The Giving Game: Google Page Rank

University of Utah Teachers' Math Circle

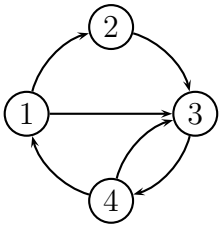
Nick Korevaar

March 24, 2009

Stage 1: The Game

Imagine a game in which you repeatedly distribute something desirable to your friends, according to a fixed template. For example, maybe you're giving away "play–doh" or pennies! (Or it could be you're a web site, and you're voting for the sites you link to. Or maybe, you're a football team, and you're voting for yourself, along with any teams that have beaten you.)

Let's play a small–sized game. Maybe there are four friends in your group, and at each stage you split your material into equal sized lumps, and pass it along to your friends, according to this template:



The question at the heart of the basic Google page rank algorithm is: in a voting game like this, with billions of linked web sites and some initial vote distribution, does the way the votes are distributed settle down in the limit? If so, sites with more limiting votes must ultimately be receiving a lot of votes, so must be considered important by a lot of sites, or at least by sites which themselves are receiving a lot of votes. Let's play!

1. Decide on your initial material allocations. I recommend giving it all to one person at the start, even though that doesn't seem fair. If you're using pennies, 33 is a nice number for this template. At each stage, split your current amount into equal portions and distribute it to your friends, according to the template above. If you have remainder pennies, distribute them randomly. Play the game many (20?) times, and see what ultimately happens to the amounts of material each person controls. Compare results from different groups, with different initial allocations.

2. While you're playing the giving game, figure out a way to model and explain this process algebraically!

Stage 2: Modeling the game algebraically

The game we just played is an example of a *discrete dynamical system*, with constant *transition matrix*. Let the initial fraction of play dough distributed to the four players be given by

$$\mathbf{x}_0 = \begin{bmatrix} x_{0,1} \\ x_{0,2} \\ x_{0,3} \\ x_{0,4} \end{bmatrix}, \quad \sum_{i=1}^{4} x_{0,i} = 1$$

Then for our game template on page 1, we get the fractions at later stages by

$$\begin{bmatrix} x_{k+1,1} \\ x_{k+1,2} \\ x_{k+1,3} \\ x_{k+1,4} \end{bmatrix} = x_{k,1} \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \\ 0 \end{bmatrix} + x_{k,2} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + x_{k,3} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} + x_{k,4} \begin{bmatrix} 0.5 \\ 0 \\ 0.5 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} x_{k+1,1} \\ x_{k+1,2} \\ x_{k+1,3} \\ x_{k+1,4} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ x_{k,3} \\ x_{k,4} \end{bmatrix}$$

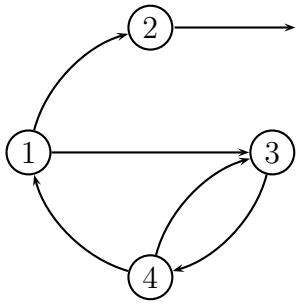
So in matrix form, $\mathbf{x}_k = A^k \mathbf{x}_0$ for the transition matrix A given above.

3. Compute a large power of A. What do you notice, and how is this related to the page 1 experiment?

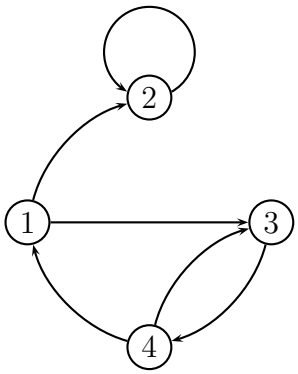
4. The limiting "fractions" in this problem really are fractions (and not irrational numbers). What are they? Is there a matrix equation you could solve to find them, for this small problem? Hint: the limiting fractions should remain fixed when you play the game.

5. Not all giving games have happy endings. What happens for the following templates?

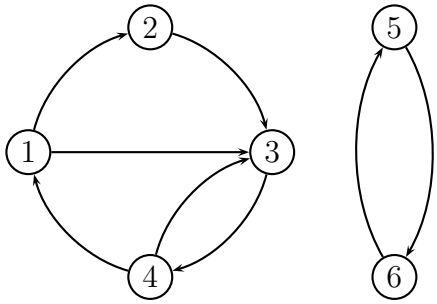
   (a)



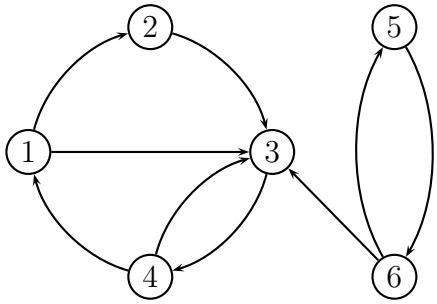
(b)



(c)



(d)



Here's what separates good giving–game templates, like the page 1 example, from the bad examples 5a,b,c,d.

Definition: A square matrix S is called *stochastic* if all its entries are positive, and the entries in each column add up to exactly one.

Definition: A square matrix A is *almost stochastic* if all its entries are non–negative, the entries in each column add up to one, and if there is a positive power k so that A^k is stochastic.

6. What do these definitions mean *vis-à-vis* play–doh distribution? Hint: if it all starts at position j, then the inital fraction vector $\mathbf{x}_0 = \mathbf{e_j}$, i.e. has a 1 in position j and zeroes elsewhere. After k steps, the material is distributed according to $A^k\mathbf{e_j}$, which is the j^{th} column of A^k.

Stage 3: Theoretical basis for Google page rank

Theorem. *(Perron–Frobenius) Let A be almost stochastic. Let $\mathbf{x}_0$ be any "fraction vector" i.e. all its entries are non–negative and their sum is one. Then the discrete dynamical system*

$$\mathbf{x}_k = A^k \mathbf{x}_0$$

has a unique limiting fraction vector $\mathbf{z}$, and each entry of $\mathbf{z}$ is positive. Furthermore, the matrix powers A^k converge to a limit matrix, each of whose columns are equal to $\mathbf{z}$.

proof: Let $A = [a_{ij}]$ be almost stochastic. We know, by "conservation of play–doh", that if $\mathbf{v}$ is a fraction vector, then so is $A\mathbf{v}$. As a warm–up for the full proof of the P.F. theorem, let's check this fact algebraically:

$$\begin{aligned}\sum_{i=1}^{n}(A\mathbf{v})_i &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}v_j = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}v_j \\ &= \sum_{j=1}^{n} v_j \left(\sum_{i=1}^{n} a_{ij}\right) = \sum_{j=1}^{n} v_j = 1\end{aligned}$$

Thus as long as $\mathbf{x}_0$ is a fraction vector, so is each iterate $A^N\mathbf{x}_0$.

Since A is almost stochastic, there is a power l so that $S = A^l$ is stochastic. For any (large) N, write $N = kl + r$, where $N/l = k$ with remainder r, $0 \le r < l$. Then

$$A^N\mathbf{x}_0 = A^{kl+r}\mathbf{x}_0 = \left(A^l\right)^k A^r\mathbf{x}_0 = S^k A^r\mathbf{x}_0$$

As $N \to \infty$ so does k, and there are only l choices for $A^r\mathbf{x}_0$, $0 \le r \le l-1$. Thus if we prove the P.F. theorem for stochastic matrices S, i.e. $S^k\mathbf{y}_0$ has a unique limit independent of $\mathbf{y}_0$, then the more general result for almost stochastic A follows.

So let $S = [s_{ij}]$ be an $n \times n$ stochastic matrix, with each $s_{ij} \ge \varepsilon > 0$. Let 1 be the matrix for which each entry is 1. Then we may write:

$$B = S - \varepsilon 1; \quad S = B + \varepsilon 1. \tag{1}$$

Here $B = [b_{ij}]$ has non–negative entries, and each column of B sums to

$$1 - n\varepsilon := \mu < 1. \tag{2}$$

We prove the P.F. theorem in a way which reflects your page 1 experiment: we'll show that whenever $\mathbf{v}$ and $\mathbf{w}$ are fraction vectors, then $S\mathbf{v}$ and $S\mathbf{w}$ are geometrically closer to each other than were $\mathbf{v}$ and $\mathbf{w}$. Precisely, our "metric" for measuring the distance "d" between two fraction vectors is

$$\mathrm{d}(\mathbf{v},\mathbf{w}) := \sum_{i=1}^{n} |v_i - w_i|. \tag{3}$$

Here's the magic: if $\mathbf{v}$ is any fraction vector, then for the matrix 1, of ones,

$$(1\mathbf{v})_i = \sum_{j=1}^{n} 1v_j = 1.$$

So if $\mathbf{v}, \mathbf{w}$ are both fraction vectors, then $1\mathbf{v} = 1\mathbf{w}$. Using matrix and vector algebra, we compute using equations (1), (2):

$$\begin{aligned} S\mathbf{v} - S\mathbf{w} &= (B + \varepsilon 1)\mathbf{v} - (B + \varepsilon 1)\mathbf{w} \\ &= B(\mathbf{v} - \mathbf{w}) \end{aligned} \tag{4}$$

So by equation (3),

$$\begin{aligned} \mathrm{d}(S\mathbf{v}, S\mathbf{w}) &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} b_{ij}(v_j - w_j) \right| \\ &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} |v_j - w_j| \\ &= \sum_{j=1}^{n} |v_j - w_j| \sum_{i=1}^{n} b_{ij} \\ &= \mu \sum_{j=1}^{n} |v_j - w_j| \\ &= \mu \mathrm{d}(\mathbf{v}, \mathbf{w}) \end{aligned} \tag{5}$$

Iterating inequality (5) yields

$$\mathrm{d}(S^k\mathbf{v}, S^k\mathbf{w}) \leq \mu^k \mathrm{d}(\mathbf{v}, \mathbf{w}). \tag{6}$$

Since fraction vectors have non–negative entries which sum to 1, the greatest distance between any two fraction vectors is 2:

$$\mathrm{d}(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{n} |v_i - w_i] \leq \sum_{i=1}^{n} v_i + w_i = 2$$

So, no matter what different initial fraction vectors experimenters begin with, after k iterations the resulting fraction vectors are within $2\mu^k$ of each other, and by choosing k large enough, we can deduce the existence of, and estimate the common limit $\mathbf{z}$ with as much precision as desired. Furthermore, if all initial material is allotted to node j, then the initial fraction vector $\mathbf{e_j}$ has a 1 in position j and zeroes elsewhere. $S^k\mathbf{e_j}$, (or $A^N\mathbf{e_j}$) is on one hand the j^{th} column of S^k (or A^N), but on the other hand is converging to $\mathbf{z}$. So each column of the limit matrix for S^k and A^N equals $\mathbf{z}$. Finally, if $\mathbf{x}_0$ is any initial fraction vector, then $S(S^k\mathbf{x}_0) = S^{k+1}(\mathbf{x}_0)$ is converging to $S(\mathbf{z})$ and also to $\mathbf{z}$, so $S(\mathbf{z}) = \mathbf{z}$ (and $A\mathbf{z} = \mathbf{z}$). Since the entries of $\mathbf{z}$ are non–negative (and sum to 1) and the entries of S are all positive, the entries of $S\mathbf{z}$ $(= \mathbf{z})$ are all positive. ■

Stage 4: The Google fudge factor

Sergey Brin and Larry Page realized that the world wide web is not almost stochastic. However, in addition to realizing that the Perron–Frobenius theorem was potentially useful for ranking URLs, they figured out a simple way to guarantee stochasticity—the "Google fudge factor."

Rather than using the voting matrix A described in the previous stages, they take a combination of A with the matrix of 1s we called 1. For (Brin an Pages' choice of) $\varepsilon = .15$ and n equal the number of nodes, consider the Google matrix

$$G = (1 - \varepsilon)A + \frac{\varepsilon}{n}1.$$

(See [Austin, 2008]).

If A is almost stochastic, then each column of G also sums to 1 and each entry is at least ε/n. This G is stochastic! In other words, if you use this transition matrix everyone gets a piece of your play–doh, but you still get to give more to your friends.

7. Consider the giving game from 5c. Its transition matrix

$$A = \left[\begin{array}{cccc:cc} 0 & 0 & 0 & .5 & 0 & 0 \\ .5 & 0 & 0 & 0 & 0 & 0 \\ .5 & 1 & 0 & .5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right]$$

   is not almost stochastic. For $\varepsilon = .3$ and $\varepsilon/n = .05$, work out the Google matrix G, along with the limit rankings for the six sites. If you were upset that site 4 was ranked as equal to site 3 in the game you played for stage 1, you may be happier now.

Historical notes

The Perron–Frobenius theorem had historical applications to input–output economic modeling. The idea of using it for ranking seems to have originated with Joseph B. Keller, a Stanford University emeritus mathematics professor. According to a December 2008 article in the Stanford Math Newsletter [Keller, 2008], Professor Keller originally explained his team ranking algorithm in the 1978 Courant Institute Christmas Lecture, and later submitted an article to Sports Illustrated in which he used his algorithm to deduce unbiased rankings for the National League baseball teams at the end of the 1984 season. His article was rejected. Utah professor James Keener visited Stanford in the early 1990s, learned of Joe Keller's idea, and wrote a SIAM article in which he ranked football teams [Keener, 1993].

Keener's ideas seem to have found their way into some of the current BCS college football ranking schemes which often cause boosters a certain amount of heartburn. I know of no claim that there is any direct path from Keller's original insights, through Keener's paper, to Brin and Pages' amazing Google success story. Still it is interesting to look back and notice

that the seminal idea had been floating "in the air" for a number of years before it occurred to anyone to apply it to Internet searches.

Acknowledgement: Thanks to Jason Underdown for creating the graph diagrams and for typesetting this document in LaTeX.

References

David D. Austin. How Google Finds Your Needle in the Web's Haystack. 2008. URL `http://www.ams.org/featurecolumn/archive/pagerank.html`.

Sergey Brin and Lawrence Page. The anatomy of a large–scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 33:107–117, 1998. URL `http://infolab.stanford.edu/pub/papers/google.pdf`.

James Keener. The Perron–Frobenius Theorem and the ranking of football teams. *SIAM Rev.*, 35:80–93, 1993.

Joseph B. Keller. Stanford University Mathematics Department newsletter, 2008.