Probability, Computer Networks, and Simulation

Maurice F. Aburdene, Thomas J. Goodman
Electrical Engineering Department
Bucknell University

Abstract

We present one project used in our random signals and noise course focusing on the applications of probability to the area of computer networks. The project requires students to apply their knowledge of probability that includes applications in electrical and computer engineering. In addition, students analyze the performance of a computer network, simulate a system, and look into some design issues.

Introduction

ABET evaluation criteria for electrical engineering programs state "The program must demonstrate that graduates have: knowledge of probability and statistics, including applications appropriate to the program name and objectives; and knowledge of mathematics through differential and integral calculus, basic sciences, computer science, and engineering sciences necessary to analyze and design complex electrical and electronic devices, software, and systems containing hardware and software components, as appropriate to program objectives."(See http://www.abet.org/criteria.html) We will present an electrical engineering project in our random signals and noise course (required for all electrical engineering students) to emphasize the application of probability and statistics[1-7] in computer communication networks area. This project is the second of five course projects in a random signals and noise course[8].

Computer communication networks are ubiquitous and have many configurations including local area networks (LANs), wireless networks, satellite networks, and Internet[9-15].We will consider network models shown in Figure 1. The probability that a packet is damaged on a computer link is p. We will consider each of the network models and analyze the performance of the network based on the values of p. Specifically, we are interested in the probability of packet losses in the network and the expected number of packet transmissions for a large number of packets.

We begin with an e-mail that is broken into K packets and then transmitted over a computer link. The probability of losing a packet is p. If a packet is damaged, it is retransmitted. Initially, we address the factors that determine p, the average number of packet transmissions for a successful packet transmission, and plots of the average number of transmissions vs. p. Here we also discuss the Bernoulli and geometric distributions, and their means and variances. In addition, we address the probability that a packet is transmitted successfully in at most two tries and move on to N tries, and then address the design factors that influence the choice of N. Students write a simulation program to verify the analytical results. We then raise the issue of the number of simulation runs used to estimate the average number of transmissions required for successful transmissions of packets and show histograms. Next we move on to the entire e-mail and address the average number of packets sent for a successful e-mail over serial and parallel links addressing reliability issues. Project quiz questions extend the project and introduce the binomial distribution (exactly m successful transmissions in *K* transmissions) and the Pascal distributions (the probability that m packets will be transmitted correctly on the K^{th} transmission).

Single-Link Network

Consider an e-mail message being transmitted across a single network link. The e-mail is broken up into *K* packets, each of which is transmitted individually across the link. The probability that a transmission fails is *p*; in the case of a failed transmission, the packet is retransmitted. The probability *p* may depend on many factors, including the type of link such as wireless or optical), the number of users sharing the link, the physical distance traveled across the link (longer distances may be more prone to data losses), environmental conditions, and the error correction techniques. If there are many users sharing a network, then several users might compete for network resources and initial transmission attempts might fail due to packet collisions.

We introduce the idea of Bernoulli distribution of a random variable at this point by looking at a single transmission and we note that the outcome of a single transmission (the random variable X) is either a success or a failure. We discuss the probability mass function and the cumulative distribution function. We also present the mean and variance of X. We move on to the geometric distribution and its probability mass function and cumulative distribution function.

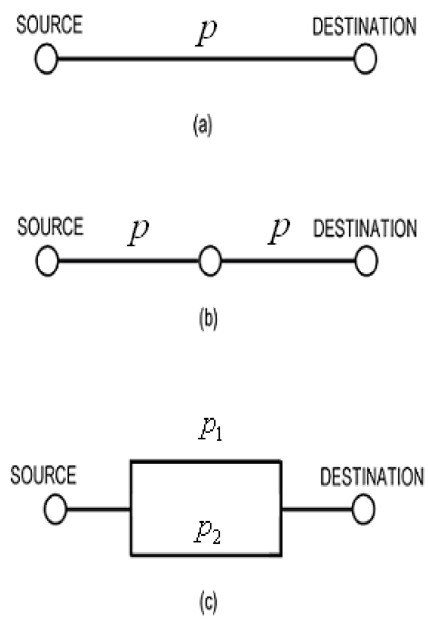

Figure 1. Basic network structures: (a) single link; (b) series link; (c) parallel link

The probability that the packet is successfully sent after *n* transmissions is given by

$$p^{n-1}(1-p) \qquad (1)$$

(the transmission must fail $n-1$ times, each with probability *p*, and then succeed on the n^{th} try).

Here we define the random variable Y to be the number of transmissions for a successful transmission of a packet. We note that Y is geometrically distributed random variable and plot its probability mass function and its cumulative distribution function.

The average number of transmissions E(Y) for a successful transmission of a packet is

$$E(Y) = \sum_{i=0}^{\infty} ip^{i-1}(1-p) = 1/(1-p). \qquad (2)$$

A quick check of boundary conditions shows that if $p = 0$ (all transmission succeed), then the expected number of transmissions is 1, while if $p = 1$ (all transmissions fail), infinitely many transmissions are required. Figure 2 shows a plot of the expected number of transmissions vs. *p*.

Ideally, packets are transmitted in as few transmissions as possible. Suppose for instance that we would like the packet to travel across the link after no more than two tries. The probability of this occurring is

$$q_{1,2} = (1-p) + p(1-p) = 1 - p^2 \qquad (3)$$

(the sum of the probabilities that the packet will arrive after exactly one or exactly two attempts or 1 minus the probability of two failures in two tries). For instance, if $p = 0$ and all attempts succeed, then only one transmission is required. However, if $p = 1$ (all attempts fail), the probability is zero. Figure 3 shows a plot of the probability of success in at most two tries vs. *p*.

Generally speaking, the probability that a packet is successfully transmitted across one network link in at most *N* tries is

$$q_{1,N} = \sum_{i=1}^{N} p^{i-1}(1-p) = 1 - p^N. \qquad (4)$$

Therefore, for values of *p* less than one (a non-zero probability of a single transmission success), the probability that the packet will be *eventually* transmitted successfully approaches one as *N* increases without bound. Figure 4 shows a plot of the probability of success vs. *p* for the specific cases of *N* = 1, 2, 3, 4, and 5. At this point we discuss the effects of *p* and *N* on the network performance.

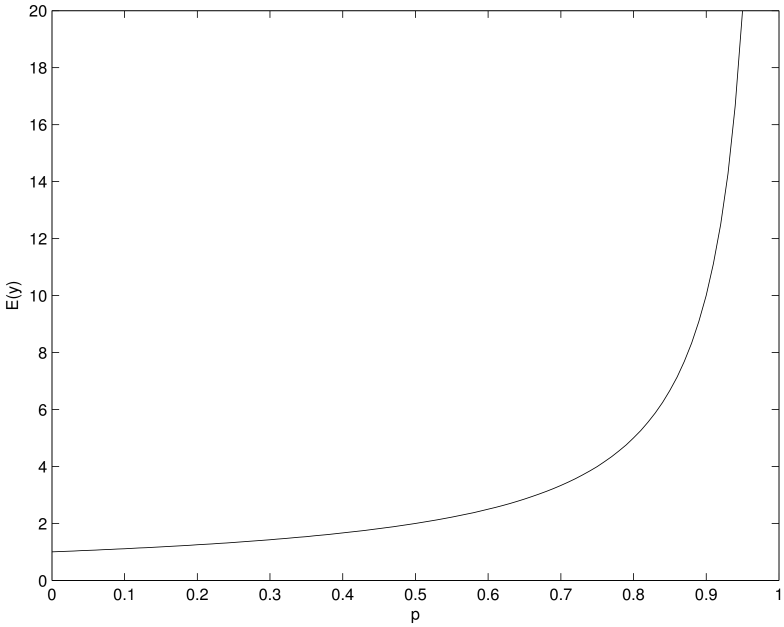

Figure 2. The average number of transmissions required to send one packet across a single network link.

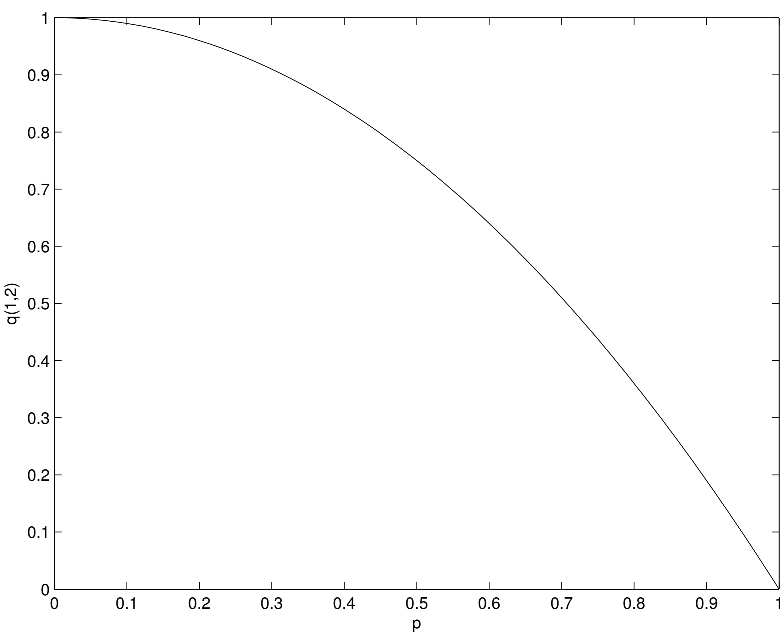

Figure 3. The probability that a single packet will be sent without error across one network link in at most two tries.

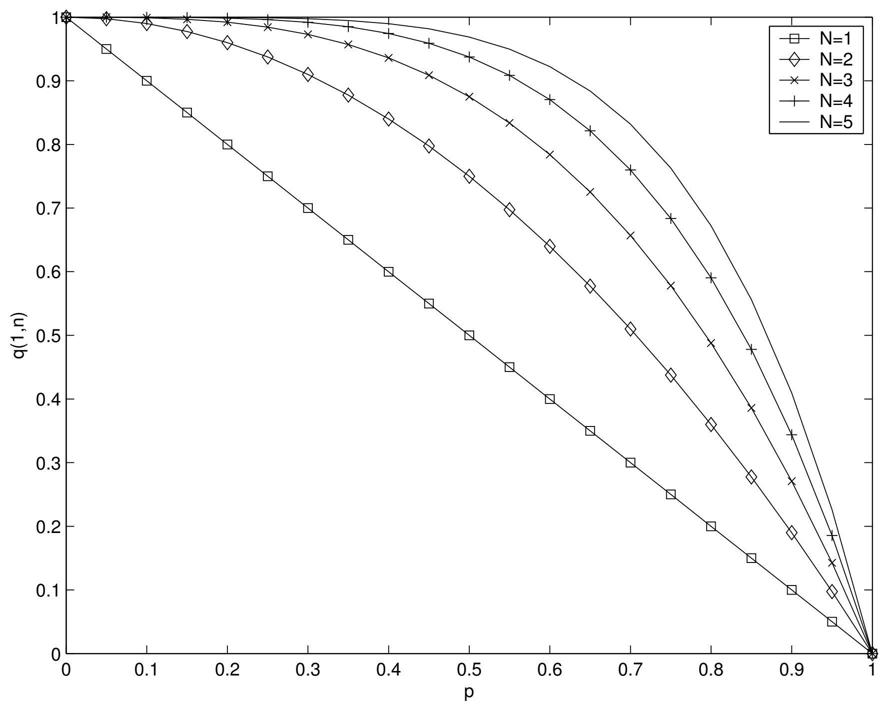

Figure 4. The probability that one packet will be transmitted across a single link without error in at most N tries, N=1 to 5.

If a packet is transmitted N times without success, the sender should be notified that a failure has occurred and to try again later. The choice of N depends on the number of users on the network and traffic patterns observed by the network managers, transmission medium, environmental conditions, and network delay. For instance, if a user is working late at night when there is virtually no competing network traffic, and repeated attempts at sending a packet fail, then the error is probably due to a network failure instead of competition. On the other hand, if multiple users are attempting to send information across a network at the same time, more transmission attempts should be made before deciding that the problem is due to a network failure.

On the project quiz questions, we ask students to apply the binomial distribution and introduce them to the Pascal distribution and exponential distribution with questions similar to the following:

1. Assume that you are transmitting 10 packets and each packet is transmitted only once.

   What is the probability that exactly three packets are transmitted successfully in 10 attempts? (Binomial distribution)

2. To test a particular computer link or line, you transmit the same packet repeatedly.

   a) What is the probability that five packets will be transmitted correctly on the 7^{th} transmission?

b) What is the probability that two packets will be damaged on the 7th transmission? (Pascal distribution)

Single Link: Multiple packets

An e-mail message will be partitioned into several packets before being transmitted over the network. Suppose that there are K packets, each of which is transmitted separately. Depending on the values of p and K, it may be very unlikely for the entire message to transmit successfully on the first attempt. All packet transmissions must succeed, making this probability

$$Q_K = (1-p)^K . \qquad (5)$$

Q_K is plotted in Figure 5 for several values of K.

Suppose that packets are sent over the network in succession, and if any packet transmission fails, the failed packet is retransmitted until success occurs. Packets successfully transmitted before a later packet fails are not resent. We define the random variable Z to be the average number of transmissions required for a successful transmission of the entire e-mail.

E(Z) can be computed easily since the success or failure of any packet transmission is independent from the other $K-1$ transmissions, and we found earlier that a single packet takes (on average) $1/(1-p)$ transmissions to succeed, so we expect that the complete set of K packets will be transmitted in $K/(1-p)$ attempts. Figure 6 shows the expected number of transmissions, E(Z), vs. p for several values of K using a logarithmic scale for readability.

We can use a simulation to verify experimentally that $K/(1-p)$ is the correct probability expression. The MATLAB program in Appendix A runs N simulations for set values of p and K and counts the number of transmissions required, then averages over the N experiments. Sample runs were performed with the following parameter sets:

- N=1000, K=5, p=0.5
- N=1000, K=11, p=0.47

The mean transmission counts for these two experiments were, respectively, 9.961 and 20.847. Theoretical values derived from our previous discussion are 10.0 and 20.8, which agree with the results of our simulation. A histogram of the transmission count for the 1000 simulations with p=0.5 is shown in Figure 7. This plot demonstrates how the values are concentrated around the theoretical value of 10.

As we have observed, this type of network raises some serious concerns about system reliability. For any file partition consisting of a large number of packets, many transmissions are expected before the file is sent in its entirety. This number grows rapidly with increasing probability of packet damage on a link. When considering the two remaining network structures, we will return to this issue as a basis for comparison.

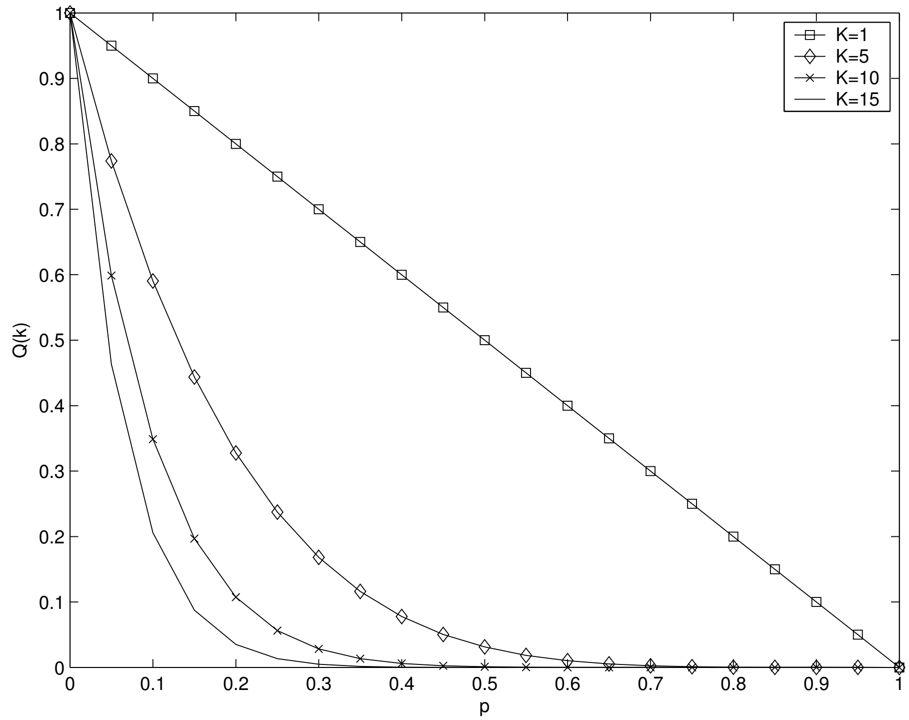

Figure 5. The probability that *K* packets of an e-mail will all transmit across one link successfully in the first attempt for *K* =1,5,10,15.

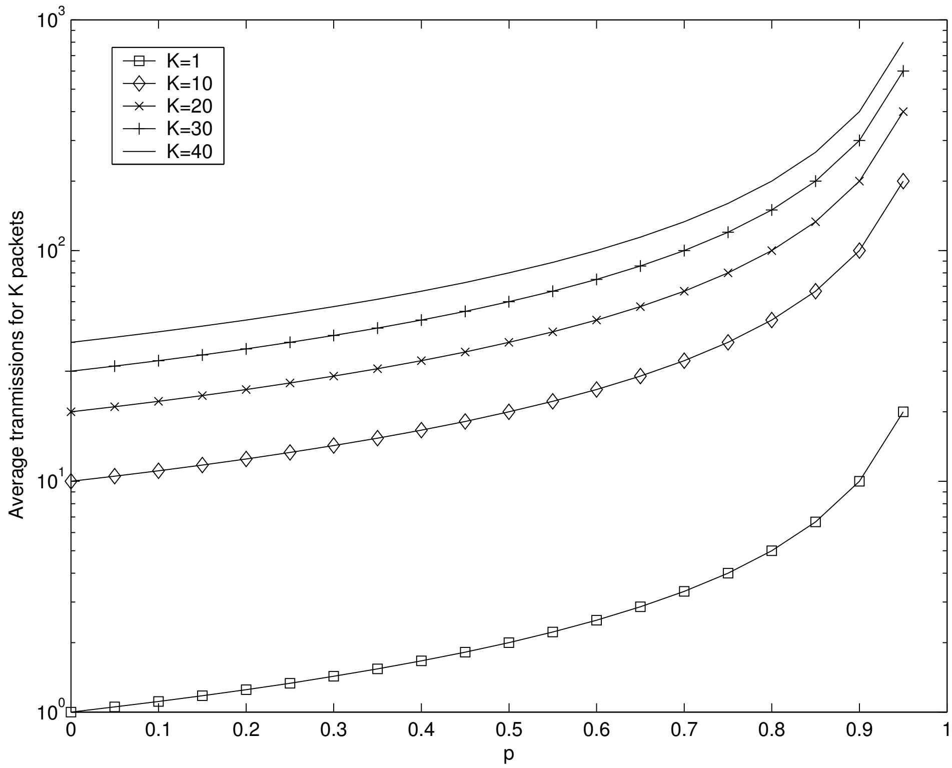

Figure 6. The expected number of transmissions required to send *K* packets across one network link without error for *K*=1,10,20,30,40.

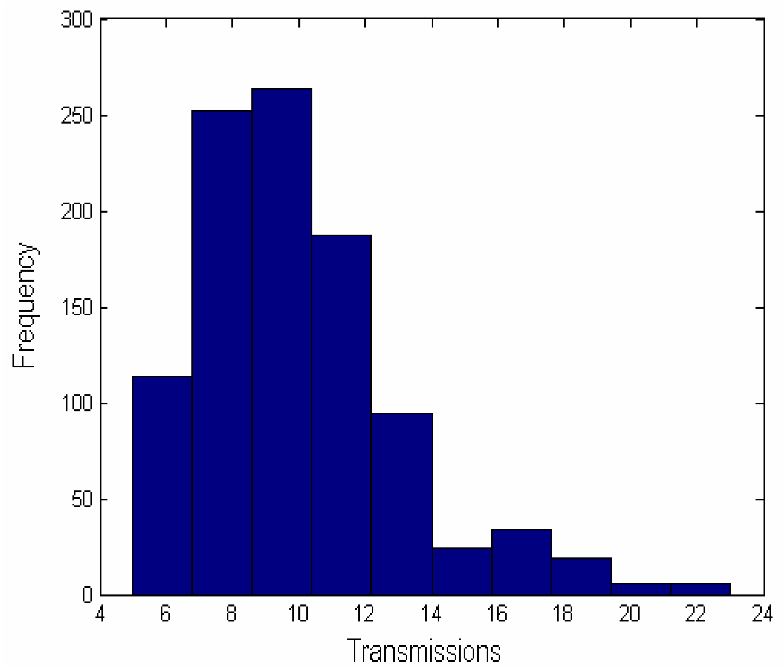

Figure 7. Histogram of transmissions required to send 5 packets across a single link, whose probability of failure is 0.5. 1000 simulations were performed.

At this point, we introduce the exponential distribution function and the normal distribution and assume K, the length of an e-mail message, is a random variable with an exponential distribution. We usually ask students the following.

Assume the length of an e-mail message is exponentially distributed with a mean of 1000 bytes.

a) What is the probability density function of K?

b) What is the cumulative distribution of K?

c) What is the variance of K?

d) What is the probability that $100 \le K \le 1000$?

Series Two-Link Network

Suppose that each packet in the K-packet e-mail must be transmitted across two network links connected in series. The probability of losing a packet on any one link is still p. During the transmission process, if a packet is damaged on any one of the two links, it must be retransmitted across the first link again.

The probability that a packet is transmitted successfully across both links in the network is $(1-p)^2$ because a successful transmission with probability $(1-p)$ must succeed in two consecutive, independent trials. Figure 8 shows a plot of this probability and compares it to the probability that a transmission across a single network link succeeds.

An e-mail composed of *K* packets is much less likely to transmit successfully on the first try. In such a case, all *K* packets must transmit successfully across two links, or a total of 2*K* successful transmissions. This occurs with probability $(1-p)^{2K}$ and is plotted vs. *p* in Figure 9 for several values of *K*.

Now we ask the following question: Let Z be the random variable representing the number of times a packet is transmitted over the first link until it is received successfully. What is the average number of transmissions over the first link ,E(Z), required to send a *single* packet across this network? We use the same rules as stated above: if a packet transmission fails (with probability *p*), we must send the packet across both links again. Let $q_{2,k}$ be the probability that the packet is transmitted successfully in exactly *k* tries. Then $q_{2,k}$ can be expressed as the following recursion:

$$q_{2,1} = (1-p)^2$$

$$\text{and } q_{2,k} = \left(1-\sum_{j=1}^{k-1} q_{2,j}\right)(1-p)^2 \qquad (6)$$

We note that in order for the attempt to succeed on the k^{th} try, all previous attempts must have failed. The probability that the attempt succeeded for some $j<k$ is $\sum_{j=1}^{k-1} q_{2,j}$, so the probability that it failed is $1-\sum_{j=1}^{k-1} q_{2,j}$. Finally, the k^{th} attempt must succeed, so we have the $(1-p)^2$ term at the end. Through some manipulation we can derive an explicit, non-recursive form for (6):

$$q_{2,k} = \left(1-\sum_{j=1}^{k-1} q_{2,j}\right)(1-p)^2$$
$$q_{2,k} = (1-p)^2 - (1-p)^2 \sum_{j=1}^{k-1} q_{2,j}$$
$$q_{2,k} = q_{2,k-1} - (1-p)^2 q_{2,k-1}$$
$$q_{2,k} = p_{k-1}\left(1-(1-p)^2\right)$$
$$q_{2,k} = q_{2,2}\left(1-(1-p)^2\right)^{k-1}$$

$$q_{2,k} = (1-p)^2\left(1-(1-p)^2\right)^{k-1} \qquad (7)$$

Then the average number of transmissions required to send a single packet across is given by

$$E(Z) = \sum_{k=1}^{\infty} k(1-p)^2\left(1-(1-p)^2\right)^{k-1} = 1/(1-p)^2 . \quad (8)$$

This can be easily shown using equation (2) when we note that the probability of a damaged packet along both links is $1-(1-p)^2$.

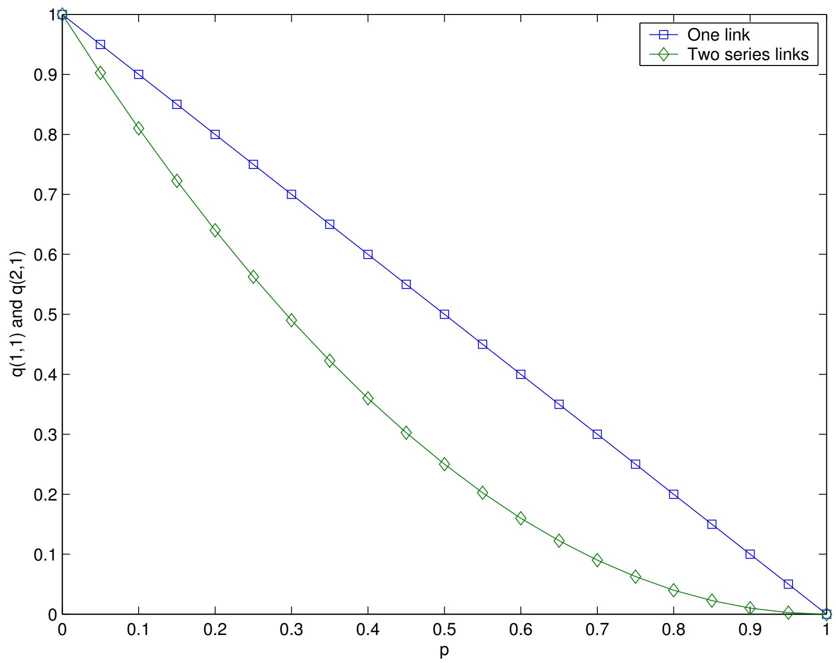

Figure 8. The probability that one packet will transmit across a network without error on the first attempt for a one-link and series two-link connection.

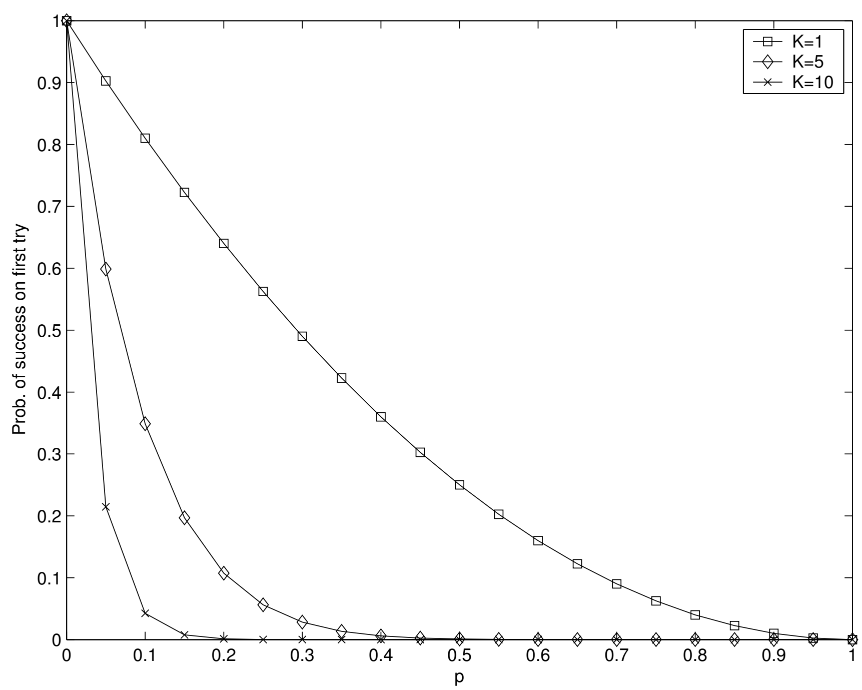

Figure 9. The probability that all K packets of an e-mail will transmit across the series network on the first attempt for K=1,5,10,15,20.

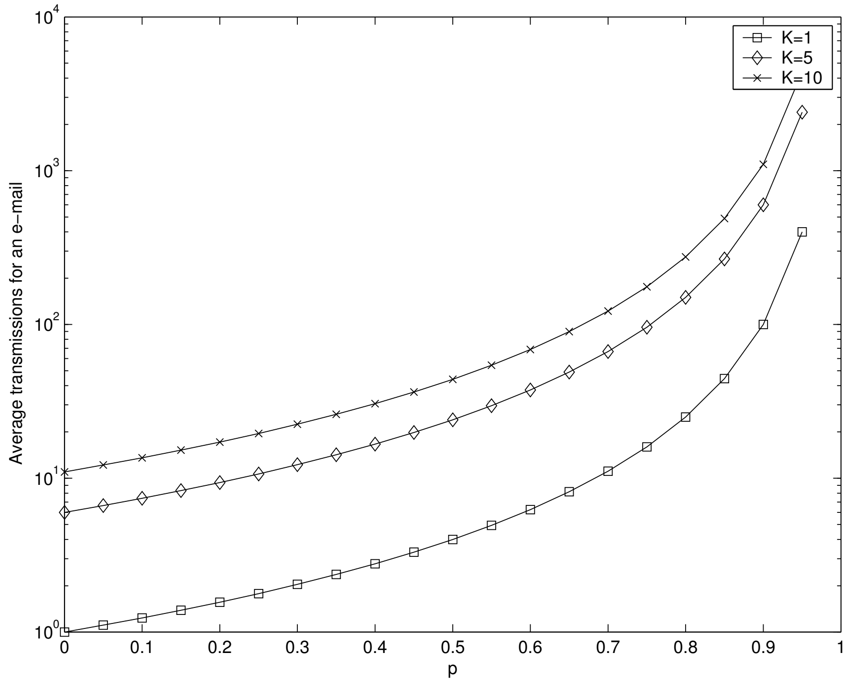

Figure 10. The average number of transmissions required to send all *K* packets across the series network without error (log scale) for *K*=1,6,11,16,21.

However, our e-mail has *K* separate packets, each of which takes an average of $1/(1-p)^2$ transmissions (we still assume that a packet is not transmitted until the previous packet has crossed the network successfully). Therefore, the total number of expected packet transmissions for the e-mail is $K/(1-p)^2$. Figure 10 shows a plot of this function vs. *p* for several values of *K*. Simulations were performed to verify this expression; the results agreed strongly with the analytical prediction. The program in Appendix A was used, with modifications to include two uniformly distributed random numbers instead of one and *both* be less than *p* for a successful transmission.

If we already assume that a packet is transmitted successfully over the first link, it is much easier to compute the probability of a successful transmission across the network. The probabilities of successful transmission across either link are independent; then the probability that the packet is transmitted without error over the second link, given that it was transmitted correctly over the first link, is $(1-p)$.

The series-link connection exhibits a poorer overall performance than the single-link network. The simple explanation for this is that a greater number of transmissions must take place, increasing the chances that failures will occur. As Figure 8 demonstrates, this is true even if only one packet is considered.

Let us define the random variable W to be the number of transmissions across both links for successful transmission. What is E(W)? Simulation can be a powerful tool to estimate E(W). Figure 11 shows a plot of E(W) vs. p obtained through simulation.

We can compute E (W) analytically by noting that

$$\begin{aligned} q_{2,2} &= (1-p)^2 \\ q_{2,3} &= p(1-p)^2 \\ q_{2,4} &= pq_{2,3} + (1-p)pq_{2,2} \qquad (9) \\ &\ldots \\ q_{2,k} &= pq_{2,k-1} + p(1-p)q_{2,k-2} \end{aligned}$$

where $q_{2,k}$ is the probability of success in k transmissions on both links and p is the probability of failure on a link.

$$E(W) = \sum_{k=2}^{\infty} kq_{2,k} = \frac{2-p}{(1-p)^2} \qquad (10)$$

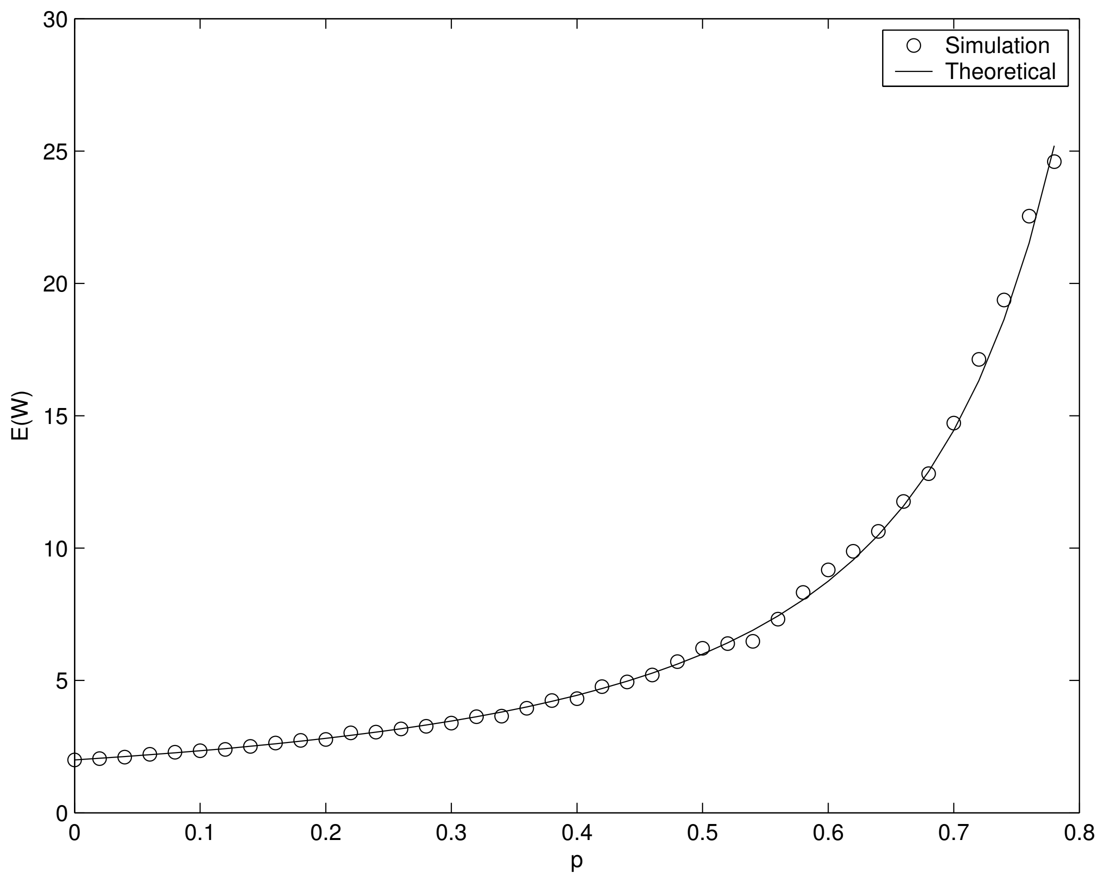

Figure 12. The average number of transmissions across two links vs. p

Series Multiple-Link Network

Now suppose that we have a network of *k* links, all in series, and that we would like to transmit a single packet across all the links. As before, damage occurs with probability *p*, and once a packet is damaged, it is lost along the entire network and must be retransmitted from the first node. How many packets N_k should we expect to send?

One way to answer this question is to notice the following recursion:

$$N_{k+1} = N_k N_1 + N_1 \qquad (11)$$

We expect $N_1 - 1$ failed transmissions across the final link (since only one out of the N_1 can succeed), which requires a total of $N_k(N_1 - 1)$ transmissions across the first *k* links that lead to a failure. Furthermore, we require one transmission across the final link for success, and N_k transmissions across the first *k* links for that success. This yields $N_k N_1 + N_1$ transmissions in total for *k*+1 links.

The initial condition for the recursion is $N_1 = (1-p)^{-1}$ which we have already derived. The recursion can be encoded in the matrix equation

$$\begin{bmatrix} N_{k+1} \\ 1 \end{bmatrix} = \begin{bmatrix} N_1 & N_1 \\ 0 & 1 \end{bmatrix}^k \begin{bmatrix} N_1 \\ 1 \end{bmatrix}. \qquad (12)$$

The eigenvalues and eigenvectors of the recursion matrix are

$$\begin{aligned} \lambda &= \{1, N_1\} \\ \varphi &= \left\{ \begin{bmatrix} \frac{N_1}{1-N_1} & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix} \right\}' \end{aligned} \qquad (13)$$

which leads to the following solution:

$$\begin{bmatrix} N_k \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{N_1}{1-N_1} \\ 1 \end{bmatrix} + \frac{N_1^{\,2}}{N_1 - 1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} N_1^{\,k-1} \qquad (14)$$

$$N_k = \frac{N_1}{1-N_1} - \frac{N_1^{\,k+1}}{1-N_1} \qquad (15)$$

$$N_k = \frac{1-(1-p)^k}{p(1-p)^k} \qquad (16)$$

A quick check of the boundary conditions shows that $N_k \to \infty$ as *p* nears 1 (all transmissions fail), and $N_k \to k$ as *p* approaches zero (all transmissions succeed), which is what we expected.

Parallel Two-Link Network

Suppose that our network consists of two links connecting a pair of common nodes (a parallel network). The probability that a packet is damaged during a transmission across the two links is p_a and p_b, respectively. The packet travels across both links to arrive at the destination. The packet will reach its destination through both links with a probability of $(1-p_a)(1-p_b)$, since the transmission must succeed on both links simultaneously. If $p_a = p_b$, then the probability reduces to $(1-p_a)^2 = (1-p_b)^2$; this behavior is plotted in Figure 12 for $p_b = 0$ to 1.

The packet transmission is considered successful if the packet travels across either parallel link without getting damaged. This occurs with probability $(1-p_a p_b)$ because the only chance that the transmission is unsuccessful is if failure occurs in both links (which occurs with probability $p_a\ p_b$). If $p_a = p_b$, the packet will arrive with probability $(1-p_a{}^2)$; Figure 13 shows an updated version of Figure 8 to include the parallel network. We can argue from these results that of the three network types considered in this paper, the parallel network provides the highest probability of transmission success on the first try. The intuitive reason for this is that there are the most independent paths to the destination in this type of network design, which increases the chances that the packet will travel successfully across *one* of them.

Because the transmission successes across the two links are statistically independent events, we can summarize with the following observations:

- Given that the packet is correctly received across the first link, the probability that the packet is transmitted correctly on the second link is $(1-p_b)$.
- Given that the packet is damaged along the first link, the probability that it is transmitted correctly on the second link is still $(1-p_b)$.

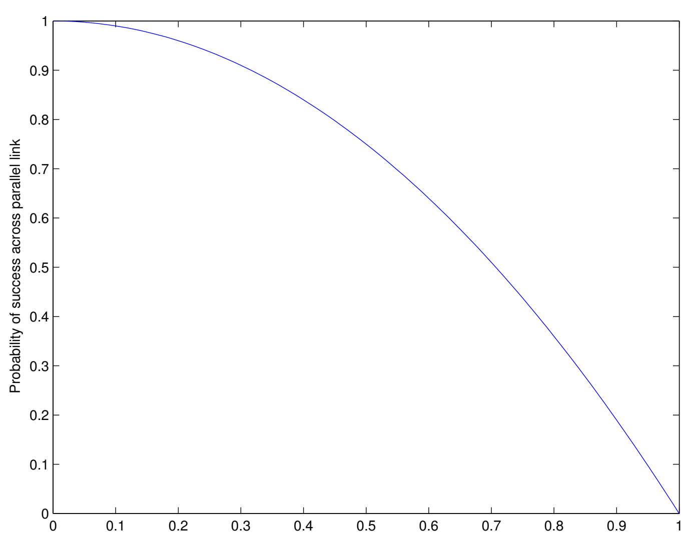

Figure 12. The probability that a single packet travels without error through both links of a parallel network on the first attempt.

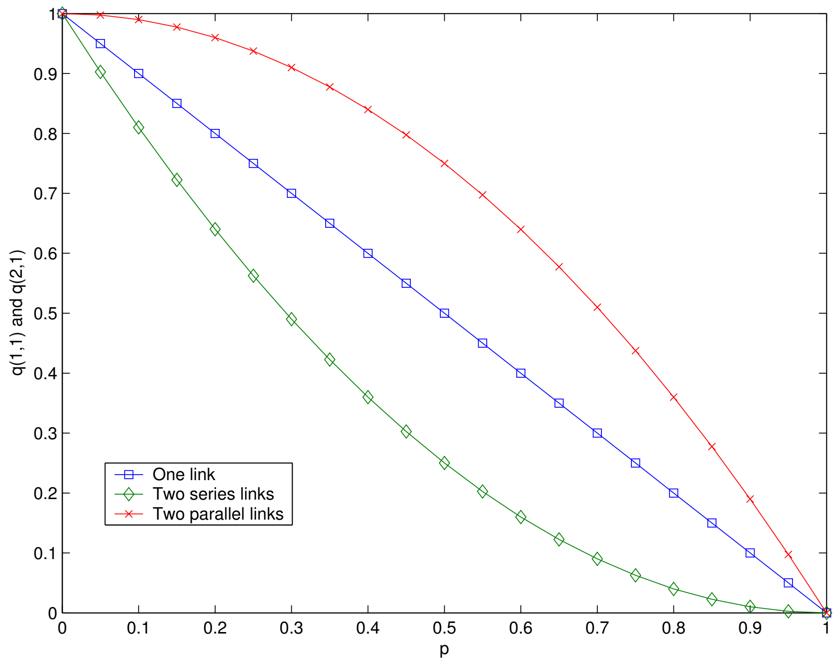

Figure 13. The probability that one packet will betransmited across a network without error on the first attempt for one-link, series two-link, and parallel two-link connections.

The parallel-link network exhibits the best behavior of the three structures we have examined. Because more links connect a common pair of routers, there is a higher chance that a transmission will succeed because only one of the links needs to pass the packet through undamaged. The situation is analogous to the two-input digital OR and AND logic gates; assuming that all input combinations are equally likely, there is a higher chance that the output of the OR gate will be logical HIGH because only one of its inputs needs to be HIGH, while both of the inputs to the AND gate need to be HIGH for the same output condition. In this case, the series network acts like an AND gate and the parallel network acts like an OR gate.

On quiz questions we address similar issues when a packet is sent over N links (in series) and when sending a packet over M links simultaneously (links are in parallel). We also consider a variety of N series links in series with M links in parallel.

Summary

We presented a project used in our random signals and noise course with applications to computer networks. The project introduced random variables with Bernoulli distributions, geometric distributions, binomial distributions, and Pascal distributions. It also introduced students to the computations of means of random variables. Simulation is a powerful tool and can enhance student learning and clarify fundamental concepts. The examples we presented can introduce design concepts and parameter sensitivity and reliability.

References

1. Roy D. Yates and David J. Goodman, *Probability and Stochastic Processes: A Friendly Introduction for Electrical & Computer Engineers*, 2^{nd} Edition, John Wiley & Sons, 2005.
2. T.T. Soong, *Fundamentals of Probability and Statistics For Engineers*, John Wiley& Sons, 2004
3. Rodger E. Ziemer, *Elements of Engineering Probability and Statistics*, Prentice-Hall, 1997. (ISBN: 0-02-431620-2)
4. Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, Belmont, MA, 2002.
5. Athanasios Papoulis and S. Unnikrishan Pillai,*Probability, Random Variables and Stochastic Processes*, 4th Edition, McGraw-Hill, 2002.
6. Jay L. Devore, *Probability and Statistics: For Engineering and the Sciences,* Duxbury, Pacific Grove, CA, 2000.
7. George R. Cooper and Clare D. McGillem, *Probablistic Methods of Signal and System Analysis*, Third edition, Oxford University Press, 1999.
8. Maurice F. Aburdene and Richard J. Kozick, "A project-oriented course in probability and statistics for undergraduate electrical engineering students" Proceedings of Frontiers in Education Conference, 1997, pp.598 - 603
9. Leonard Kleinrock, *Queueing Systems, Computer Applications*, The Academic Press, Vol. 2, 1976.
10. William Stallings, *Data and Computer Communications*, 6th Edition, Prentice-Hall, 2000.
11. D. Bertsekas and R. Gallager, *Data Networks*, 2nd Edition, Prentice Hall, Upper Saddle River, NJ, 1992.
12. Douglas E. Comer, *Computer Networks and Internets with Internet Applications*, 4th edition, Prentice Hall, Upper Saddle River, NJ, 2004.
13. James. F. Kurose and Keith. W. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet,* 2nd edition, Addison-Wesley, Boston, MA, 2005.
14. Larry L. Peterson and Bruce S. Davie, *Computer Networks – A Systems Approach*, Morgan Kaufman Publishers, San Francisco, CA, 2003.
15. Andrew S. Tanenbaum, *Computer Networks*, Prentice Hall Publishing, 2003.

APPENDIX A: MATLAB Simulation Program

```
% Parameter declaration
K=11;    % number of packets
P=0.47;  % probability of packet damage
n=1000;  % number of simulations

% Main loop
data = ones(1,n); % a place to store the results
for k=1:n,

transmissions=0;     % transmission count
packetsThrough=0;    % number of packets that have made it across

while packetsThrough < K,
      r=rand;
      transmissions = transmissions + 1;
      while r < P,
          r=rand;
          transmissions = transmissions + 1;
      end
      packetsThrough = packetsThrough + 1;
end
data(k)=transmissions;
end

% show the results
mean(data)
hist(data)

% Done
```