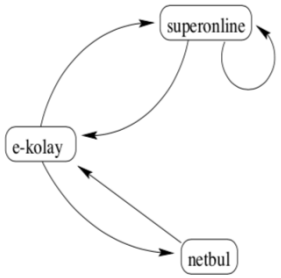
How Google Ranks Web Pages

The aim of these notes is to illustrate how linear algebra can be made use of in internet search engines. When several pages match a search query, the pages should be displayed in order of their *importance*. So how is this page importance determined? Consider the patent pending `PageRank` scheme used by `Google` to determine the importance of each web page: A page is important if important pages link to it. This is a recursive definition of importance which needs to be solved. If we imagine that each page has one unit importance initially, then we can iterate a process of each page sharing whatever importance it has among its successors and receiving new importance from its predecessors. This can be represented as a matrix vector product as illustrated in the following example: Consider the following link structure:



Let s, n and e denote the importance of Superonline, Netbul and E-kolay respectively. Then, in the following iteration, a column means that the corresponding site is giving $1/m$th of its importance (with m being the outdegree of the site) to the sites it points to. The iteration for this example is given as:

$$\begin{bmatrix} s^{(t+1)} \\ n^{(t+1)} \\ e^{(t+1)} \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} s^{(t)} \\ n^{(t)} \\ e^{(t)} \end{bmatrix}$$

where $s^{(0)} = n^{(0)} = e^{(0)} = 1$.

Eventually, the above iteration will reach a limit which happens to be *its component* in the principal eigenvector of this *example* matrix. In the limit, the solution will be $s = 6/5$, $e = 6/5$ and $n = 3/5$, i.e. Superonline and E-kolay will have the same importance. Netbul, on the other hand, will have half the importance.