Error Detecting and Error Correcting Codes

R. W. Hamming

1950

1 Introduction

The author was led to the study given in this paper from a consideration of large scale computing machines in which a large number of operations must be performed without a single error in the end result. This problem of "doing things right" on a large scale is not essentially new; in a telephone central office, for example, a very large number of operations are performed while the errors leading to wrong numbers are kept well under control, though they have not been completely eliminated. This has been achieved, in part, through the use of self-checking circuits. The occasional failure that escapes routine checking is still detected by the customer and will, if it persists, result in customer complaint, while if it is transient it will produce only occasional wrong numbers. At the same time the rest of the central office functions satisfactorily. In a digital computer, on the other hand, a single failure usually means the complete failure, in the sense that if it is detected no more computing can be done until the failure is located and corrected, while if it escapes detection then it invalidates all subsequent operations of the machine. Put in other words, in a telephone central office there are a number of parallel paths which are more or less independent of each other; in a digital machine there is usually a single long path which passes through the same piece of equipment many, many times before the answer is obtained.

In transmitting information from one place to another digital machines use codes which are simply sets of symbols to which meanings or values are attached. Examples of codes which were designed to detect isolated errors are numerous; among them are the highly developed 2 out of 5 codes used extensively in common control switching systems and in the Bell Relay Computers, the 3 out of 7 code used for radio telegraphy, and the word count sent at the end of telegrams.

In some situations self checking is not enough. For example, in the Model 5 Relay Computers built by Bell Telephone Laboratories for the Aberdeen Proving Grounds, observations in the early period indicated about two or three relay failures per day in the 8900 relays of the two computers, representing about one failure per two to

three million relay operations. The self-checking feature meant that these failures did not introduce undetected errors. Since the machines were run on an unattended basis over nights and week-ends, however, the errors meant that frequently the computations came to a halt although often the machines took up new problems. The present trend is toward electronic speeds in digital computers where the basic elements are somewhat more reliable per operation than relays. However, the incidence of isolated failures, even when detected, may seriously interfere with the normal use of such machines. Thus it appears desirable to examine the next step beyond error detection, namely error correction.

We shall assume that the transmitting equipment handles information in the binary form of a sequence of 0's and 1's. This assumption is made both for mathematical convenience and because the binary system is the natural form for representing the open and closed relays, flip-flop circuits, dots and dashes, and perforated tapes that are used in many forms of communication. Thus each code symbol will be represented by a sequence of 0's and 1's.

The codes used in this paper are called systematic codes. Systematic codes may be defined as codes in which each code symbol has exactly n binary digits, where m digits are associated with the information while the other $k = n^{-}m$ digits are used for error detection and correction. This produces a redundancy R defined as the ratio of the number of binary digits used to the minimum number necessary to convey the same information, that is,

$$R = n/m.$$

This serves to measure the efficiency of the code as far as the transmission of information is concerned, and is the only aspect of the problem discussed in any detail here. The redundancy may be said to lower the effective channel capacity for sending information.

The need for error correction having assumed importance only recently, very little is known about the economics of the matter. It is clear that in using such codes there will be extra equipment for encoding and correcting errors as well as the lowered effective channel capacity referred to above. Because of these considerations applications of these codes may be expected to occur first only under extreme conditions. Some typical situations seem to be:

a. unattended operation over long periods of time with the minimum of standby equipment.

b. extremely large and tightly interrelated systems where a single failure incapacitates the entire installation.

c. signaling in the presence of noise where it is either impossible or un- economical to reduce the effect of the noise on the signal.

These situations are occurring more and more often. The first two are par- ticularly true of large scale digital computing machines, while the third occurs, among other places, in "jamming" situations.

The principles for designing error detecting and correcting codes in the cases most likely to be applied first are given in this paper. Circuits for implementing these principles may be designed by the application of well- known techniques, but the problem is not discussed here. Part I of the paper shows how to construct special minimum redundancy codes in the following cases:

a. single error detecting codes

b. single error correcting codes

c. single error correcting plus double error detecting codes.

Part II discusses the general theory of such codes and proves that under the assumptions made the codes of Part I are the "best" possible.