Pergamon

Information Processing & Management, Vol. 32, No. 5, pp. 555–562, 1996
Copyright © 1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0306-4573/96 $15+0.00

S0306-4573(96)00009-X

CHAOTIC BEHAVIOR IN COMPUTER MEDIATED NETWORK COMMUNICATION

HERBERT SNYDER[1]* and DOUGLAS KURTZE[2]

[1] School of Library and Information Science, Indiana University, Bloomington, IN 47405, U.S.A.
[2] Dept. of Physics, North Dakota State University, Fargo, North Dakota, U.S.A.

(Received 10 April 1995; accepted in final form 20 February 1996)

Abstract—The work examines the use of chaos theory in modelling time series data generated by computer mediated communication (CMC). Data generated by CMC bulletin boards is examined for the presence of chaotic behavior and to assess the variance which can be accounted for by the deterministic mechanism. The study regards the time series data generated from a CMC discussion group as the sum of two components. One is a deterministic "signal" which presumably obeys some unknown chaotic dynamics. The other component is truly random "noise". The study's overall goal is to assess the relative importance of these two components, using techniques devised by Procaccia and Grassberger for studying chaos in time series data. Chaotic time series data had increasingly larger fractions of noise added until chaotic behavior was no longer found. Analysis of the data with added noise indicates that from 20 to 30% of the variation in the data is the result of noise. Conversely, 70 to 80% of the variation in the data can be accounted for by deterministic chaos. Implications for future research using chaos and CMC are discussed. Copyright © 1996 Elsevier Science Ltd

INTRODUCTION

It has been estimated that there are currently more than two million Internet hosts, a number that can be expected to grow even larger in the future (Internet Society, 1993). With such a large population of users, the networks that make up the Internet have already begun to show signs of congestion. Given this, there is a need for innovative techniques for modelling traffic flow on the Internet. The goal of this paper is to explore the usefulness of non-linear dynamics, or chaos, for describing the phenomenon of Internet network traffic.

Chaos theory has been well-documented in both the popular and scientific presses, and this paper will not cover basic theory in depth. Briefly, chaos theory deals with random-appearing behavior that is the product of simple, rigorously defined systems. The key to the apparently random behavior is that chaotic systems are characterized by what is known as sensitive dependence on initial conditions. That is, for the same deterministic system, two sets of initial conditions that are very close will produce widely divergent results over time. The classic example of this aspect of chaotic behavior is the so-called "butterfly effect" of global weather systems (Lorenz, 1978). Prediction for small weather systems (e.g., thunderstorms or tornadoes) becomes worthless more than 6 or 7 days in advance. The reason for this is that extremely small perturbations in the atmosphere (the "flap of a butterfly's wings in Brazil") multiply until they produce much larger turbulent outcomes ("tornadoes in Texas") in a very short period of time. This leads to the counter-intuitive situation in which simple ordered systems can be capable of extremely complex, or even random, behavior. (Readers who are interested in a more comprehensive discussion of chaos are directed to the Gleick, 1987 and Kellert, 1993 books in the references.)

Chaos theory is most useful for investigating phenomena in which there is very little or no

* To whom all correspondence should be addressed.

periodicity, but in which there is some evidence of an underlying causal mechanism. This appears to be the case for computer-mediated communication (CMC) traffic. A preliminary exploration of chaos theory and its applicability to CMC traffic was conducted using message postings to the IBM-PC Digest. A Fourier analysis of CMC traffic indicated a weak periodicity at 7-day intervals (slightly higher message traffic on Mondays) but no other indication of periodic behavior which can be exploited. There is some indication, however, that CMC may still have an underlying causal mechanism, specifically the phenomenon of message posting.

Messages were chosen as a unit of analysis because in order for a phenomenon to be a useful subject for study, it should be possible (at least as a thought experiment) to posit some simple, underlying mechanism that causes the behavior in the system. In the case of messages to an electronic bulletin board, the mechanism is posting a message, which in turn produces other replies, which produce further postings, etc. This is not to say that the process is this simple or linear, or even that it behaves chaotically, but only that experience tells us the behavior of the system is close enough to a simple dynamic to make it useful to investigate using chaos theory.

In order to evaluate the relevance of chaos theory to analyzing the phenomenon of CMC traffic it is necessary to do the following things:

- Demonstrate that CMC traffic behaves chaotically.
- Explore the utility of chaos analysis for analyzing data generated by behavior as disorderly as human communication. Chaos analysis is usually applied to systems that behave relatively simply (e.g. biological populations, crystal growth, etc.). There is a concern, however, that chaos analysis may indicate chaotic behavior is present when applied to human-generated data not because there is underlying deterministic chaos, but because the data are extremely complex.
- Estimate the amount of variance in CMC traffic that can be accounted for by deterministic chaos. Chaos may be a major factor in explaining CMC or a minor factor overlaid with large amounts of noise. The presence of chaotic behavior does not indicate how strongly or completely it can explain the phenomenon in which it is found.
- Determine the predictive power of chaos analysis in the context of CMC. If CMC is governed by extremely complex chaotic processes it may not be possible to identify them with finite amounts of data or to extract the deterministic mechanism that describes the communication behavior.

This paper expands on the results of several earlier, preliminary studies (Kurtze *et al.*, 1992; Snyder & Kurtze, 1992; Kurtze & Snyder, 1993) that have examined the first three issues mentioned above.

METHODOLOGY

In order to address the first three concerns listed above our study undertakes to do three things: (1) examine a sample of time-series data generated by CMC for evidence of chaotic behavior; (2) randomly reorder the time-series and reanalyze the data to ascertain if evidence of chaotic behavior is a spurious finding as the result of data complexity rather than some underlying deterministic mechanism; and (3) estimate the amount of variance in the time series data that can be accounted for by an underlying deterministic mechanism. The rationale for the samples selected and the specific techniques for data analysis are described in greater detail in the following sections.

Sample selection

Computer networks. Computer mediated discussion groups typically deal with a specific area of interest such as philosophy or computer-use. Participants in the discussion groups read and respond to messages which have been previously posted on the network; single messages frequently generate a number of other messages in reply.

An examination of the discussion groups available on the USENET system was made. There were three criteria for selection: (1) scholarly or technical discussion subject, (2) availability of discussion group archives (i.e., copies of all messages sent to the discussion group), and (3) continuous use of the group for at least 5 years. The discussion group which was finally selected was IBM-PC Digest. IBM-PC Digest deals with issues concerning IBM personal computer use. The group draws its participants almost exclusively from academic or research institutions, and has been in operation continuously since 1982. To the best of the authors' knowledge, this places it among the longest continuously-running groups on the USENET system. Complete archives were available for all traffic on the network, and copies of these were obtained for the period January 1982–November 1989. (There is no reason to believe that the methodology might not be applicable to a variety of other discussion groups, however, at the time of the research there was no other archive of comparable length on which to test further. Additional work is planned by the authors as more large data sets become available.)

All messages sent to the system have a standard heading which includes the data and time of transmission. A computer program was written which identified the date and counted the number of messages posted each day, including an extra day for leap years and days for which there were zero postings. The resulting data was in the form of a time-series of 2707 separate days, each paired with the total number of messages sent that day. The mean and standard deviation of the sample was 4.82 and 4.26, respectively (calculated on the basis of single-day data).

Investigating deterministic chaos in time-series data

A time series obtained from a purely deterministic, nonlinear dynamical system can be "chaotic," resembling a purely rarndom time series so closely that the eye cannot distinguish between the two. The two can be distinguished, however, by calculating their respective fractal dimensions (i.e. a non-integral dimension of the strange attractors that characterize chaotic systems). To see how this is done, suppose the dynamics produce each new datum x in the time series using the information contained in the m-most recent data:

$$x_{\{i\}}=f(x_{\{i-1\}}, x_{\{i-2\}}, \ldots, x_{\{i-m\}}). \tag{1}$$

Another way to say this involves arranging the data into m-tuples of consecutive data, of the form $(x_{\{i-1\}}, x_{\{i-2\}}, \ldots, x_{\{i-m\}})$. Then each m-tuple contains all the information needed to compute the next m-tuple:

$$(x_{\{i-1\}}, x_{\{i-2\}}, \ldots, x_{\{i-m\}}) \rightarrow x_{\{i\}}, (x_{\{i-1\}}, x_{\{i-m+1\}}) \tag{2}$$

where the new datum $x_{\{i\}}$ is given by the eqn (1) above.

Geometrically, then, the dynamics take each point in the m-dimensional space of m-tuples of consecutive data to another point in that space. Note that the m-tuples overlap; this is part of the definition of their dynamics. If the underlying function f uses only the m most recent data (or fewer) to determine the next, then each m-tuple contains all the information needed to determine its successor. Chaotic dynamics characteristically concentrate the m-tuples on a subset of the full m-dimensional space, called an attractor. Typically, the attractor is a fractal, with a dimension d which is less than m, and which is not an integer. On the other hand, if the data are actually random (independent and identically distributed), then the m-tuples will be distributed randomly in their m-dimensional space, and will not concentrate on any lower-dimensional subspace.

It is thus possible to distinguish a purely deterministic, chaotic time series from a random one by calculating the fractal dimension of the figure formed by the m-tuples of data points. For random data, the dimension will be m, while for chaotic, deterministic data, it will be less than m. One small complication in this prescription is the fact that m, the number of data points which the putative chaotic dynamics actually needs to calculate the next point in the time series, is unknown. This is actually an advantage, however. If we choose a value for m which is too high, then the chaotic dynamics actually uses fewer than m values to compute the next datum, and so the dimension of the attractor is virtually unaffected. Random data will always give a dimension equal to m, whatever value of m we have chosen.

The standard method for calculating the fractal dimension of an attractor from a time series of points on the attractor is the correlation algorithm of Grassberger and Procaccia (Grassberger & Procaccia, 1983). Conceptually, this counts the average number of m-tuples which lie within a distance r of a given m-tuple. The volume of a fractal of dimension d lying within a distance r of a given point on the fractal is proportional to r^d, so we can estimate the dimension d by fitting the dependence of the counts on r by a power law. The calculation is actually done by computing the correlation sum.

$$C(r, N) \approx (1/N^2)\ [\text{number of pairs}(i, j) \text{ for which } |\|v_i\| - \|v_j\|| <= r] \qquad (3)$$

where N is the total number of m-tuples, then plotting log (C) as a function of log(r), and fitting a straight line through as large a section of the plot as possible. The slope of the line is then an estmate of d.

There are several technical difficulties with the procedure, mostly having to do with the fact that an N of 2707 is actually a very small sample for these techniques. Another difficulty arose because the data themselves are small integers. We used the 1_∞ definition of the distance between two m-tuples, namely the absolute value of the greatest difference between corresponding elements of the two. Since the data are small, the greatest distance between two m-tuples, namely the absolute value of the greatest difference between corresponding elements of the two. Since the data are small, the greatest distance between any two m-tuples comes out to be 31, so that the distance r only ranges from 1 to 31, a range of 1–1/2 decades. As a result, the correlation plot was well approximated by a straight line over less than a decade of r. In order to increase the range of r, we used data for consecutive 3-day periods instead of data for consecutive days. This increases the size of the data and hence the range of possible r values, at the cost of decreasing the size N of the data set; the choice of 3-day intervals, rather than, e.g. 4- or 2-day intervals, represents a compromise.

Examining noise and deterministic chaos in CMC data

If CMC data exhibit chaotic behavior we may regard the data as the sum of two components. One is a deterministic "signal" which presumably obeys some unknown chaotic dynamics. This signal is itself chaotic, yet if the dynamics can be unravelled the signal should be predictable at least for short times. The other component is truly random "noise", which is not predictable at all, even in principle. Our goal for this portion of the study is to assess the relative importance of these two components.

Until we have actually identified which part of the time series is random, it is clearly impossible to subtract off the random component and isolate the deterministic, chaotic part. What we can do, however, is to add random noise to the data. By adding successively larger random components to the data, we obtain new time series with larger random components than the original data. We then analyze these new time series to determine the effect of the added noise on the estimates of the fractal dimension of the attractor.

We have repeated our calculations applying the Grassberger–Procaccia algorithm to "spoiled" data sets obtained by adding varying amounts of random noise to the original data set. The random noise is obtained by generating a set of independent, gaussian distributed random numbers with mean zero and standard deviation equal to some prescribed multiple of the standard deviation of the original data, which is 10.26 (calculated on the basis of the 3-day binned data). For each spoiled data set we apply the correlation algorithm to estimate the apparent dimension of the attractor.

RESULTS

Computer network data

As noted above, the choice was made to sum the data in consecutive 3-day intervals. This increased the length of the data set to 902, while increasing the maximum value of r to 55. In

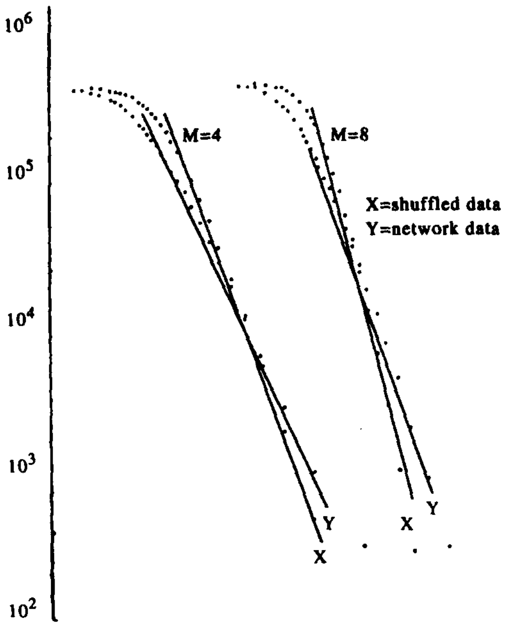

Fig. 1. Log–log plots of the correlation counts $C(r)$ defined in eqn (3) for the network data (Y) and for shuffled data (X). The curves on the right are for an embedding dimension $m=8$, those on the left (which are displaced by one decade in r) are for $m=4$. The slope of the straight-line portion of each graph gives an estimate of the dimension d of the attractor.

addition, it tended to average out any dependence of the data on weekly work schedules. As Fig. 1 illustrates, the log–log plots are straight over slightly less than one decade in r, and we obtain estimates of $d=3.1$ for $m=8$, $d=23$ and for $m=4$.

Theiler (1986) has shown that the Grassberger–Procaccia algorithm can give inaccurate estimates of the attractor when it is used to analyze a finite set of data which has a short-range autocorrelation. To avoid this problem, he recommends a simple modification to the algorithm, namely that in calculating the autocorrelation sums in eqn (3) above, one could exclude all pairs of m-tuples which are too close together in time. Thus, we included a "waiting parameter" w in our calculation, omitting from the correlation counts all pairs of m-tuples which are separated in time by w days or fewer. The dimension estimates, however, were unaffected by changes in the waiting parameter from $w=0$ to $w=4$, so we concluded the data were not plagued by autocorrelation problems. The insensitivity to w is a common feature of all of our calculations, so henceforth we will no longer quote values of w.

Analysis of reordered data

Chaos is a property of a time series of data which are produced by a deterministic rule; this rule must have generated those data in a particular order. Thus it should be possible to test whether a given time series is chaotic or just random by simply reordering the same numbers, then applying the Grassberger–Procaccia algorithmn to the new, rearranged time series. We have done this, using a linear congruential random number generator to generate a random permutation of the original 2707-day time series. We then binned the resulting, shuffled time series into 3-day bins and computed the correlation sums. These are also plotted in Fig. 1 for $m=4$ and $m=8$. The slopes of the correlation plots for the shuffled data are clearly very different, in fact much larger, than the slopes of the plots for the original data. Moreover, the slopes for the shuffled data are much more strongly dependent on the embedding dimension m, as random data would be. Specifically, we estimate $d=3.7$ for $m=4$ and $d=6.5$ for $m=8$. These values do not differ significantly from the result $d=m$ expected for random, non-chaotic data. Thus we conclude that the temporal ordering of the data is critically important to the dimension estimates.

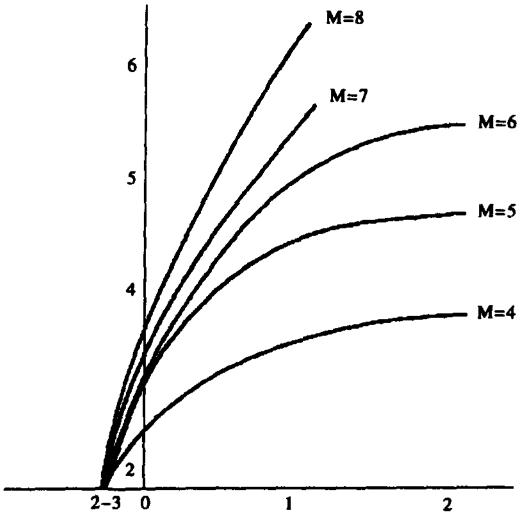

Fig. 2. Plots of the estimates of the attractor dimension vs strength of added noise for network data contaminated with Gaussian random noise. Plots are given for embedding dimensions *m* running from 4 through 8. The curves appear to intersect at a noise strength of − 0.2 to − 0.3 sigma, where sigma is the standard deviation of the original data set.

This suggests strongly that there is a significant component of deterministic chaos in the original time series.

Deterministic chaos and noise

The graph of successive additions of noise versus changes in the calculated dimension of the attractor (*d*) can be seen in Fig. 2. As noted earlier, the dimensions were calculated after adding successively larger amounts of noise, represented in the graph as fractions of the standard deviation of the original data (sigma). The dimension of the attractor can be seen to increase significantly for the first 0.2 sigmas of noise and flatten out and approach *m* for large (> 1.0 sigma) values of noise. It is known that *m*-tuples of random numbers have a fractal dimension of *m*, so each curve should approach its respective *m* value for large amounts of added noise. We have extrapolated to negative values of added noise by fitting an exponential curve to each set of dimension estimates.

Extrapolating from the contaminated data (the equivalent of subtracting the noise), the curves intersect at a point in the range of − 0.2 to − 0.3 sigma. A variation of 0.1 sigma is probably the most accurate estimate of the amount of noise we can achieve using the graph technique. The

Noise *	M=4	M=5	M=6	M=7	M=8
0.00	2.40	2.88	3.08	3.16	3.30
0.20	2.67	3.37	3.41	3.42	4.04
0.40	3.05	3.81	4.24	4.39	4.40
0.60	3.24	4.02	4.35	4.81	5.20
0.80	3.34	4.26	4.79	4.92	5.80
1.00	3.47	4.30	4.94	5.52	6.33
2.00	3.64	4.60	5.36		

Fig. 3. Values of the dimension estimates (*d*) for each embedding dimension (*m*) at successive levels of noise. Successive addition of noise are in increments of the standard deviation of the data set.

curves become extremely steep (probably vertical) prior to intersecting, and it is generally difficult to get more accurate estimates from correlation plots.

What the results indicate, however, is that from 20 to 30% of the variation in the data is the result of noise. Conversely, 70 to 80% of the variation in the data can be accounted for by deterministic chaos. Even allowing for the higher estimate of noise, this indicates a strong deterministic signal.

DISCUSSION

The data analysis indicates strongly that the traffic on CMC discussion groups behaves chaotically, that is, although the time-series data appear to have little periodic behavior, there is some underlying mechanism driving the system. A possibility suggested by this finding is that some type of non-linear dynamic may be a general characteristic of CMC. However, as noted earlier, there is some question concerning the reliability of the findings given the size limitations of the data set. As such, the results can only be viewed as an indication of chaotic behaviour; more generalizable conclusions will require further research with larger data sets and with other discussion groups.

Analysis also demonstrates that human-generated data are not too complex or disorderly, *per se*, for analysis using chaos theory. This is not to assert that all human-generated data behave chaotically or that chaos theory applies equally well to all human-generated data, only that chaos theory should not be excluded as a method of analysis simply because the data may be the product of human communication. The theory underlying the correlation algorithm provides an adequate explanation for why chaos should not be found if no mechanism exists. However, it is easier in some ways simply to formulate a demonstration. More problematic is that the applicability of chaos to human-generated data may be limited due to the comparatively large data sets needed for investigating chaos. It is likely that many human phenomena do not produce sufficiently large data sets for chaos theory to be applicable as a modelling technique even if they are controlled to some degree by simple systems that are sensitive to initial conditions.

Finally, the results indicate that a large part of the variance in CMC time series data may be described by a deterministic chaos mechanism. In particular, the relative importance of the deterministic signal is significantly higher than that of the noise component of the signal.

CONCLUSIONS

It can be seen that the study has addressed only the first three of the four concerns that were outlined in the first section. What remains is to find an approximation of a nonlinear function from the CMC data which can then be tested for its predictive power. The predictive power of systems which behave chaotically is generally limited to short periods. Beyond relatively short time horizons (e.g. 3–7 days for weather systems) the system becomes so dependent on the cumulative effects of small differences in its initial conditions that prediction becomes impractical. However, for many systems there is value in being able to predict behavior for comparatively short times. In the case of CMC, even several days of predictive power is potentially useful in areas such as allocating bandwidth capacity for communication networks.

Current research in the field of deterministic chaos indicates that it may be possible to estimate deterministic mechanisms from time-series data, thus making prediction possible for systems that exhibit chaotic behavior. (Kaplan & Glass, 1992) To date, the researchers have not found a useful approximation of the function that underlies chaotic behavior in the CMC system investigated. Future research to extract the deterministic mechanism from CMC time series data continues.

It may still be problematic to use deterministic chaos for predicting the behavior of CMC systems, given the difficulty of extracting deterministic mechanisms from a time series. However, to the extent that the data are random, there is no predictive power. To the extent that data may be described as chaotic, there is an underlying order which has the possibility of being exploited.

It is also possible that for this and a variety of other systems we may need to revise our definitions of what constitutes prediction. As Kellert (1993) has noted, the idea of prediction as control may simply not apply in the case of chaotic systems, but that there is still predictive power in our ability to understand the range of values that a chaotic system passes through and the reasons for change between values.

REFERENCES

Gleick, J. (1987). *Chaos: Making a New Science.* New York: Viking Penguin.

Grassberger, P. & Procaccia, I. (1983). Measuring the strangeness of strange attractors, *Physica D, 9,* 189–208.

Internet Society. (1993). Gopher://gopher.isoc.org/[internet society]internet. summary.gif.

Kaplan, D. & Glass, L. (1992). *Physics Review Letters 68,* p 427.

Kellert, S. (1993). *In the Wake of Chaos: Unpredictable Order in Dynamical Systems.* Chicago: University of Chicago Press.

Kurtze, D & Snyder, H. (1993). How strongly deterministic is chaotic behavior in computer mediated communication. *Proceedings of the Fourth International Conference on Bibliometrics, Informetrics and Scientometrics.*

Kurtze, D., Snyder, H. & Newby, G. (1992). An investigation of chaos in computer mediated communication. *Informetrics-91,* Sarada Ranganathan Endowment for Library Science, Bangalore, India. 332–342.

Lorenz, E. (1988). Does the flap of a butterfly's wings in Brazil set off a tornado in Texas. Address presented at the *Annual Meeting of the American Association for the Advancement of Science,* Washington DC, December 29, 1979.

Snyder, H. & Kurtze, D. (1992). An examination of the utility of non-linear dynamics techniques for analyzing human information behaviors. *Proceedings of the 55th Annual Meeting of the American Society for Information Science,* Learned Information, Medford, NJ. 98–100.

Theiler, J. (1986). *Physical Review, A34* 2427.