A Methodology for Managing Risk in Electronic Transactions over the Internet

Rajan M. Lukose and Bernardo A. Huberman

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

Abstract

The explosive growth of the Internet has made possible an economic transformation in which commerce will become largely electronic. In this vision, the medium of communication and negotiation becomes a distributed network of computers equipped with robust, secure and trusted "cash" mechanisms. It is thus of paramount importance to design mechanisms that will ensure timely and reliable transactions in cyberspace. In this paper we present a methodology for quantitatively managing the risk and cost of executing transactions in a distributed network environment, and illustrate it with several concrete examples under different pricing schemes.

To be presented at the Third International Conference on
Computational Economics, Stanford University. June 30, 1997.

1. Introduction

The explosive growth of the Internet has made possible an economic transformation in which commerce will become largely electronic. In this vision, the medium of communication and negotiation becomes a distributed network of computers equipped with robust, secure and trusted "cash" mechanisms. Given the far-reaching implications of the shift toward electronic commerce, there is at present an enormous amount of interest in providing such a mechanism for the Internet in its present state. However, since the Internet is plagued by congestion problems, there is also a great deal of interest in finding ways of relieving congestion in order to preserve the Internet as a reliable and efficient infrastructure [1].

Most of the current proposals focus on developing some form of usage-based pricing in order to manage congestion, and so it is likely, if not inevitable, that in the future the delivery channels for electronic commercial transactions will be a network in which users will be charged for transmitting the information necessary to complete those transactions. Such costs would be analogous to the cost of postage when sending a bill by ordinary mail, or to the cost of maintaining Automatic Teller Machines in order to deliver cash to bank depositors, or to highway tolls, etc. They are part of the cost of doing business.

For a high transaction volume institution like a bank, such new and potentially comprehensive delivery channels as the Internet raise important issues of how to manage efficiently the costs and any risks associated with them. Just as banks have tried to encourage the use of Automatic Teller Machines in order to cuts costs and set up contingency plans in case unforeseen problems affect the physical delivery of important materials such as checks, so will banks have to consider various options in order to manage the cost and risk involved in completing electronic transactions. However, assuming the existence of a comprehensive electronic delivery channel, the range of possible risks and costs that need attention is considerably reduced, since issues such as the mechanical wear in Automatic Teller Machines are not relevant. In fact, we assume that the time it takes to complete a transaction will be the parameter that needs to be optimized, and so we can identify that time with the cost of a transaction (which would in turn determine the monetary cost of executing the transaction in a usage-priced network). The variance in the transaction time can be considered to be the risk associated with the transaction, which would be an especially important quantity for a high-volume institution such as a bank that has to coordinate many different transactions which may have interdependencies.

In what follows, we present a methodology for quantitatively managing the risk and cost of executing transactions in a distributed network environment. By associating the cost with the time it takes to complete the transaction and the risk with the variance in that time, we exploit an analogy with the modern theory of financial portfolio management first suggested, in a more general context, in a previous paper [2]. In modern portfolio theory [3, 4], the fact that investors are risk-averse means that they may prefer to hold assets from which they expect a lower return if they are compensated for the lower return with a lower level of risk exposure. Furthermore, it is a non-trivial result of portfolio theory that simple diversification can yield portfolios of assets which have higher expected return as well as lower risk. Similarly, in this case we consider different methods of executing transactions that are analogous to asset diversification, and which yield mixed strategies that allow an efficient trade-off between the average time a transaction will take, and the variance or risk in that time. Just as in the case of financial portfolios, we find that some of these mixed strategies can execute transactions faster on average and with a smaller variance in their speed.

In Section 2, we present a detailed description of the risk-cost framework, and describe how it would be implemented in the Internet as it exists today without any pricing, and in Sections 3 and 4 we consider situations for a network equipped with different kinds of pricing schemes. Section 5 is a general discussion of the risks which stem from congestion in the Internet using empirical measurements of Internet "storms." In Section 6 we considers implementation issues in general, and we conclude with a discussion of directions for further research.

2. Strategies Without Pricing

A transaction consists of a round-trip message sent from one computer to another. The process is considered to be wholly automated so that when a message is received by one computer, it is assumed to be processed in a fixed amount of time. After processing, an acknowledgment is sent back to the computer which originated the message.

In a large congested network, the time it takes to send and receive a message is a random variable. In this section we assume that the probability density function $p(t)$ for this time is known. Thus, the expected duration for the transaction is simply

$$< t >= \int_0^\infty tp(t)dt \tag{1}$$

and the risk is

$$\sigma = \sqrt{Var[t]} = \sqrt{< (t- < t >)^2 >} \tag{2}$$

However, consider a situation in which a message has been sent and no acknowledgment has been received for some time τ. This time can be very long in cases where

$p(t)$ has a long tail, and one is faced with the choice of either continuing to wait for the acknowledgment or sending out another message or, if the network protocols allow, to cancel the original message before sending out another. The target computer can always be programmed to ignore duplicate messages so that sending another need not cause any confusion. For simplicity, we consider the case where it is possible to cancel the original message before sending out another at the time τ. This case is analogous to the familiar situation of using a World Wide Web browser such as Netscape, where requests for files which are taking a long time to download can be restarted, often to advantage. For such a "restart" strategy, in terms of the density $p(t)$, an expected duration and risk for a transaction can be calculated for each possible value of τ. The formal results are given in Appendix A.

As $\tau \to \infty$ we recover the situation in which no restart strategy is used at all. As τ decreases, different expected times and risks result, and the dependence can be shown on a plot of expected time versus risk. For some distributions $p(t)$, a restart strategy can make accessible points in the expected time versus risk plane which are unequivocally preferable to what would result without the policy. Such a situation is shown in Fig. 1 where the distribution is a lognormal. The probability density for a lognormally distributed random variable is $p(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{-(\log t-\mu)^2}{2\sigma^2}}$, which has a long tail compared with a normal distribution of the same mean and variance. We discuss in Section 5 the significance of the lognormal in the context of Internet traffic, but it is here used to illustrate the risk-expected time construction.

The cusp of the curve Fig. 1, corresponding to $\tau \approx 12.5$, yields significant speed and risk improvements. In fact, the cusp represents the strategy which is preferable to *all* other points, since none exist in this case which have lower expected time (at, possibly, the cost of higher risk) or lower risk (at, possibly, the cost of higher expected time).

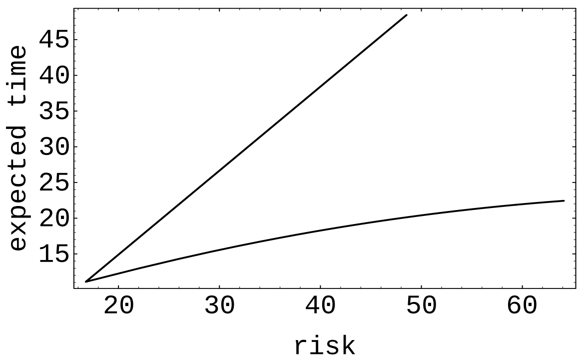

Fig. 1. Expected transaction completion time versus risk for a restart strategy. The lower right endpoint of the curve corresponds to no restart strategy ($\tau \to \infty$) and the upper left corresponds to very short restart times. The parameters for the lognormal used in constructing the curve are $\mu = 2$ and $\sigma = 1.5$

Of course, the restart strategy is not the only possible one. If cancelling a message is not possible under the network protocol, then one could simply send another message after waiting a time τ, as was mentioned above. Indeed, an even better strategy would be to send many messages to begin with. The resulting expected time and risk, parametrized by N, the number of messages sent in the latter case, is also given in Appendix A, to illustrate how such calculations can be done.

Clearly, in a network without any kind of usage-based pricing, sending many transaction messages to begin with would be the best strategy as long as N is not so large as to overwhelm the target computer. On the other hand, others using the network can reason in exactly the same way, resulting in congestion that would render the network useless. This paradoxical situation, sometimes called a social dilemma, arises often in the consideration of "public goods" such as natural resources, highways, agricultural community grazing lands, to name a few, and have been studied extensively [5–7]. The paradox disappears if users are charged in proportion to their use [8]. This is why, of course, there is so much current interest in determining the details of an appropriate pricing scheme for the Internet, since users currently do consume Internet bandwidth greedily when they download large multimedia files, for example, without consideration of the congestion caused by such activity.

3. Strategies With Simple Pricing

Consider the simplest pricing scheme in which users are charged in proportion to the amount of data they send. Such a system is currently used in New Zealand [9]. Restart strategies, parametrized by τ, would now have an expected monetary cost associated with them. Specifically, for any τ, an expected number of restarts can be calculated, which when multiplied by the fixed cost of each message sent (since the messages are of fixed size and are just copies of one another), gives the expected monetary cost for each τ strategy. In general, the effect of this kind of pricing scheme is to add a third dimension, in addition to expected time and risk, along which a strategy must be evaluated. In other words the relevant curve, containing all the information necessary to choose a τ, is a curve in the three dimensions expected time, risk, and expected monetary cost.

Fig. 2 shows the three-dimensional extension of Fig. 1, assuming that each restart has unit monetary cost. For clarity the curve is drawn as a surface, but only points on the upper edge of the curved "bannister" are accessible strategies. The lower edge, lying in the risk-expected time plane, is exactly Fig. 1. Thus, a particular risk-expected time strategy (corresponding to a particular τ) has an expected cost equal to its expected number of restarts, which is simply the hight of the surface in Fig. 2.

The crucial point, when pricing is taken into account, is that with the addition of the extra expected monetary cost dimension, the cusp in Fig. 1, is no longer necessarily

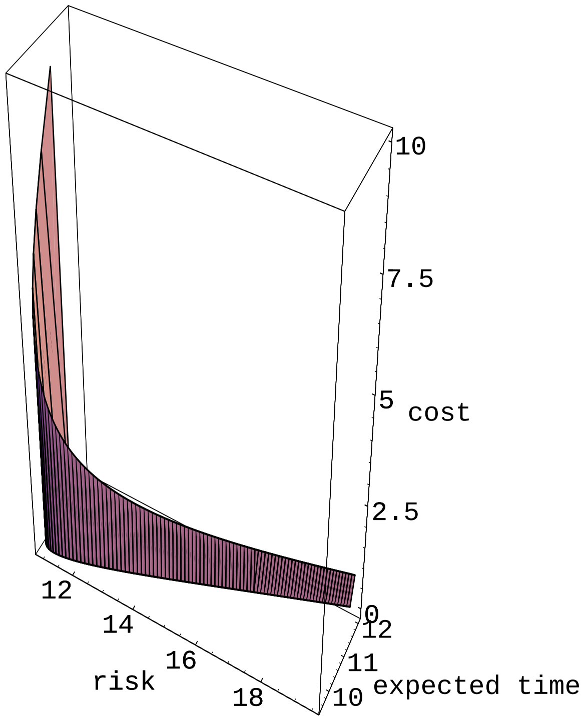

Fig. 2. A plot of expected monetary cost for various restart strategies. The upper edge of the surface gives the actual expected cost. The lower edge is the curve Fig. 1. Here, however, we show only the strategies near the cusp, for which $.3 < \tau < 15$.

optimal. Although it is the point with the lowest expected transaction time and the lowest risk, it is also now more expensive than using no restart strategy at all. It would be up to the user to decide then if the faster execution and lower risk is worth the extra expense.

Similarly, for the strategy of sending N messages to begin with, the cost would be a simple linear function of N, and a different three-dimensional curve could be calculated for which the same argument applies, i.e. even if sending 100 messages yields a very fast expected execution, the monetary cost would be 100 times higher than sending only one message. There is thus a relative disincentive to flood the network with transaction messages. The monetary cost must also be weighed, and this quantitative framework allows the sender the freedom to choose among a range of possibilities all parameterized either by τ or N, possibilities which are inaccessible without strategies like restarting.

4. Strategies With Priority Pricing

The previous section dealt with situations in which it was necessary to introduce a third dimension in order to evaluate strategies. However, for some kinds of pricing in which it is possible to attach a priority level to sent data [1], it may be possible to gain improvements in expected execution time and/or risk for a fixed budget, simply by sending, for example, two copies of the message at two different priorities, in such a way that the total cost of such a mixed strategy would still be equal to the cost of sending one copy at the maximum priority. In this case, the third monetary cost dimension is unnecessary, since all such strategies cost the same.

Specifically, consider a situation in which the probability density for the time it takes to send a message of fixed size at the highest priority is known to be $p(t)$. Suppose that it is possible to send data at some lower priority denoted by $0 < f < 1$, for which the probability density for a message sent at that priority is $p(ft)$. In order to get some intuitive sense of what this means, consider the situation in which

$$p(t) = \frac{1}{2}\delta(t-1) + \frac{1}{2}\delta(t-10) \tag{3}$$

i.e. the probability that the transaction time takes *1 s* is equal to the probability that it takes *10 s*. For a priority level $f = \frac{1}{10}$, the probability density would become

$$p(t) = \frac{1}{2}\delta(t-10) + \frac{1}{2}\delta(t-100) \tag{4}$$

so that the shortest completion time is *10 s* which is equally likely as a completion time of *100 s*. Thus, at a lower priority, the transaction has a higher expected completion time as well as a higher risk.

Assuming that the monetary cost of sending a message (for messages of the same size) is proportional to the priority f, let two copies of the same message be sent at different priorities such that the sum of the priorities equals 1. Letting, for simplicity, the constant of proportionality between monetary cost and priority equal unity, the monetary cost for this mixed strategy is equal to the cost of sending one copy at the highest priority (which is itself equal to 1 in arbitrary cash units). The duration of the transaction is then given by the random variable,

$$t = \min\left[\frac{t_1}{f}, \frac{t_2}{1-f}\right] \tag{5}$$

where $t_{1,2}$ are independent random variables each with the distribution given in Eq. 3. By allowing the priority f to vary, the expected time versus risk curve shown in Fig. 3 can be calculated, following the procedure outlined in Appendix B. Each point on the curve corresponds to a particular value of f, in the same way that the curve in Fig. 1 is parameterized by τ.

The curve has a remarkable feature, the "efficient frontier", indicated by the solid segment , which is defined by the fact that for every point not on the efficient frontier,

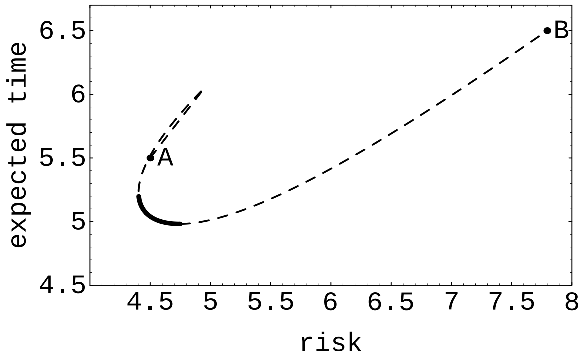

Fig. 3. Expected transaction completion time versus risk. Each point on the curve represents the risk and performance of the mixed priority strategy when the priority of one message is f and the other is $1 - f$ with the priority ranging from 0 to 1. The solid segment corresponds to the efficient frontier. The point labelled A corresponds to the expected completion time and risk for sending one message at maximum priority and the point labelled B corresponds to sending two messages with equal priority.

there is at least one point on the efficient frontier that is *always preferable* , i.e., has a lower risk as well as faster expected completion time. Note that the endpoint of the curve labelled *A*, which corresponds to sending the message at the highest priority, is not on the efficient frontier and would therefore never be preferred. This result is analogous to the result in modern portfolio theory which says that simple diversification can increase the expected return and decrease the risk of a portfolio beyond the levels of the highest returning asset or the least risky asset in the portfolio. Thus, under such a priority pricing scheme, mixed strategies constructed according to the risk management methods presented here can lead to much better strategies for executing electronic transactions.

5. Internet Storms

In the previous sections we assumed that round-trip times in networks have a well-defined probability distribution which can then be used to design particular efficient mixed-strategies. In what follows we will show that this is indeed the case for the Internet in its present pattern of use.

We have measured round-trip times for Internet Control Message Protocol (ICMP) ping packets [10], which are almost exactly analogous to the fixed-size transaction messages discussed here over a number of days and geographical locations. Fig. 4 shows an example 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.

Fig. 5 shows a histogram of the ping times with a superimposed lognormal distribution of the same mean and variance as the data depicted in the time series. The lognormal distribution appears to be a good fit. It reflects the dynamics of millions of

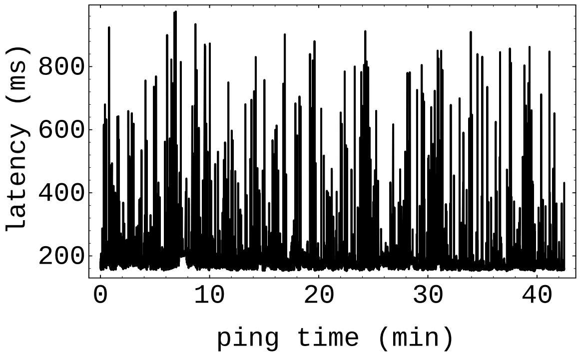

Fig. 4. Time series of round-trip times for 56 byte ICMP ping packets sent from **elaine.leland.stanford.edu** to **www.cranfield.ac.uk**. The pings took place starting at about 7:00 P.M. PST on January 28, 1997 and took about 45 minutes to complete. Of the $10,000$ packets sent, 179 of them timed-out 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.

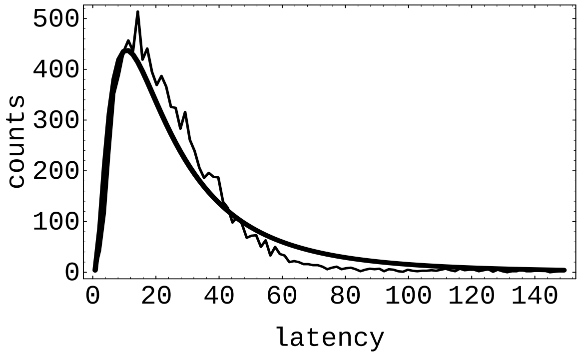

Fig. 5. Histogram of sampled ping latencies (thin line) and a lognormal distribution with the sample mean and variance (thick line). The baseline for the ping latencies has been subtracted, so that real latency values in ms can be obtained by adding the constant value150.

users confronted with the dilemma of either consuming bandwidth greedily and thus creating congestion, or postponing access to remote information and paying the corresponding cost. These empirical findings and the analytical theory underlying them, discussed in detail in Ref. [10], therefore justifies the use of a lognormal distribution in the examples used in the restart strategy above.

6. Implementation Issues

There are several important implementation issues worth mentioning. First, in the preceding sections we assumed that all the relevant probability densities were known beforehand. In reality however, the probabilities will have to be estimated dynamically on-line following methods that are presented elsewhere [11].

Second, we note that there is evidence that the lognormal distribution used in Sections 2 and 3 may be very general, as it arises from the fundamental dynamics of individuals using what amounts to a public good [10]. This means that any online method to determine the probability distribution can make use this information to establish a portfolio in a efficient and prompt fashion. In addition, we point out that there is a wide range of distributions which give similar results to those presented above [2].

Third, while the pricing scheme presented in Section 3 is realistic and is actually currently in use , the priority pricing scheme used in Section 4, and the assumptions underlying it, are untested. Even so, it contains the basic elements of more comprehensive priority pricing systems and the added feature of intuitive appeal.

7. Conclusion

The Internet is rapidly becoming the medium through which a large share of communications and commerce takes place. While it is still too early to tell which elements of banking will be dominated by the existence of such an ubiquitous medium, several scenarios are worth mentioning. The first one is the impending appearance of stock and commodities transactions over the Internet, performed by users with or without the intervention of brokers and intermediaries. It is obvious that prompt receipt of buy and sell orders is imperative in such an environment, for otherwise transaction delays could lead to significant losses. Thus acceptance of such a system will hinge on having good bandwidth and tolerable latencies. Another one is the appearance of private e-mail delivery systems that will guarantee, in exchange for a fee, delivery of important data sets and mail. Here, congestions that can be circumvented can give a strategic advantage to any provider that can have access to the technology.

In this paper we presented a methodology for quantitatively managing the risk and cost of executing transactions in a distributed network environment, and illustrated it with several concrete examples under different pricing schemes. We showed that it is indeed possible to construct specific strategies that can deal with the fluctuating congestions and ensuing delays of distributed networks. This methodology is thus directly applicable to a number of problems relevant to banking and its implementation does not require additional investments in hardware. Rather, software controls of the type that we discussed can speed up transactions while increasing their reliability.

As is always the case with the appearance of a new method, its advantage relies on having only a small share of the actors in the economy using it. However, at present there is no data that allows one to show the deterioration of any such advantages as more and more agents adopt this methodology. Thus, it might take a long time for the market to become efficient, thus making this methodology very useful.

Appendix A

Restart Strategy

Given the probability density $p(t)$ for the completion time, a choice of restart time τ yields a new probability density $P(t)$ for that strategy. It is given by the probability that the task does not complete in exactly $floor\left[\frac{t}{\tau}\right]$ restarts before finally completing. Here the $floor[]$ function simply gives the nearest integer less than its argument. Thus,

$$P(t) = p\left(t - floor\left[\frac{t}{\tau}\right]\right)\left[1 - \int_0^\tau dt' p\left(t'\right)\right]^{floor\left[\frac{t}{\tau}\right]} \tag{6}$$

For a density of this form, the expectation of any function can be written as a series,

$$< f(t) >= \sum_{i=0}^{\infty}\left[1 - \int_0^\tau dt' p\left(t'\right)\right]^i \int_0^\tau dt'' f\left(t'' + i\tau\right) p\left(t''\right) \tag{7}$$

If, for ease of notation, we define $a \equiv 1 - \int_0^\tau dtp(t)$, $b \equiv \int_0^\tau dtp(t)$, $c \equiv \int_0^\tau dttp(t)$, and $d \equiv \int_0^\tau dtt^2 p(t)$, then the first and second moments for the restart strategy can be calculated by summing series, gotten from Eq. 7, which turn out to be derivatives of geometric series. The result is

$$< t >= \frac{1}{b}(c + \tau a) \tag{8}$$

for the expected time, and using the result

$$< t^2 >= \frac{1}{b}\left[d + \tau a\left(2\frac{c}{b} + \tau\left(1 + \frac{2}{b}\right)\right)\right] \tag{9}$$

in the definition of risk in Eq. 2 gives the risk.

Multiple Send Strategy

If we send out N messages independently and each have the probability density $p(t)$ for its completion time, then the probability that all of them will take longer than some time t is given by

$$G(t) = Prob[> t] = \left(\int_t^\infty dt' p\left(t'\right) \right)^N \tag{10}$$

Then, the probability density $P(t)$ for this multiple send strategy is simply the negative of the time derivative of the cumulative distribution function $G(t)$ so that

$$P(t) = -\frac{dG(t)}{dt} = N \left(\int_t^\infty dt' p\left(t'\right) \right)^{N-1} p(t) \tag{11}$$

from which the risk and expected time can be calculated.

Appendix B

What follows is a calculation similar to the calculation for the multiple send strategy given in Appendix A. Here, we find the probability density function for the time to completion when that time is given by

$$t = \min\left[\frac{t_1}{f_1}, \frac{t_2}{f_2}\right] \tag{12}$$

the generalization of Eq. 5 for any constants f_1 and f_2.

First, consider two random variables t_1 and t_2 which have continuous probability density functions $p_1(t)$ and $p_2(t)$. We wish to know the probability density function $P(t)$ for the random variable defined by $t = \min[t_1, t_2]$. We can define the distribution function $G(t)$ as the probability of the random variable being greater than some t. It is simply the probability that both t_1 and t_2 are greater than t

$$G(t) = Prob[> t] = \left(\int_t^\infty dt' p_1\left(t'\right) \right) \left(\int_t^\infty dt' p_2\left(t'\right) \right) \tag{13}$$

Then, the probability density function $P(t)$ is given by:

$$P(t) = -\frac{dG}{dt} = p_1(t) \int_t^\infty dt' p_2\left(t'\right) + p_2(t) \int_t^\infty dt' p_1\left(t'\right) \tag{14}$$

We need one more general result. If $p(t)$ is the density function for the random variable t, what is the density function $p^{'}(t)$ for the random variable defined by $t^{'} = t/k$ where k is just a number? This is a basic result, and the probability density function is simply

$$p^{'}(t) = kp(kt) \tag{15}$$

Use of the two results given above leads finally to

$$P(t) = f_1 f_2 \left\{ p_1(f_1 t) \int_t^{\infty} dt^{'} p_2\left(f_2 t^{'}\right) + p_2(f_2 t) \int_t^{\infty} dt^{'} p_1\left(f_1 t^{'}\right) \right\} \tag{16}$$

the probability density for the random variable defined in Eq. 12, where $p_1(t)$ and $p_2(t)$ are the probability densities for the random variables t_1 and t_2 (for the priority pricing example in Section 4, $p_1(t) = p_2(t)$). Using these results, the means and variances of the distribution of completion times can be calculated, and the risk-expected time curve can be produced for two arbitrary algorithms combined in this way.

References

[1] http://www.sims.berkeley.edu/resources/infoecon/Pricing.html.

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

[3] Harry Markowitz. Portfolio selection. *Journal of Finance*, pages 77–91, March 1952.

[4] William F. Sharpe. *Portfolio Theory and Capital Markets*. McGraw-Hill Inc., New York, 1987.

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

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

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

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

[9] http://www2.waikato.ac.nz:81/its/charges.html.

[10] Bernardo A. Huberman and Rajan M. Lukose. Social dilemmas and internet congestion. *Science*, 1997. In press.

[11] George Roumeliotis and Bernardo A. Huberman. In preparation.

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