

A Dynamic Model of Traffic on the Web for Analyzing Network Response to Attack

John A. Tomlin
IBM Almaden Research Center
650 Harry Road K53/80-2
San Jose, CA 95120
tomlin@almaden.ibm.com

Abstract

The world wide web (WWW) has rapidly become an important component of the economy, and as such invites malicious manipulation and attack. To analyze the vulnerability of the web in general, and major sites in particular, to such attacks, it is useful to have a dynamic model of traffic on the web—that is a model which reflects the impact on web traffic over time. We develop such a model, generalizing an existing equilibrium (entropy maximizing) model by following an approach previously used in road traffic theory.

1 Introduction

The Internet and its child, the World Wide Web (WWW), have become essential parts of our social infrastructure and the modern economy. Understanding the stability of these vast networks, and their expected behavior in the event of attack, is therefore an important consideration in preparedness to combat cyber-terrorism. We propose an approach to modeling and analyzing the dynamic behavior of the traffic on the WWW, which allows us to study in particular the expected response to suddenly imposed changes, such as those which might be associated with disabling critical servers or links in the network. We focus here on a model built on an existing equilibrium model of WWW traffic ([14]), extending it dynamically as a large system of differential equations, which contain the equilibrium model as a particular case. Experience with the latter indicates that it is computationally feasible. We also briefly review some alternative approaches.

2 Traffic on the World Wide Web

We model the WWW in the now standard way[5] as a graph $G = (V, E)$, where V is the set of pages, corresponding to n vertices or nodes, and E is the set of hyperlinks (henceforth referred to simply as links) corresponding to *directed* edges in the graph, such that if page i has a link to page j then edge (i, j) exists. For simplicity, we will assume here that the entire graph G is strongly connected.

We shall be modeling the behavior of users as what have come to be called “random surfers” - that is web surfers who, at each tick of a notional clock, follow a random out-link (edge) from the page they are currently browsing. It is in

some ways clear that this is not completely representative of actual web surfer behavior. Many surfers have specific “home” or bookmarked pages they wish to access, or click through to one of the results produced by a search engine, but the random surfer model remains dominant, not least in the methods used for static ranking by such search engines ([4, 12]). Opting again for simplicity, we shall therefore adopt it here, recognizing that extensions may have to be made to handle other behaviors.

In our random surfer model let us define y_{ij} to be the number of surfers currently browsing page i who, at each clock-tick, click through to page j . We shall refer to this as the “traffic” following link (i, j) per unit time. Then to satisfy conservation requirements on the flows[8] we need the constraints:

$$(2.1) \quad \sum_{j|(i,j) \in E} y_{ij} - \sum_{j|(j,i) \in E} y_{ji} = 0 \quad (i = 1, \dots, n)$$

$$(2.2) \quad \sum_{i,j} y_{ij} = Y$$

where Y is the total traffic (number of surfers).

The Markov Chain model of random surfer behavior, as used in the well-known “PageRank” model[12], assumes that the probability of a user following an out-link to j from page i is a fixed and known value μ_{ij} , which is usually assumed to be the inverse of the out-degree of i and thus independent of j . The success of this model for the static ranking of web pages in the Google search engine[4] is manifest to all. Such a model implies that the flows y_{ij} satisfy:

$$(2.3) \quad y_{ij} = \mu_{ij} \sum_{(h,i) \in E} y_{hi} \quad \forall (i,j) \in E.$$

Despite its success, the Markov Chain model does suffer from the drawback that the assumption of fixed (usually equal) probabilities of following out-links is somewhat arbitrary, as is the particular solution (2.3) of the conservation equations (2.1). An alternative approach is suggested by work modeling the behavior of road traffic (see [3, 13, 15, 18, 19]). It may be argued on either statistical mechanical or information theoretic grounds (see [10, 14]) that the appropriate y_{ij} are those which maximize the *entropy* of the distribution

of the y_{ij} subject to satisfying the constraints. That is we should

$$(2.4) \quad \text{Maximize} \quad - \sum_{(i,j) \in E} y_{ij} \log y_{ij}$$

The unique maximizing solution is readily seen to be of the form:

$$y_{ij} = Y \exp[-\lambda_0 - \lambda_i + \lambda_j] \quad \forall (i,j) \in E$$

where the λ_i are the Lagrange multipliers.

A more general model is obtained if we allow *a priori* estimates ω_{ij} of the y_{ij} , and cost or benefit values c_{ij} to be associated with the links (i,j) and add the constraint:

$$(2.5) \quad \sum_{(i,j) \in E} c_{ij} y_{ij} = C,$$

where C is the total cost or benefit available. Assigning a Lagrange multiplier β to this constraint, we obtain the solution to this more general form of the model as:

$$y_{ij} = Y \omega_{ij} \exp[-\lambda_0 - \lambda_i + \lambda_j - \beta c_{ij}] \quad \forall (i,j) \in E$$

Writing

$$\alpha_i = e^{-\lambda_i},$$

we see that the solution in terms of the y_{ij} may be written

$$(2.6) \quad Z = \sum_{(i,j) \in E} \omega_{ij} \alpha_i \alpha_j^{-1} e^{-\beta c_{ij}}$$

$$(2.7) \quad y_{ij} = Z^{-1} Y \omega_{ij} \alpha_i \alpha_j^{-1} e^{-\beta c_{ij}} \quad \forall (i,j) \in E$$

where $Z = e^{\lambda_0}$ is the *partition function* for the distribution. This model has already been used successfully as the basis for a new method of ranking web pages[14]. We now apply it in a dynamic context.

3 Dynamic Model of Web Traffic

This model follows the approach of S.G. Tomlin[16, 17], who formulated kinetic and dynamic origin-destination models for road traffic on a bipartite graph. In web terms we introduce *transition coefficients* a_{ijpq} defined as the rate at which surfers will switch from link (i,j) to link (p,q) . Then in an open system the rate of change of the y_{ij} is given by:

$$\frac{dy_{ij}}{dt} = \sum_{(p,q)} (a_{pqij} y_{pq} - a_{ijpq} y_{ij}) + f_{ij}$$

Here dy_{ij}/dt denotes the total rate of change of the y_{ij} from all causes, while f_{ij} denotes the contribution to this change from exogenous sources. When there are no such exogenous influences, we have a (homogeneous) closed system:

$$(3.8) \quad \frac{dy_{ij}}{dt} = \sum_{(p,q)} (a_{pqij} y_{pq} - a_{ijpq} y_{ij})$$

These differential equations can be viewed as forms of the Boltzmann transport equation (see e.g. [1]) and their

solution depends on the form of the coefficients a_{ijpq} . We shall concentrate on the homogeneous system.

Again following the approach in [16, 17] we can initially derive expressions for the a_{ijpq} which result in our previous equilibrium solution, where by definition the dy_{ij}/dt in (3.8) must be zero. It can be shown that if we choose the transition coefficients to be of the special form:

$$a_{ijpq} = \alpha_p \alpha_q^{-1} \omega_{pq} e^{-\beta c_{pq}}$$

(which are independent of the link (i,j)), we obtain the same solution:

$$y_{ij} = Z^{-1} Y \omega_{ij} \alpha_i \alpha_j^{-1} e^{-\beta c_{ij}}$$

as we did for the entropy maximization model in the previous section.

4 General Form of Solution

To examine more general solutions of (3.8) it is convenient to enumerate the links by a single index $k = 1, \dots, N$, where $N = |E|$, so that each k corresponds to a link (i,j) (denoted $k \leftrightarrow (i,j)$), and if $k \leftrightarrow (i,j)$ and $l \leftrightarrow (p,q)$, then if $k < l$ then $i \leq p$, and if $i = p$, then $j < q$.

Corresponding to this numbering, if $k \leftrightarrow (i,j)$ and we denote

$$\begin{aligned} x_k &= y_{ij} \\ u_k &= \omega_{ij} \alpha_i \alpha_j^{-1} e^{-\beta c_{ij}} \end{aligned}$$

then after some algebraic manipulation (3.8) can be rewritten as:

$$\frac{dx_k}{dt} = u_k (\mathbf{e}^T \mathbf{x}) - Z x_k \quad \forall k$$

or in matrix form:

$$(4.9) \quad \frac{d\mathbf{x}}{dt} = (\mathbf{u}\mathbf{e}^T - Z\mathbf{I})\mathbf{x}$$

where $\mathbf{u}^T = (u_1, \dots, u_N)$ and \mathbf{e} is the vector of 1's of conforming dimension. Note in particular that

$$Z = \mathbf{e}^T \mathbf{u}.$$

Following standard methods for simultaneous ordinary differential equations (see [2],[6]), we look for a fundamental matrix $\Phi = \Phi(t)$ that satisfies

$$\frac{d\Phi(t)}{dt} = (\mathbf{u}\mathbf{e}^T - Z\mathbf{I})\Phi(t), \quad \Phi(0) = \mathbf{I}$$

When \mathbf{u} is constant, this is obtained by observing that the eigensystem of the rank-one matrix $\mathbf{u}\mathbf{e}^T$ may be derived from the identity

$$(\mathbf{u}\mathbf{e}^T)\mathbf{V} = \mathbf{V}\mathbf{J}$$

where, defining $\bar{\mathbf{u}}^T = (u_2, \dots, u_N)$:

$$\mathbf{V} = \left(\begin{array}{c|c} u_1 & -\mathbf{e}^T \\ \hline \bar{\mathbf{u}} & \mathbf{I} \end{array} \right),$$

and

$$\mathbf{J} = \left(\begin{array}{c|c} Z & \mathbf{0}^T \\ \hline \mathbf{0} & \mathbf{O} \end{array} \right)$$

and that the eigenvalues of $(\mathbf{u}\mathbf{e}^T - Z\mathbf{I})$ are then simply shifted by Z , so that

$$(\mathbf{u}\mathbf{e}^T - Z\mathbf{I})\mathbf{V} = \mathbf{V}\hat{\mathbf{J}}$$

with \mathbf{V} as above, and

$$\hat{\mathbf{J}} = \begin{pmatrix} 0 & & & \\ & -Z & & \\ & & \ddots & \\ & & & -Z \end{pmatrix}$$

It is straightforward to verify that \mathbf{V}^{-1} exists and is given by

$$\mathbf{V}^{-1} = \left(\begin{array}{c|c} Z^{-1} & Z^{-1}\mathbf{e}^T \\ \hline -Z^{-1}\bar{\mathbf{u}} & \mathbf{I} - Z^{-1}\bar{\mathbf{u}}\mathbf{e}^T \end{array} \right)$$

The fundamental matrix is now given by

$$(4.10) \quad \Phi = e^{(\mathbf{u}\mathbf{e}^T - Z\mathbf{I})t} = \mathbf{V}e^{\hat{\mathbf{J}}t}\mathbf{V}^{-1}$$

where

$$e^{\hat{\mathbf{J}}t} = \begin{pmatrix} 1 & & & \\ & e^{-Zt} & & \\ & & \ddots & \\ & & & e^{-Zt} \end{pmatrix}.$$

Thus if we are given an initial condition $\mathbf{x}(0) = \mathbf{x}^0$ the solution at time t is given by

$$(4.11) \quad \mathbf{x}(t) = \Phi\mathbf{x}^0$$

for the fundamental matrix (4.10).

5 Relaxation Time of the System

Armed with the above machinery, we are now in a position to examine the behavior of solutions over time. Suppose in particular that we have an equilibrium solution for the web traffic, and that a disturbance is induced—say the disabling of a major host or site. Then we may model this by changing the ω_{ij} to very small values for the links into (and out of) the host, and/or by assigning very large costs c_{ij} to those links. The modified model will have a different equilibrium solution, and the behavior of the system will be modeled by (4.11) where \mathbf{x}^0 is the old (undisturbed) equilibrium solution, and the fundamental matrix corresponds to the new (disturbed) model. The evolution of traffic over time after the disturbance can be obtained from:

$$(5.12) \quad \mathbf{x}(t) = \Phi\mathbf{x}^0 = \mathbf{V}e^{\hat{\mathbf{J}}t}\mathbf{V}^{-1}\mathbf{x}^0$$

where the \mathbf{V} , \mathbf{V}^{-1} and $\hat{\mathbf{J}}$ are now defined in terms of the \mathbf{u} corresponding to the final equilibrium state \mathbf{x}^f of the disturbed system. Letting $\bar{\mathbf{x}}^0 = (x_2^0, \dots, x_n^0)^T$, and noting that $Y = \mathbf{e}^T\mathbf{x}^0 = \mathbf{e}^T\mathbf{x}^f$, we obtain:

$$\mathbf{x}(t) = \mathbf{V}e^{\hat{\mathbf{J}}t}\mathbf{V}^{-1}\mathbf{x}^0$$

$$\begin{aligned} &= \mathbf{V}e^{\hat{\mathbf{J}}t} \left(\frac{Z^{-1}Y}{\bar{\mathbf{x}}^0 - Z^{-1}Y\bar{\mathbf{u}}} \right) \\ &= \left(\frac{Z^{-1}Y\mathbf{u}_1 - e^{-Zt}(\mathbf{e}^T\bar{\mathbf{x}}^0 - Z^{-1}Y\mathbf{e}^T\bar{\mathbf{u}})}{Z^{-1}Y\bar{\mathbf{u}} + e^{-Zt}(\bar{\mathbf{x}}^0 - Z^{-1}Y\bar{\mathbf{u}})} \right) \\ &= Z^{-1}Y\bar{\mathbf{u}} + e^{-Zt} \left(\frac{x_1^0 - Z^{-1}Y\mathbf{u}_1}{\bar{\mathbf{x}}^0 - Z^{-1}Y\bar{\mathbf{u}}} \right) \\ &= \mathbf{x}^f + e^{-Zt}(\mathbf{x}^0 - \mathbf{x}^f) \end{aligned}$$

which demonstrates that the solution will deform exponentially from the initial solution to the final, with relaxation time $1/Z$ - the inverse of the (final) value of the partition function. Conversely, we may obtain a picture of the return of traffic to “normal” by letting \mathbf{x}^0 be our estimate of the perturbed traffic after a particular attack scenario, and then using the above method, with the equilibrium Φ , to compute the recovery path $\mathbf{x}(t)$ to the equilibrium state (now represented by \mathbf{x}^f).

Since the value of the partition function is clearly of great importance it is useful to form an idea of its order of magnitude. In the event that we choose the uniform values $\omega_{ij} = 1/|E|$, assume that (2.5) is absent, and the graph is complete (or at least regular), so that we may assume the α_i are uniform, then it is easy to see from (2.6) that $Z = 1$ in this case, and the y_{ij} are uniformly of the value $Y/|E|$. In practice, with a 19 million node IBM intranet graph, and a 178 million node subset of the WWW, the experiments reported in [14] resulted in values which were of the order of $1/2$ to $1/4$.

6 Alternate Approaches

While the approach we have outlined above leads to an elegant description of dynamic behavior, there are potential alternatives. One obvious such alternative is to take the solution implied by the PageRank assumption (2.3), that is:

$$(6.13) \quad \frac{dy_{ij}}{dt} = \mu_{ij} \sum_{(h,i) \in E} y_{hi} - y_{ij} \quad \forall (i,j) \in E$$

A major practical disadvantage here is that instead of a system of dimension $|E|$ which reduces to a shifted rank one system (4.9), the system (6.13) is in general a shifted system of rank n (the number of pages), which greatly complicates the derivation, not to mention the computation, of the fundamental matrix. However, given the popularity of the Markov Chain model it would be interesting to pursue it further.

Another alternative approach is to look, not at migrations of surfers from one link to another, but of the probability mass of surfers at one page to another. Such an approach is reminiscent of the “compartmental models” [9] used in other disciplines, but is beyond the scope of this paper.

7 Further Work and Conclusion

The dynamic model presented here provides a framework for the investigation of vulnerability to attack, but clearly requires considerable computational resources. Each scenario requires the solution of the appropriate maximum entropy model, in addition to the “baseline” solution. We suggest, however, that this is not as onerous as it sounds if the level of granularity is chosen correctly. It would seem most appropriate to work at the level of the “host graph” [7], since we are mainly concerned with attacks on hosts rather than on individual pages on the host. While the web graph available to us has over a billion nodes (pages), the number of hosts is less formidable—currently of the order of 20 million [7]. While this number of nodes is much reduced, the number of links between hosts is still large—the corresponding number here is about 1.5 billion—but still more manageable, especially since we may modify the formulation to specifically handle the large number of parallel links between hosts. If the number of parallel links between hosts i and j is m_{ij} , then (if their ω and c values are the same) we may aggregate them as a single link, and modify the objective function (2.4) to be:

$$\text{Maximize } \sum_{(i,j) \in E} (\log m_{ij} - \log y_{ij}) y_{ij}$$

The algorithm in [14] is easily modified to handle the linear term. Preliminary experiments are currently being undertaken on this host graph data set.

Finally it should be noted that our model has concentrated entirely on the link structure of the WWW, and that the random surfer “traffic” makes up only a fraction of the total internet traffic (as measured by actual packets transmitted between hosts), which has its own traffic models (see e.g. Kelly [11]). It would be most interesting to map the dynamic model we have presented on to this underlying internet, and to examine the security aspects of this interaction.

8 Acknowledgements

I am indebted to Kevin S. McCurley and my other colleagues at IBM Research for several helpful discussions in the course of this work, and most especially to my late father, S.G. Tomlin, who pioneered this approach.

References

- [1] R. Balescu, *Equilibrium and Nonequilibrium Statistical Mechanics*, Wiley, NY (1975).
- [2] R. Bellman, *Stability Theory of Differential Equations*, Dover Edition (1969).
- [3] D.E. Boyce, L.J. Leblanc, K.S. Chon, “Network Equilibrium Models of Urban Location and Travel Choices: A Retrospective Survey.” *Journal of Regional Science* **28**, 159-183 (1988).
- [4] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Proc. of WWW7, Brisbane, Australia, June 1998. See: <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>

- [5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, “Graph Structure in the Web”, Proc. WWW9 conference, 309–320, May 2000. See also: <http://www9.org/w9cdrom/160/160.html>
- [6] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, NY (1955).
- [7] N. Eiron, K.S. McCurley and J.A. Tomlin, “Ranking the Web Frontier”, To appear in Proc. of WWW2004, New York, NY, May 2004.
- [8] L.R. Ford, Jr. and D.R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, NJ, (1962).
- [9] K. Godfrey, *Compartmental Models and their Application*, Academic Press, London (1983).
- [10] E. Jaynes, “Information Theory and Statistical Mechanics”, *Physical Review* **106**, 620–630 (1957).
- [11] F.P. Kelly, “Mathematics of the Internet”, in *Mathematics Unlimited - 2001 and Beyond* (Editors B. Engquist and W. Schmid), 685–702. Springer-Verlag, Berlin, (2001).
- [12] L. Page, S. Brin, R. Motwani and T. Winograd “The PageRank Citation Ranking: Bringing Order to the Web”, Stanford Digital Library working paper SIDL-WP-1999-0120 (version of 11/11/1999). See: <http://www.diglib.stanford.edu/cgi-bin/get/SIDL-WP-1999-0120>
- [13] R.B. Potts and R.M. Oliver, *Flows in Transportation Networks*, Academic Press, New York (1972).
- [14] J.A. Tomlin, “A New Paradigm for Ranking Pages on the World Wide Web”, Proc. World Wide Web Conference 2003 (WWW2003), 350–355, Budapest, May 2003. See also: <http://www2003.org/cdrom/papers/refereed/p042/paper42.html/p42-tomlin.htm>
- [15] J.A. Tomlin and S.G. Tomlin, “Traffic Distribution and Entropy”, *Nature* **220**, 974-976 (1968).
- [16] S.G. Tomlin, “A Kinetic Theory of Traffic Distribution and Similar Problems”, *Environment and Planning* **1**, 221-227, (1969).
- [17] S.G. Tomlin, “Time-Dependent Traffic Distributions”, *Transportation Research* **4**, 77-86 (1970).
- [18] A.G. Wilson, “Notes on Some Concepts in Social Physics”, Regional Science Association: Papers, XXII, Budapest Conference, 1968.
- [19] A.G. Wilson, *Entropy in Urban and Regional Modeling*, Pion Press, London (1970).