberman and Tad Hogg. The behavior of computational ecologies. In B. A. Huberman, editor, *The Ecology of Computation*, pages 77–115. North Holland, Amsterdam, 1988.

[10] Bernardo A. Huberman and Natalie S. Glance. Diversity and collective action. In H. Haken and A. Mikhailov, editors, *Interdisciplinary Approaches to Complex Nonlinear Phenomena*, pages 44–64. Springer, New York, 1993.

[11] Bernardo A. Huberman and Natalie S. Glance. Beliefs and cooperation. In P. Danielson, editor, *Modelling Rational and Moral Agents*, Cognitive Science Series. Oxford University Press, 1997.

[12] R. L. Stratonovich. *Topics in the Theory of Random Noise*. Gordon and Breach, 1967.

[13] Richard C. Lewontin and D. Cohen. On population growth in a randomly varying environment. *Proceedings of the National Academy of Sciences, USA*, 62:1056–1060, 1969.

[14] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes*. Cambridge Univ. Press, Cambridge, 1986.

[15] Bernardo A. Huberman, Rajan M. Lukose, and Tad Hogg. An economics approach to hard computational problems. *Science*, 275:51–54, 1997.

[16] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic. In *ACM/SIGCOMM Computer Communications Review*, volume 23, pages 183–193. ACM/SIGCOMM, 1993.

[17] William Vickrey. Congestion theory and transport investment. *American Economic Review*, 59:251–260, 1969.

[18] Kenneth A. Small. *Urban Transportation Economics*. Hardwood Academic Publishers, 1992.

[19] Ron Cocchi, Scott Shenker, Deborah Estrin, and Lixia Zhang. Pricing in computer networks: Motivation, formulation, and example. *IEEE Trans. on Networking*, 1(6):614–627, 1993.

[20] D. O. Stahl A. Gupta and A. B. Whinston. Pricing of services in the internet. Technical report, University of Texas at Austin, 1995.

[21] Jeffrey K. Mackie-Mason Hal R. Varian and Scott Shenker. Network architecture and content provision: An economic analysis. *Telecommunications Policy*, 20:203–217, 1996.

[22] Part of this research was supported by NSF Grant IRI-9615111. R. Lukose acknowledges the support of a Citibank fellowship at Stanford University.