How Google works

or How linear algebra powers the search engine

M. Ram Murty, FRSC
Queen’s Research Chair
Queen’s University

Google AS A NOUN!

"They're encyclopedias, Timmy. . . they're an early form of **Google**."

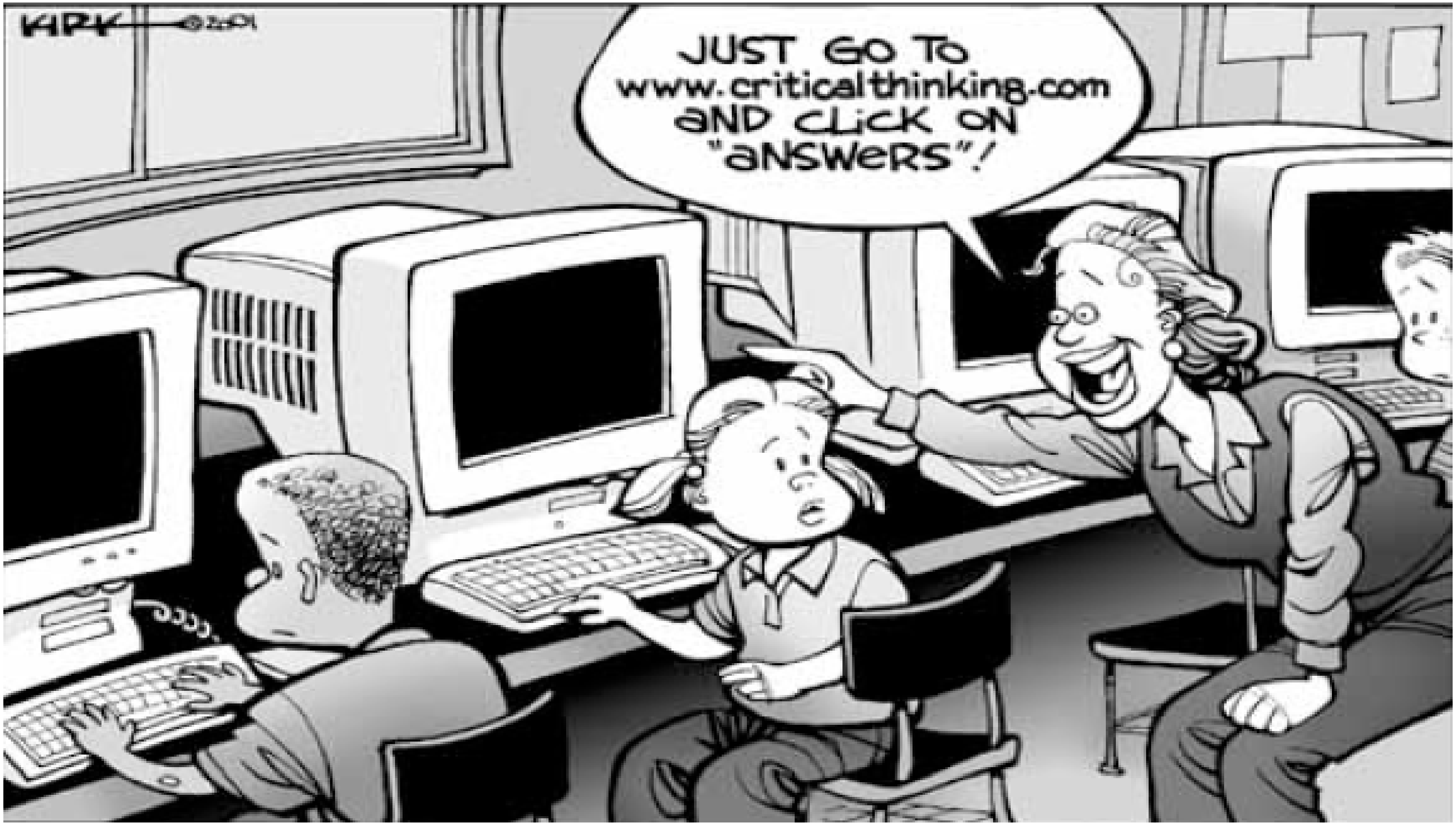

[redacted]

"I'M STUCK. CHECK IT OUT ON GOOGLE."

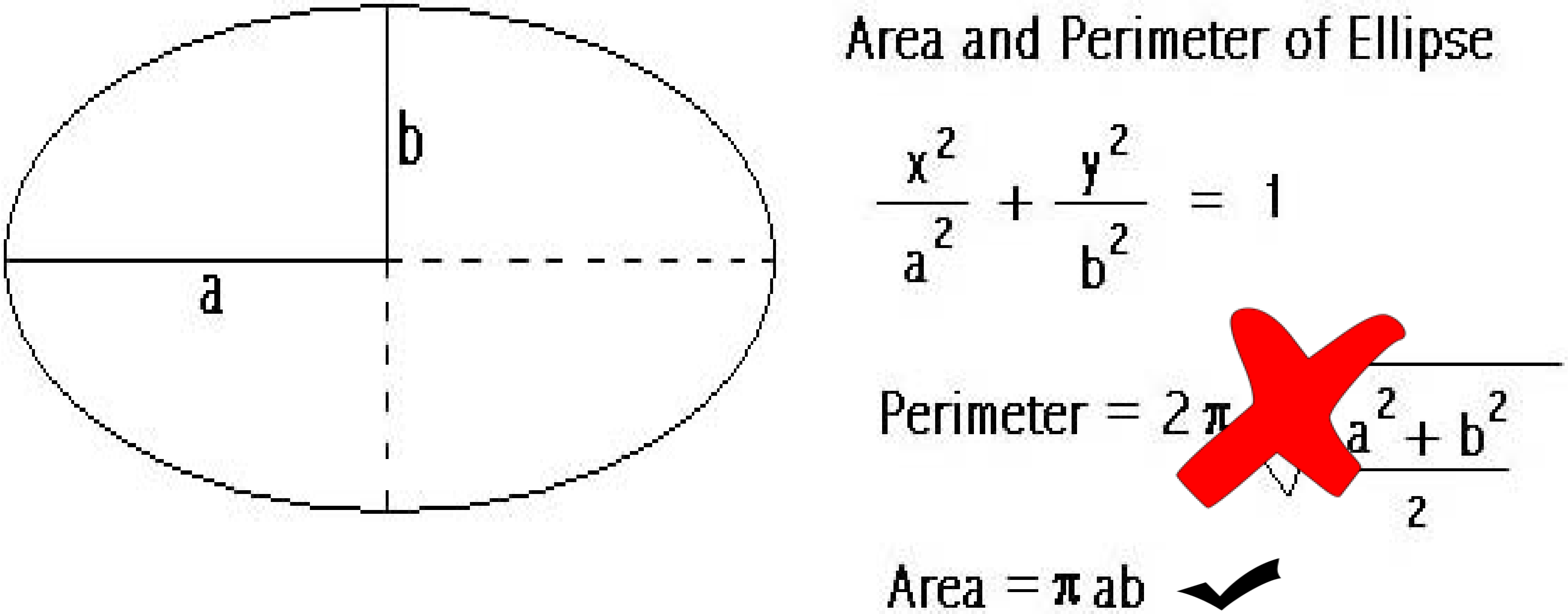
Google IS NOW A VERB!

"YOU'VE STUMPED ME WITH THAT QUESTION.
I THINK THAT'S SOMETHING YOU NEED TO GOOGLE."

Google AS AN ORACLE!



From: gomath.com/geometry/ellipse.php



Area and Perimeter of Ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Perimeter $= \cancel{2\pi\sqrt{\frac{a^2+b^2}{2}}}$

Area $= \pi ab$ ✓

Metric mishap causes loss of Mars orbiter (Sept. 30, 1999)

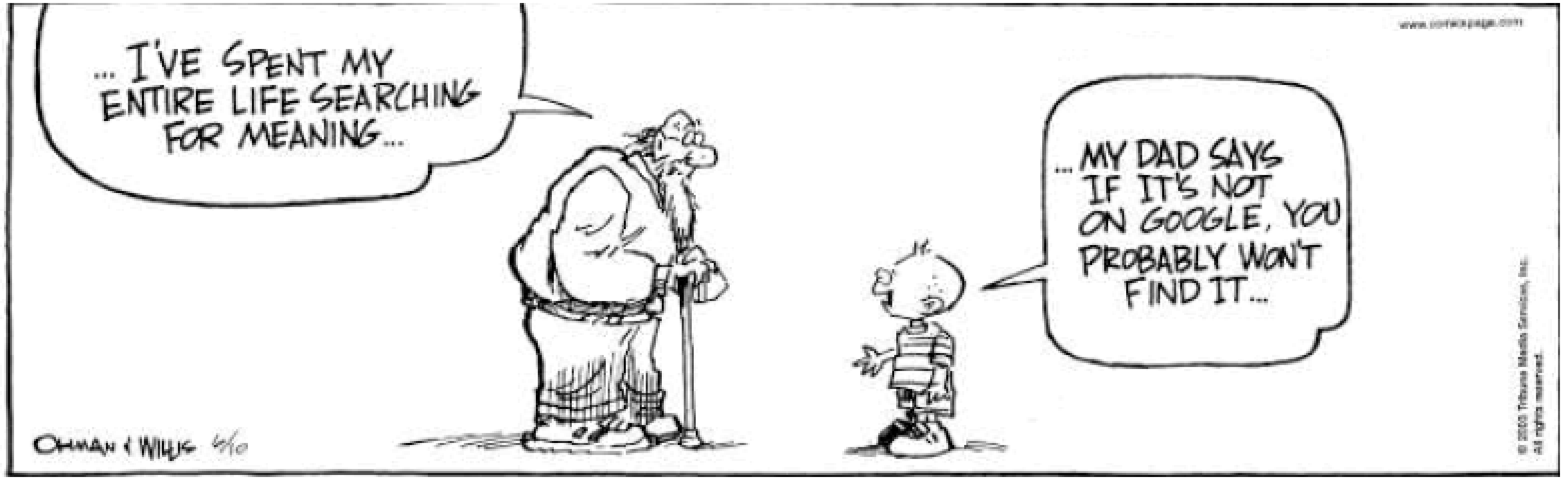
(NASA)

CNN NASA lost a 125 million Mars orbiter because a Lockheed Martin engineering team used English units of measurement while the agencys team used the more conventional metric system for a key spacecraft operation, according to a review finding released Thursday.

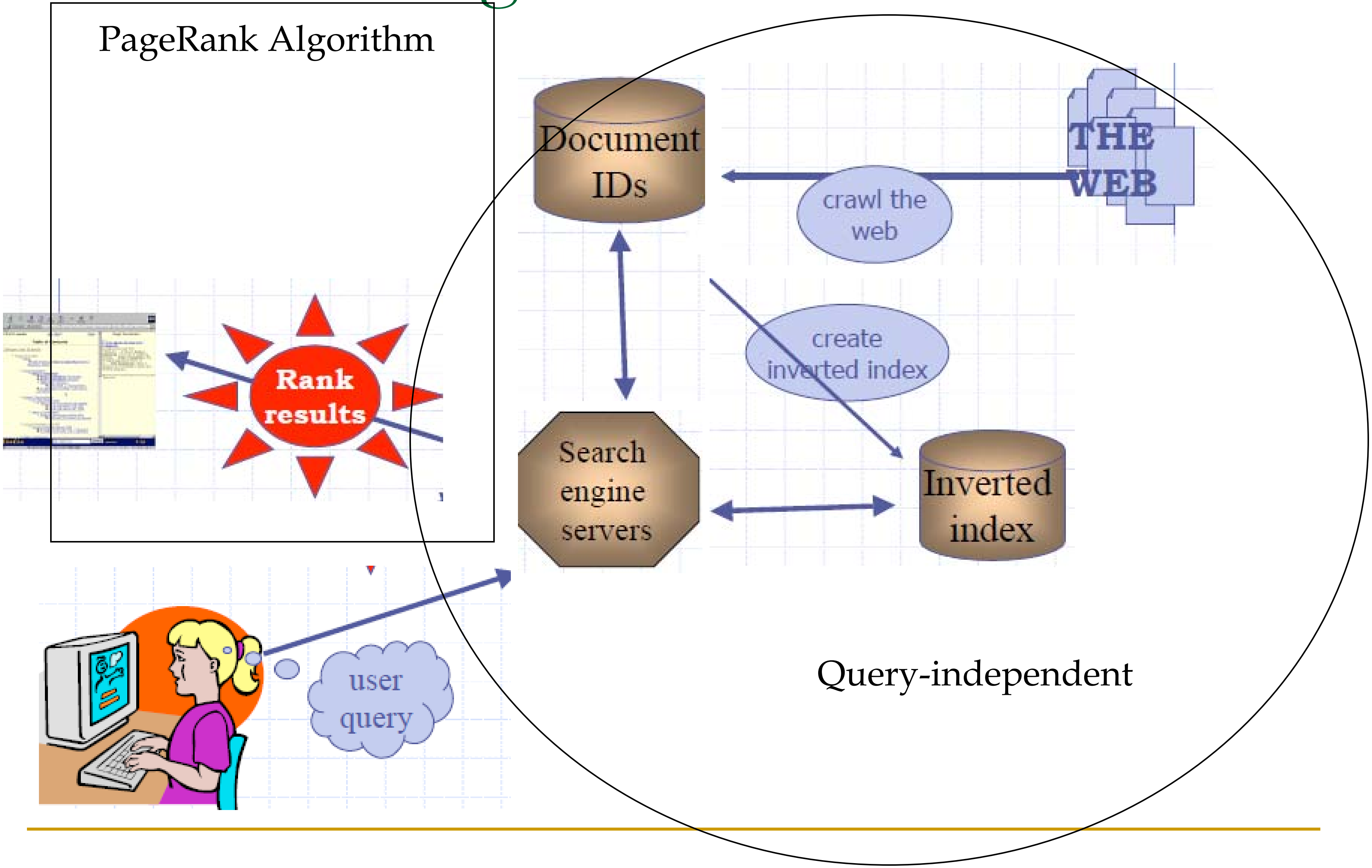
The units mismatch prevented navigation information from transferring between the Mars Climate Orbiter spacecraft team in at Lockheed Martin in Denver and the flight team at NASAs Jet Propulsion Laboratory in Pasadena, California.

Lockheed Martin helped build, develop and operate the spacecraft for NASA. Its engineers provided navigation commands for Climate Orbiters thrusters in English units although NASA has been using the metric system predominantly since at least 1990.

The limitations of Google!



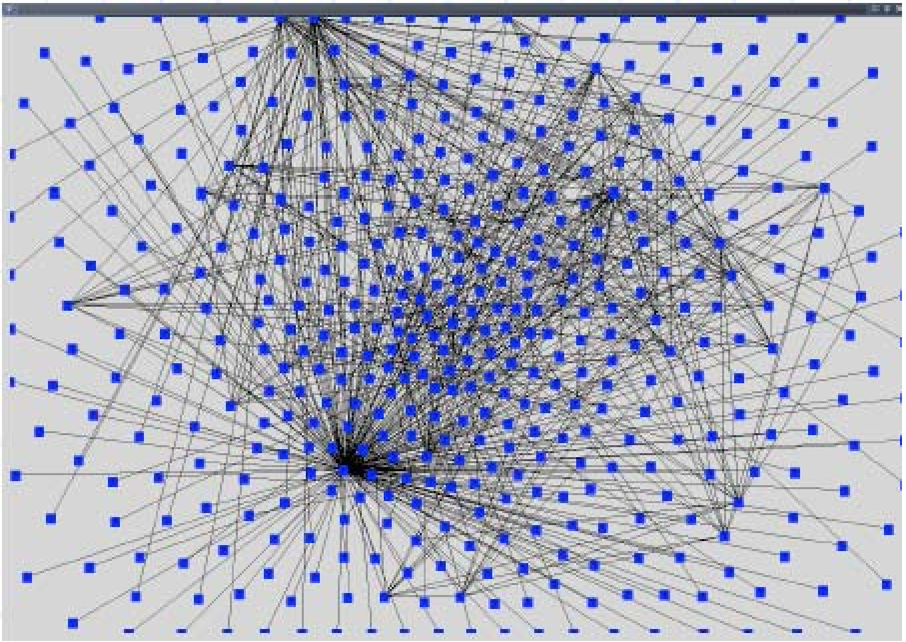
The web at a glance



The web is a directed graph

- The nodes or vertices are the web pages.
- The edges are the links coming into the page and going out of the page.

This graph has more than 10 billion vertices and it is growing every second!



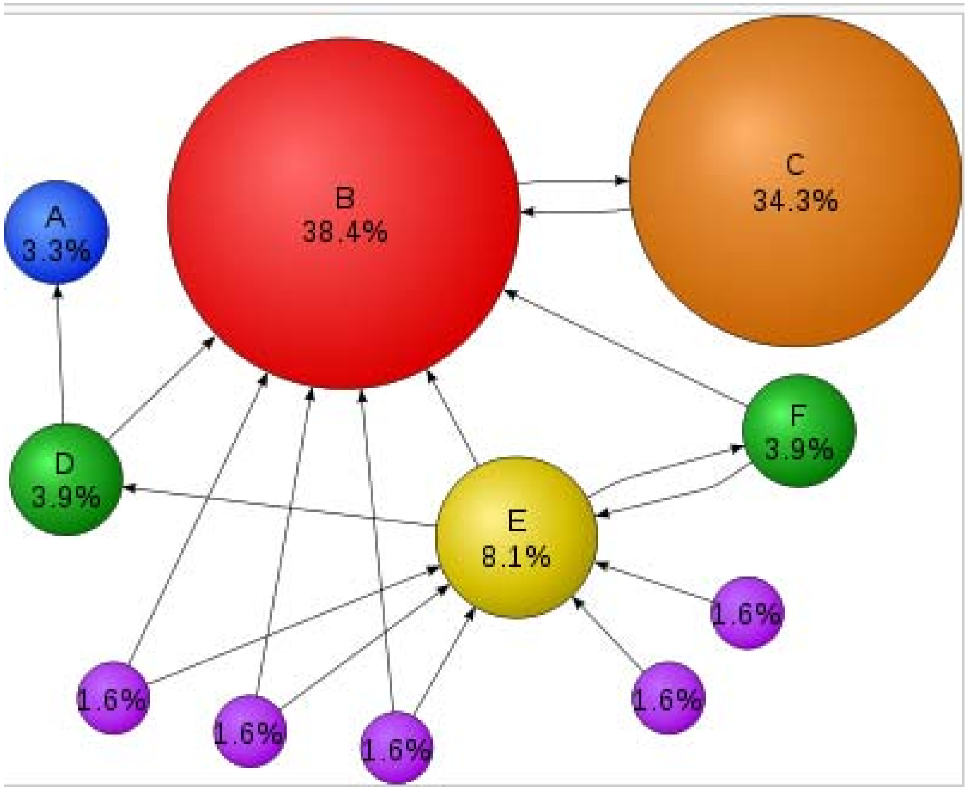
The PageRank Algorithm

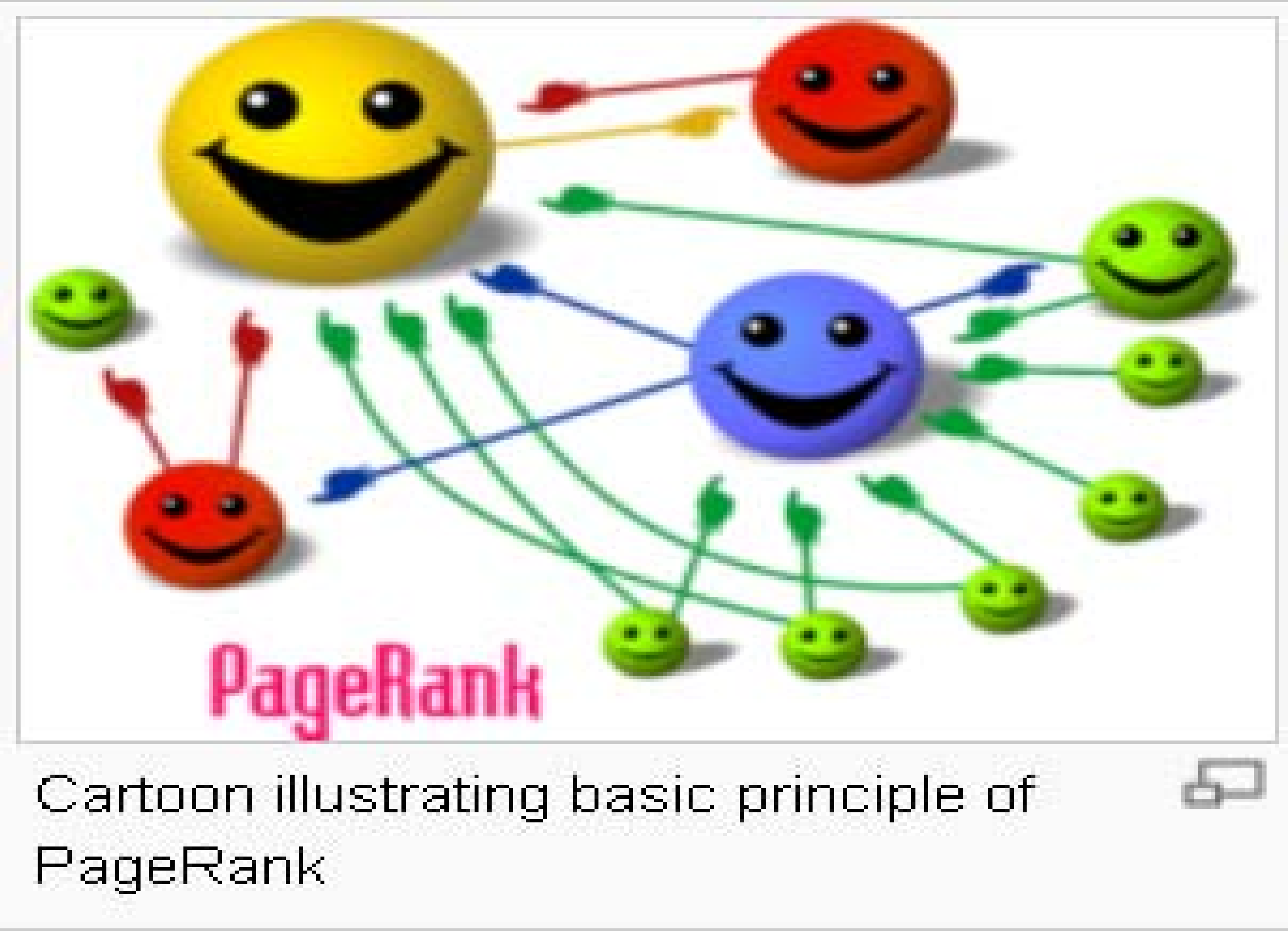
- PageRank Axiom: A webpage is important if it is pointed to by other important pages.
- The algorithm was patented in 2001.

Sergey Brin and Larry Page

Example

- C has a higher rank than E, even though there are fewer links to C since the one link to C comes from an "important" page.

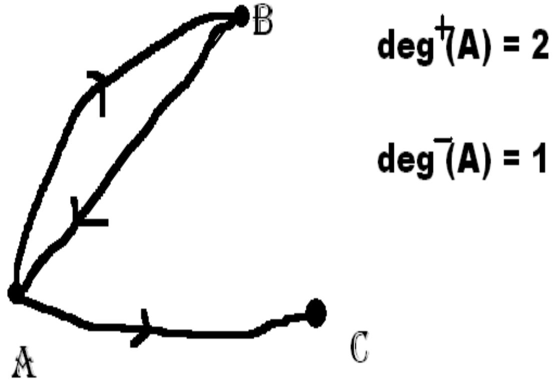




Cartoon illustrating basic principle of PageRank

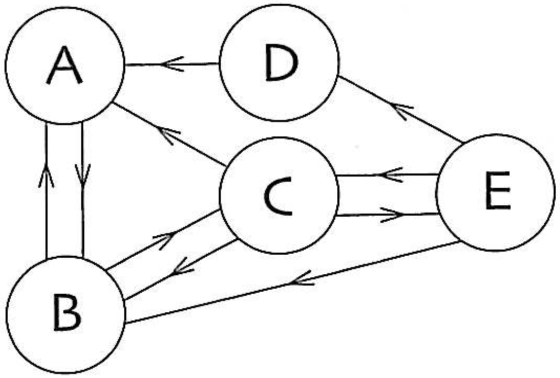
Mathematical formulation

- Let r(J) be the “rank” of page J.
- Then r(K) satisfies the equation $r(K)= \Sigma_{J\to K}\, r(J)/\deg^{+}(J)$, where $\deg^{+}(J)$ is the outdegree of J.



The web and Markov chains

- Let p_{uv} be the probability of reaching node u from node v.
- For example, $p_{AB}=1/2$ and $p_{AC}=1/3$ and $p_{AE}=0$.



Notice the columns add up to 1.
Thus, (1 1 1 1 1)P=(1 1 1 1 1).
P^t has eigenvalue 1

P is called the transition matrix.

$$P = \begin{array}{ccccccc} & A & B & C & D & E & \\ \left(\begin{array}{c} \\ \\ \\ \\ \\ \end{array}\right. & \begin{array}{c} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{array} & \begin{array}{c} \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \\ 0 \end{array} & \begin{array}{c} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{array} & \begin{array}{c} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} & \begin{array}{c} 0 \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \end{array} & \left.\begin{array}{c} \\ \\ \\ \\ \\ \end{array}\right) \begin{array}{c} A \\ B \\ C \\ D \\ E \end{array} \end{array}$$

Markov process

- If a web user is on page C, where will she be after one click? After 2 clicks? … After n clicks?

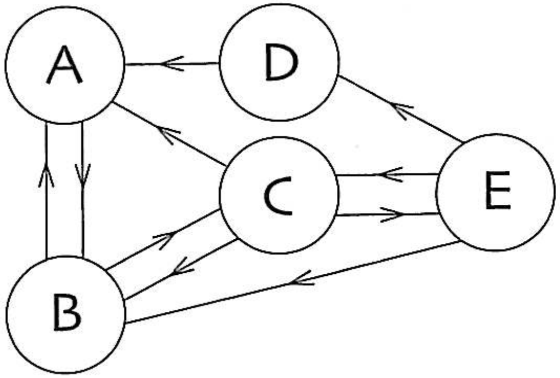
$$p^0 = \begin{pmatrix} p(X_0 = A) \\ p(X_0 = B) \\ p(X_0 = C) \\ p(X_0 = D) \\ p(X_0 = E) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

$$p^1 = \begin{pmatrix} p(X_1 = A) \\ p(X_1 = B) \\ p(X_1 = C) \\ p(X_1 = D) \\ p(X_1 = E) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix}$$

$$p^2 = \begin{pmatrix} p(X_2 = A) \\ p(X_2 = B) \\ p(X_2 = C) \\ p(X_2 = D) \\ p(X_2 = E) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ 0 \\ 0 \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ \frac{4}{9} \\ \frac{5}{18} \\ \frac{1}{9} \\ 0 \end{pmatrix}$$

After n steps, $P^n p^0$.

A.A. Markov (1856-1922)



Eigenvalues and eigenvectors of P

$$\Delta_{P^t}(\lambda) = \det(\lambda I - P^t) = \det(\lambda I - P)^t = \det(\lambda I - P) = \Delta_P(\lambda),$$

- Therefore, P and P^t have the same eigenvalues.
- In particular, P also has an eigenvalue equal to 1.

Theorem of Frobenius

- All the eigenvalues of the transition matrix P have absolute value ≤ 1.
- Moreover, there exists an eigenvector corresponding to the eigenvalue 1, having all non-negative entries.

Georg Frobenius (1849-1917)

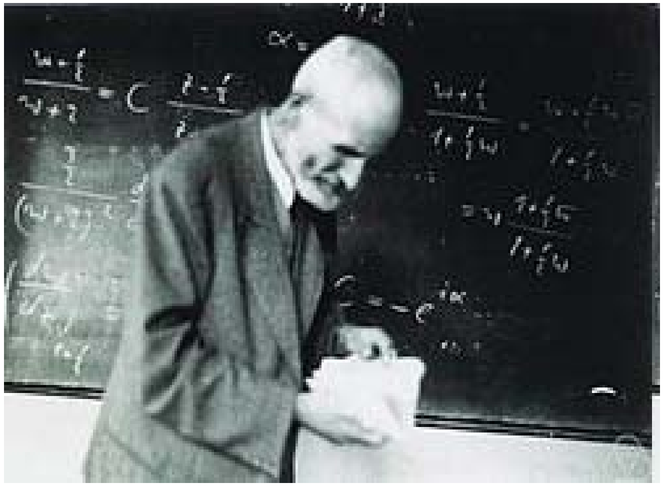
SIAM REVIEW
Vol. 48, No. 3, pp. 569–581

**The $25,000,000,000 Eigenvector:
The Linear Algebra behind Google***

Kurt Bryan[†]
Tanya Leise[‡]

Perron's theorem

- Theorem (Perron): Let A be a square matrix with strictly positive entries. Let λ* = max{ |λ|: λ is an eigenvalue of A}. Then λ* is an eigenvalue of A of multiplicity 1 and there is an eigenvector with all its entries strictly positive. Moreover, |λ|< λ* for any other eigenvalue.



O. Perron (1880-1975)

Frobenius's refinement

- Call a matrix A irreducible if A^n has strictly positive entries for some n.
- Theorem (Frobenius): If A is an irreducible square matrix with non-negative entries, then λ^* is again an eigenvalue of A with multiplicity 1. Moreover, there is a corresponding eigenvector with all entries strictly positive.

Why are these theorems important?

- We assume the following concerning the matrix P:
- (a) P has exactly one eigenvalue with absolute value 1 (which is necessarily =1);
- (b) The corresponding eigenspace has dimension 1;
- (c) P is diagonalizable; that is, its eigenvectors form a basis.
- Under these hypothesis, there is a unique eigenvector v such that Pv = v, with non-negative entries and total sum equal to 1.
- Frobenius's theorem together with (a) implies all the other eigenvalues have absolute value strictly less than 1.

Computing $P^n p^0$.

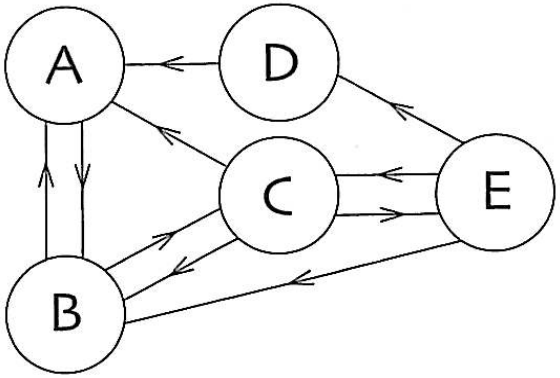
- Let v_1, v_2, …, v_5 be a basis of eigenvectors of P, with v_1 corresponding to the eigenvalue 1.
- Write $p^0 = a_1 v_1 + a_2 v_2 + \ldots + a_5 v_5$.
- It is not hard to show that $a_1=1$.
- Indeed, $p^0= a_1 v_1 + a_2 v_2 + \ldots + a_5 v_5$
- Let $J=(1,1,1,1,1)$.
- Then $1 = J\, p^0= a_1 Jv_1 + a_2 Jv_2 + \ldots + a_5 Jv_5$
- Now $Jv_1=1$, by construction.
- For $i\geq 2$, $J(Pv_i) = (JP)v_i = Jv_i$. But $Pv_i = \lambda_i v_i$.
- Hence $\lambda_i Jv_i = Jv_i$. Since $\lambda_i \neq 1$, we get $Jv_i =0$.
- Therefore $a_1=1$.

Computing $P^n p^0$ continued

- $P^n p^0 = P^n v_1 + a_2 P^n v_2 + \ldots + a_5 P^n v_5$
- $\quad = v_1 + \lambda_2^n a_2 v_2 + \ldots + \lambda_5^n a_5 v_5$.
- Since the eigenvalues $\lambda_2, \ldots, \lambda_5$ have absolute value strictly less than 1, we see that $P^n p^0 \to v_1$ as n tends to infinity.
- Moral: It doesn't matter what p^0 is, the stationary vector for the Markov process is v_1.

Returning to our example …

- The vector (12, 16, 9, 1, 3) is an eigenvector of P with eigenvalue 1.
- We can normalize it by dividing by 41 so that the sum of the components is 1.
- But this suffices to give the ranking of the nodes:B, A, C, E, D.



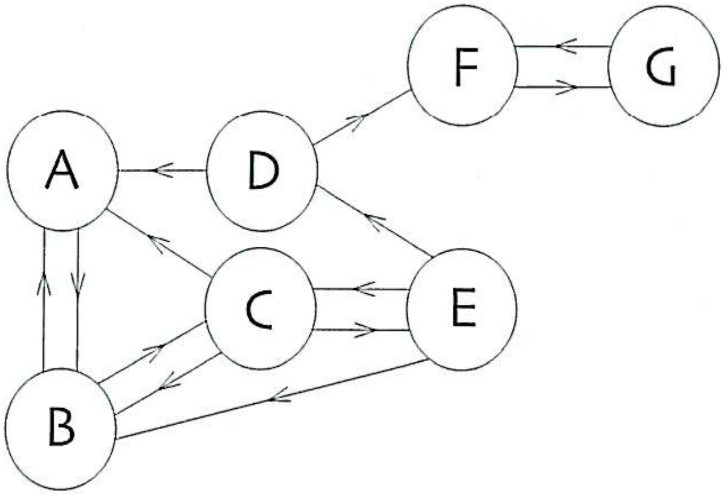
$$
P = \begin{array}{c} \\ \end{array}
\begin{array}{ccccc|c}
A & B & C & D & E & \\
0 & \frac{1}{2} & \frac{1}{3} & 1 & 0 & A \\
1 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & B \\
0 & \frac{1}{2} & 0 & 0 & \frac{1}{3} & C \\
0 & 0 & 0 & 0 & \frac{1}{3} & D \\
0 & 0 & \frac{1}{3} & 0 & 0 & E
\end{array}
$$

How to compute the eigenvector

- We can apply the power method: Compute $P^n p^0$ for very large n to get an approximation for v_1.
- This is called the power method and there are efficient algorithms for this large matrix computation.
- It seems usually 50 iterations (i.e. n=50) are sufficient to get a good approximation of v_1.

Improved PageRank

- If a user visits F, then she is caught in a loop and it is not surprising that the stationary vector for the Markov process is $(0,0,0,0,0, ½, ½)^t$.
- To get around this difficulty, the authors of the PageRank algorithm suggest adding to P a stochastic matrix Q that represents the "taste" of the surfer so that the final transition matrix is $P' = xP + (1-x)Q$ for some $0 \leq x \leq 1$.
- Note that P′ is again stochastic.
- One can take Q=J/N where N is the number of vertices and J is the matrix consisting of all 1's.
- Brin and Page suggested x=.85 is optimal.



References

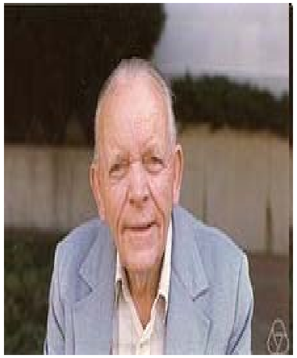
- Mathematics and Technology, by C. Rousseau and Y. Saint-Aubin, Springer, 2008.
- Google's PageRank and Beyond, The Science of Search Engines, A. Langville and C. Meyer, Princeton University Press, 2006.
- The 25 billion dollar eigenvector, K. Bryan and T. Liese, SIAM Review, 49 (2006), 569-581.

Mathematical genealogy

P.L.Chebychev
(1821-1894)

A.A.Markov
(1856-1922)

J.D.Tamarkin
(1888-1945)



D.H. Lehmer
(1905-1991)

H.M. Stark
(1939-

Thank you for your attention.

- Have a Goooooooooogle day!

"I looked up your symptoms on Google.
If you want a second opinion, I can check Yahoo."