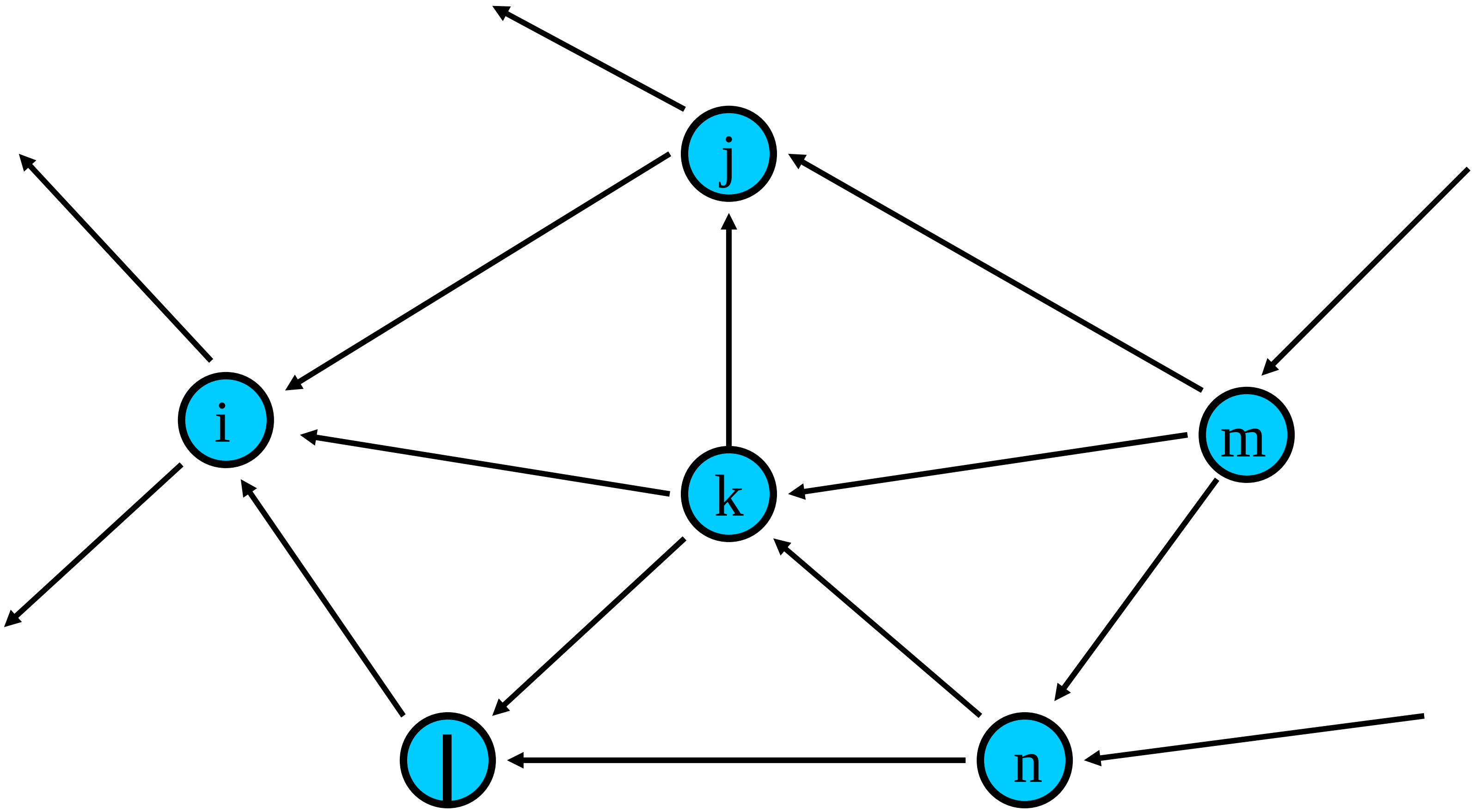
# A Dynamic Model of Traffic on the Web for Analyzing Network Response to Attack

John Tomlin
IBM Almaden Reseach Center
San Jose, CA 95120
tomlin@almaden.ibm.com

# Motivation

The WWW traffic makes up a large fraction of traffic on the internet, which has become a significant part of the national (and international) economy and infrastructure.

An understanding of how the WWW responds to imposed changes, such as an attack, is therefore of significant economic/national security interest.



Graph model of the web:

$$G = (V,E)$$

Where *V* is set of vertices (*i*,*j*,*k*,…..) or nodes or pages
And *E* is the set of edges (*i*,*j*)

# The “Random Surfer”

We postulate a notional “clock”.

At each tick of this clock, each web surfer clicks on One of the outlinks from the page he/she is browsing.

For simplicity let us initially assume that G is *strongly connected* (i.e. there is a directed path between every pair of pages in the set *V*)

# The “Markov Random Surfer”

Let

$\lambda_{ij}$ = probability that a surfer at page i will click through to page j

(It is often assumed these probabilities are uniform, i.e. if

$d_i$ = out degree (number of outlinks) of page i,

then

$\lambda_{ij} = 1/d_i$ for all $j$ such that $(i,j)$ $\bar{E}$

This is the usual assumption in computing PageRank)

$M = [\lambda_{ij}]$ defines a Markov chain. The stochastic vector $w = <w_1, \ldots, w_m>$, where $w_1 + \ldots. + w_m = 1$ is a *stationary state* of the Markov chain if

$$w^T = w^T M$$

# Beyond Markov Chains
# - A Network Flow Model

Let us maintain the graph *G* as our basic model, and the assumption that each user clicks through to a page at each tick of the “clock”, but define
flow variables:

$y_{ij}$ = flow per unit time from page *i* to page *j*.

Where $$\sum_{(i,j) \in E} y_{ij} = Y \ (const) \qquad (*)$$

(We will usually find it convenient to work with the normalized values $p_{ij} = y_{ij} / Y$ (probabilities) )

The flows must satisfy conservation equations (Kirchoff Conditions)

$$\sum_{(h,i)\ ^{-}\ E} y_{hi} \ \ - \ \ \sum_{(i,j)\ ^{-}\ E} y_{ij} = 0 \qquad\qquad i \quad ^{-}\mathrm{V}$$

as well as

$$y_{ij} \geq 0$$

and (*).

The $y_{ij}$ can take any values which satisfy these constraints. What should we estimate them to be?

The PageRank (Markov) assumption is that:

$$y_{ij} = d_i^{-1} \sum_{(h,i) \; ^{-} \; E} y_{hi} \quad \text{for all } (i,j) \quad ^{-} \; E$$

(Where $d_i$ is the out-degree of page *i*)

or more generally:

$$y_{ij} = \lambda_{ij} \sum_{(h,i) \; ^{-} \; E} y_{hi} \quad \text{for all } (i,j) \quad ^{-} \; E$$

Is is easy to see, by direct substitution, that these flows, suitably scaled to satisfy (*), satisfy the conservation equations.

Now, using normalized values (probabilities), so that:

$$\sum_{(h,i)\ \text{¯}\ E} p_{hi} \;-\; \sum_{(i,j)\ \text{¯}\ E} p_{ij} = 0 \qquad i \ \text{¯}\mathrm{V}$$

$$\sum_{(i,j)\ \text{¯}\ E} p_{\mathrm{ij}} = \mathit{1}$$

$$p_{ij} \geq 0$$

where $p_{ij} = y_{ij} / Y$

The $p_{ij}$ form a probability distribution we wish to estimate. What should we estimate them to be?

# Max Entropy Solution

Both statistical mechanics and information theory tell us that the correct estimate (given only knowledge of the network topology) is the solution of the maximum entropy problem:

Maximize $$-\sum_{(i,j)\ \bar{E}} p_{ij} \ln p_{ij}$$

subject to the constraints.

The solution is of the form:

$$p_{ij} = \exp[\, - \kappa_0 - \kappa_i + \kappa_j \,] \qquad \text{for } (i,j) \quad \bar{E}$$

and if we define:

$$\mathbf{Z} = e^{\kappa_0} = \sum_{(i,j)\quad \overline{E}} \exp\{- \kappa_i + \kappa_j \}$$

and also define

$$\alpha_i = e^{-\kappa_\mathrm{i}}$$

then

$$p_{ij} = \exp[ - \kappa_0 - \kappa_i + \kappa_j ] \qquad \text{for } (i,j) \quad \overline{E}$$

may be written as:

$$p_{ij} = \mathrm{Z}^{-1} \alpha_i \alpha_j^{-1} \qquad \text{for } (i,j) \quad \overline{E}$$

Letting $P = [p_{ij}]$ , $\text{✌} = \mathrm{diag}(\alpha_1 , \ldots, \alpha_m)$

and

$$c_{ij} = \begin{cases} \mathrm{Z}^{-1} & \text{for } (i,j) \quad \bar{E} \\ 0 & \text{otherwise,} \end{cases}$$

then

$$P = \text{✌} C \text{✌}^{-1}$$

subject to

$$\sum_{(h.i) \, - \, E} p_{hi} - \sum_{(i,j) \, - \, E} p_{ij} = 0 \qquad i \quad \bar{\mathrm{V}}$$

$$\sum_{(i,j) \, - \, E} p_{ij} = 1$$

This is a matrix balancing (iterative scaling) problem, which can be solved relatively efficiently.

General idea:

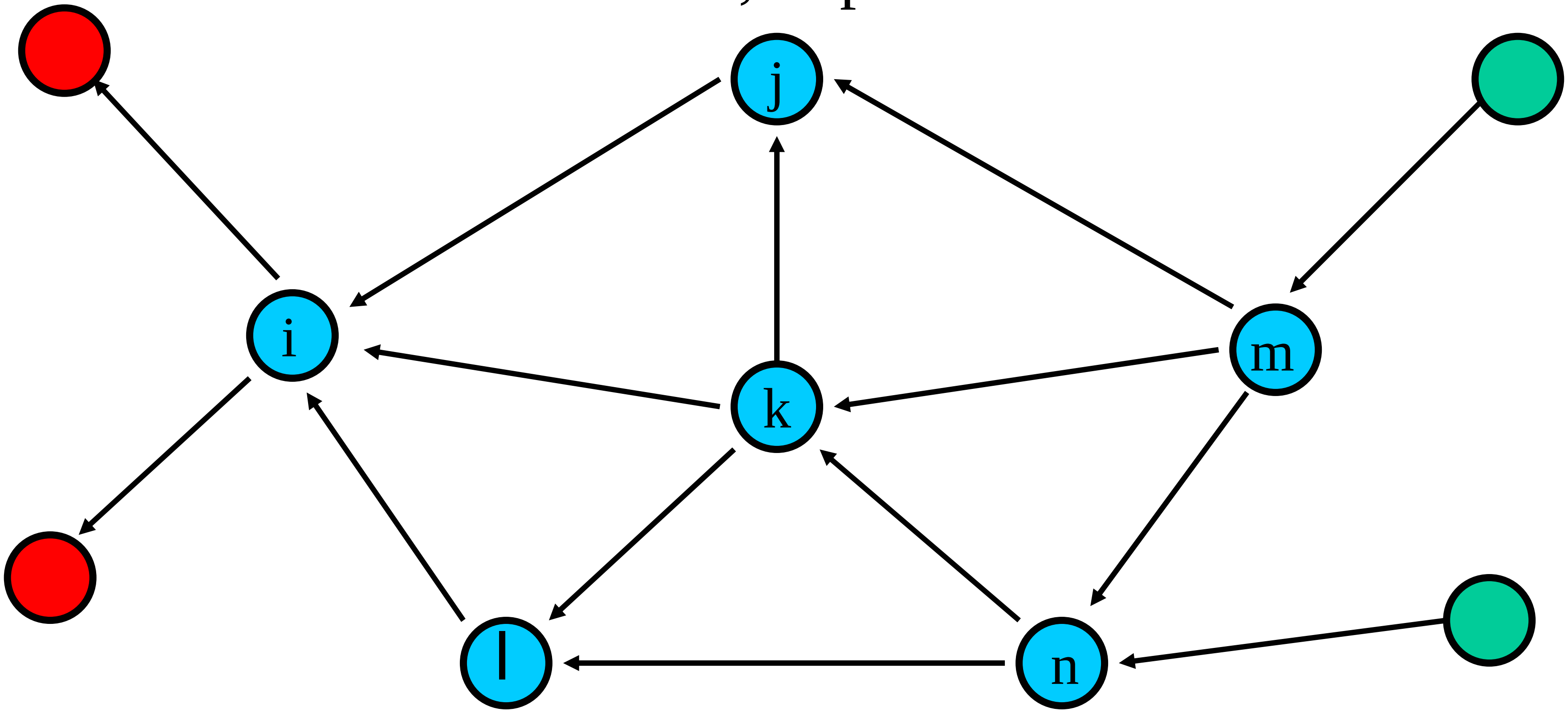
0. Guess the value of $\mathbf{Z}^{-1}$

   Start with initial values for the $\alpha_i$ (e.g. 1) , denoted $\alpha_i^{(0)}$, and let $p_{ij}^{(k)} = \mathbf{Z}^{-1} \alpha_i^{(k)} / \alpha_j^{(k)}$. At each iteration:

1. Compute $\}_i^{(k)} = \sum_j p_{ij}^{(k)}$ , $\rho_i^{(k)} = \sum_j p_{ji}^{(k)}$
2. Let $\gamma_i^{(k)} = (\rho_i^{(k)} / \}_i^{(k)})^{1/2}$
3. Update $\alpha_i^{(k+1)}$ ♌ $\gamma_i^{(k)} \alpha_i^{(k)}$ , for some or all *i*
4. Stop if $1 - \varepsilon$ ☯ $\gamma_i^{(k)}$ ☯ $1+ \varepsilon$ , else step *k* and go to 1.
5. Check if sum of the final $p_{ij}$ is 1.0. If not, adjust **Z** and go to 1.

Note the work per inner iteration is about twice an iteration of power iteration (or Gauss-Seidel).

# However, in practice:



Many pages have no in-links or no out-links.
The “random surfer” can never reach the green pages
or escape from the red pages. G is not strongly connected

# A Modified Network Formulation

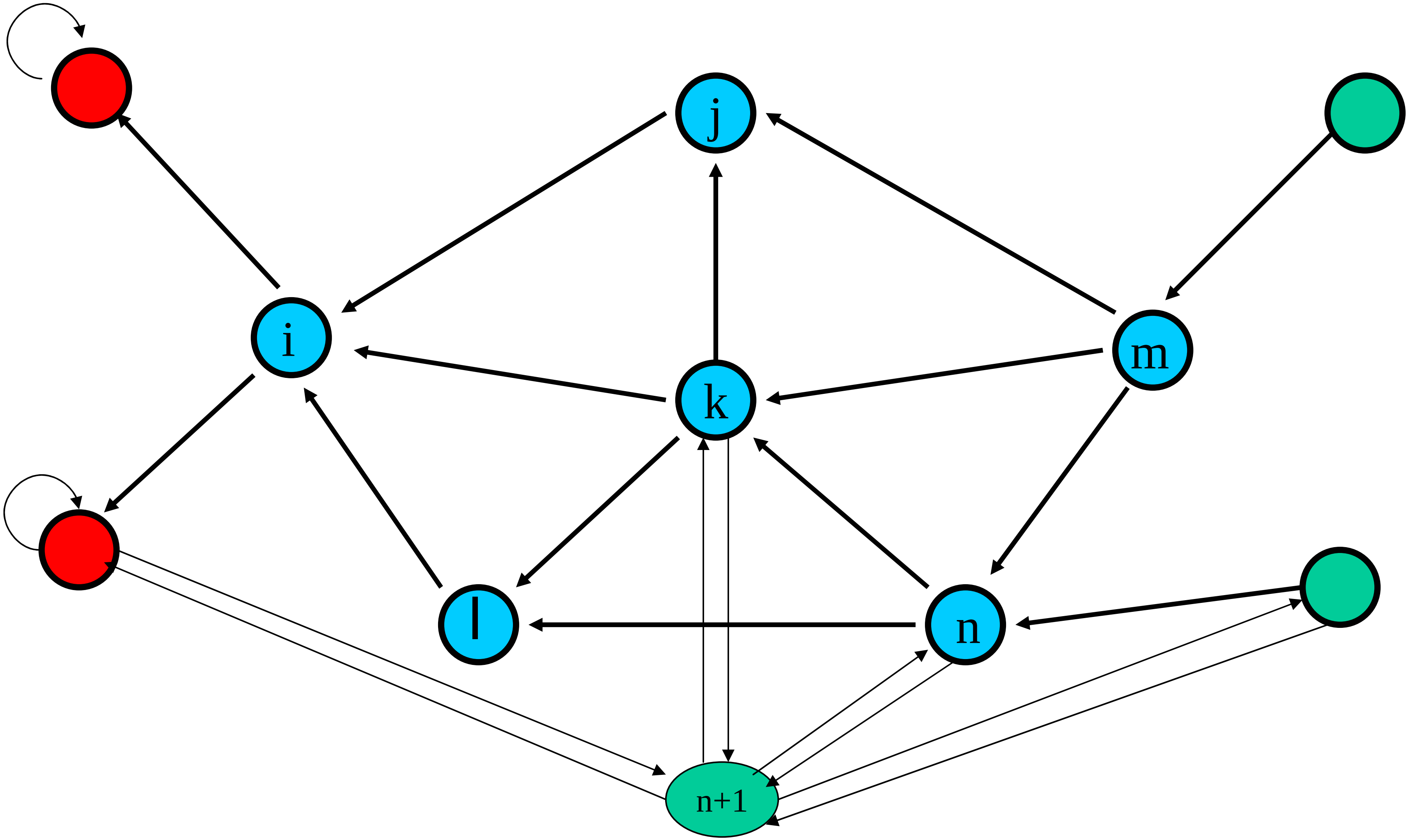
With  with frequency (*1* – α), let surfers make a random jump.

Define an additional node *n+1* and add it to *V* to get *V'*.
We then construct edges from every node in *V* to *n+1* and from *n+1* to every other node.

The total traffic through this node is required to be

$$\sum_{i} p_{i,n+1} = (1 - \alpha) = \sum_{j} p_{n+1,j}$$

The new edge set *E'* is *E* plus the new edges and also a self loop at each dead-end page.



(All nodes should be linked to and from n+1 but all such links are not drawn)

The extended model is now:

Maximize $$S = -\sum_{(i,j) \in E'} p_{ij} \ln p_{ij}$$

Subject to:

$$\sum_{(h.i) \in E'} p_{hi} - \sum_{(i,j) \in E'} p_{ij} = 0 \qquad i \in V$$

$$\sum_{i} p_{i,n+1} = (1 - \alpha)$$

$$\sum_{j} p_{n+1,j} = (1 - \alpha)$$

$$\sum_{(i,j) \in E'} p_{ij} = 1$$

The extended model still corresponds to a (hybrid) matrix balancing problem, but with some row and column sums required to take particular values – a simple extension.

Note that if we know the flow through any node we can write down the equations, just as we did for the “artificial” node, and treat them in the same way. In other words we can reduce the uncertainty in the model by applying additional information.

The solution algorithm requires only minor modification, and produces not only the “traffic” primal variables ($p_{ij}$) , but as an essential by-product the exponentials of the Lagrange multipliers:

$$\alpha_i = e^{-\kappa_i}$$

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

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

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

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

# A Kinetic/Dynamic Model

We now consider how the flows $y_{ij}$ (or the $p_{ij}$) might change over time.

We introduce *transition coefficients* $a_{ijpq}$ defined as the rate at which (random, not individual) 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:

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

Where $dy_{ij}/dt$ 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 the $f_{ij}$ are zero we have a *closed* system (which we assume from now on).

These are forms of the Boltzmann *transport equations*.

# Form of the $a_{ijpq}$

The model is determined by the form of the $a_{ijpq}$ coefficients. It is often useful to separate these into two parts – an “escape” rate $\varepsilon_{ij}$ and a “capture” rate $\varphi_{pq}$ , so that :

$$a_{ijpq} = \varepsilon_{ij}\, \varphi_{pq} ,$$

In what follows we will be assuming $\varepsilon_{ij} = 1$

# Equilibrium Solution

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 above.

# General Form of Solution

To examine more general solutions it is convenient to enumerate the links by a single index $k = 1, ..., N$, 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 the DE's can be rewritten as:

$$\frac{dx_k}{dt} = u_k(\mathbf{e}^T\mathbf{x}) - Zx_k \qquad \forall k$$

or in matrix form:

$$\frac{d\mathbf{x}}{dt} = (\mathbf{u}\mathbf{e}^T - Z\mathbf{I})\mathbf{x} \tag{1}$$

where $\mathbf{u}^T = (u_1, ..., u_N)$ 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, we look for a fundamental matrix $\mathbf{\Phi} = \mathbf{\Phi}(\mathrm{t})$ that satisfies

$$\frac{d\mathbf{\Phi}(t)}{dt} = (\mathbf{u}\mathbf{e}^T - Z\mathbf{I})\mathbf{\Phi}(t), \qquad \mathbf{\Phi}(0) = \mathbf{I} \quad (1)$$

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

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

Defining $\bar{\mathbf{u}}^T = (u_2, ..., u_N)$:

$$\mathbf{V} = \left(\begin{array}{c|c} u_1 & -\mathbf{e}^T \\ \hline \bar{\mathbf{u}} & \mathbf{I} \end{array}\right), \tag{1}$$

and

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

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

$$(\mathbf{u}\mathbf{e}^T - Z\mathbf{I})\mathbf{V} = \mathbf{V}\hat{\mathbf{J}} \tag{2}$$

with $\mathbf{V}$ as above, and

$$\hat{\mathbf{J}} = \begin{pmatrix} 0 & & & \\ & -Z & & \\ & & \ddots & \\ & & & -Z \end{pmatrix} \tag{3}$$

The fundamental matrix is now given by

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

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

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

Suppose 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. The evolution of traffic over time from an initial state $\mathbf{x}^0$ to a final state $\mathbf{x}^f$ after the disturbance can be obtained from:

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

and can be shown to be:

$$\mathbf{x}(t) = \mathbf{x}^f + e^{-Zt}(\mathbf{x}^0 - \mathbf{x}^f)$$

*The critical quantity is the final value of the partition function $Z$*

# Computational results

1. IBM Intranet crawls (made in 2002)
   (a) 19 million pages , ~200 million links
   (b) 17 million pages.
   Z ~ 0.5

2. Partial internet crawl (made in 2001).
   173 million pages (constraints)
   2 billion links (variables)
   Z ~ 0.25

3. Host Graph (from 2003)
   ~20 million pages
   ~1.1 billion links
   Z ~ 0.65

# Additional possible approaches

One alternative model is to take the solution implied by the PageRank assumption, i.e.

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

Unfortunately, instead of a system of dimension $|E|$ which reduces to a shifted rank one system this leads to a shifted system of rank $n$ (the number of pages).

Another alternative approach is to look at "compartmental models" as used in other disciplines.

# Future Research

1. Accelerated methods for the  matrix balancing problem
2. Massively parallel implementation (on Blue Gene/L ?)
3. Further work on forms of the transition coefficients
4. Mapping the changes in web traffic onto the underlying internet (or vice-versa)

# Generalized Equilibrium Model

Maximize $$S = \sum_{(i,j) \in E'} \{\zeta_{ij}\, p_{ij} - p_{ij} \ln p_{ij}\} = -\sum_{i,j} p_{ij} \ln (p_{ij} / \zeta_{ij})$$

Subject to:

$$\sum_{(h.i) \in E'} p_{hi} - \sum_{(i,j) \in E'} p_{ij} = 0 \qquad i \in U$$

$$\sum_{(h.i) \in E'} p_{hi} = H_i \qquad i \in V\text{-}U$$

$$\sum_{(i,j) \in E'} p_{ij} = H_i$$

$$\sum_{(i,j) \in E'} c_{ij}\, p_{ij} = C$$

$$\sum_{(i,j) \in E'} p_{ij} = 1$$

I