Introduction of higher order effects into page ranking*(and some other ideas)

Vasilii Shelkov

vasia@slac.stanford.edu

Proposal

*In the large cross-references information archives or documents catalogs, the **association of the document relevancy with a single number(rank) can be misleading**. The non-uniform density of number of citations across the system, as well as correlations among the sources of these citations need to be taken into consideration.*

- Introduction
- Examples of non-uniformity and correlations in page rankings
- Interpretation of non-uniformity and correlations
- Artificial directory web-pages
- Monte-Carlo of WWW and other ideas
- Conclusion

---

* Patent pending

Introduction

- the structure of WWW has dual properties of interconnected graph as well as random statistical ensemble; the introduction of PageRank™ efficiently addresses the graph part, but makes no use of its statistical random behaviour,

- it‘s interesting, that the gain in search quality obtained from application of page ranking was obtained by capitalizing on the external source of information based not on the web page content, but the topology of page-to-page links of WWW graph,

- the original global PageRank approach doesn‘t take into account neither possible non-uniformity in the distribution of the number of incoming links across WWW, nor correlations between page rankings of various web documents(i.e. second order effects),

- The obvious question is which of 2 properties are more important? I will show that the distribution of number of incoming links demonstrates remarkable statistical properties and we should either 1) use it or 2) “loose it“ (i.e. add more structure to WWW by hands)

the first goal of this note is to demonsrate that there is a simple to implement, way of extending PageRank™ to account for statistical effects where It useful,

the second goal is to show how we can add more useful stucture to WWW by introducing artificial web-directories directly into WWW graph.

Selection of random WWW documents

-if, for a given web-page, the distribution of number of incoming links demonstrates random statistical behavior(e.g. Gaussian-like shapes), the following properties could be expected: the narrower distribution of the number of incoming links increases the effective rank of the page, while, for a pair of documents, the correlations between the sources of these incoming links effectively decrease their compound importance(rank),

- the most straight-forward way to obtain a distribution of page ranking(popularity) is to take a sequence of random web-page samples, and plot the distribution of number of incoming links calculated for each of these samples,

- here, to get sequences of random web-pages, we employ a "Poor Man's trick" described on the next slide,

- to illustrate main ideas, the distributions of number of incoming links for cnn.com, ntimes.com, news.bbc.co.uk, and msnbc.com web-sites are shown on the next slide,

- all the distributions are based on 150 samples consisting of 50K-500K of random web-pages each,

- the selection of random web-pages are done with 4-digit random numerical keys which ensure that selected samples of web pages don't intersect(see next slide),

“Poor Man‘s Trick“

- in the absence of access to the full copy(index) of all the pages in WWW, it’s not clear how to acquire a sample of random web-pages,

- is there any way to rely on some random features of the page contents Instead of generating random IDs for the whole WWW index???

Problem: information is not organized, different languages, formats – hard!!!

Solution(wild guess) practically everybody around the Globe is using Latin numbering system (phones, Zip codes, byte sizes,…), which, given the size of WWW, could turn out to be a pretty good source of random digits

some random number

Example(using character string e.g. ”microsoft.com + RNDM”)

				Ratio($x10^{-3}$)
“+517265”,	N(hits) = 170”	“microsoft.com +517265”,	N(hits) = 1	**9.3**
“ +45762”,	N(hits) = 4690”	“microsoft.com +45762”,	N(hits) = 35	**7.5**
“ +3876”,	N(hits) = 151000”	“microsoft.com +3876”,	N(hits) = 852	**5.6**
“ +495”,	N(hits) = 3880000”	“microsoft.com +495”,	N(hits) = 27,100	**7.0**
“ +78”,	N(hits) = 28000000”	“microsoft.com +78”,	N(hits) = 187,000	**6.7**

- N(hits) scales almost like factor of 10
- ratio stays stable over 5! orders of magnitude in number of hits
- sample selected with 4-digit “keys” – almost 100% independent!

Examples of non-uniformity of page rankings for random WWW documents samples

Shapes:

- "bell-like" shapes appear to Indicate a high degree of randomness in finding links to a specific web-page,

- all distributions already have tails, and tails may grow longer when PageRank™ converges,

- widths, peak-vs-mean shifts are quite different,

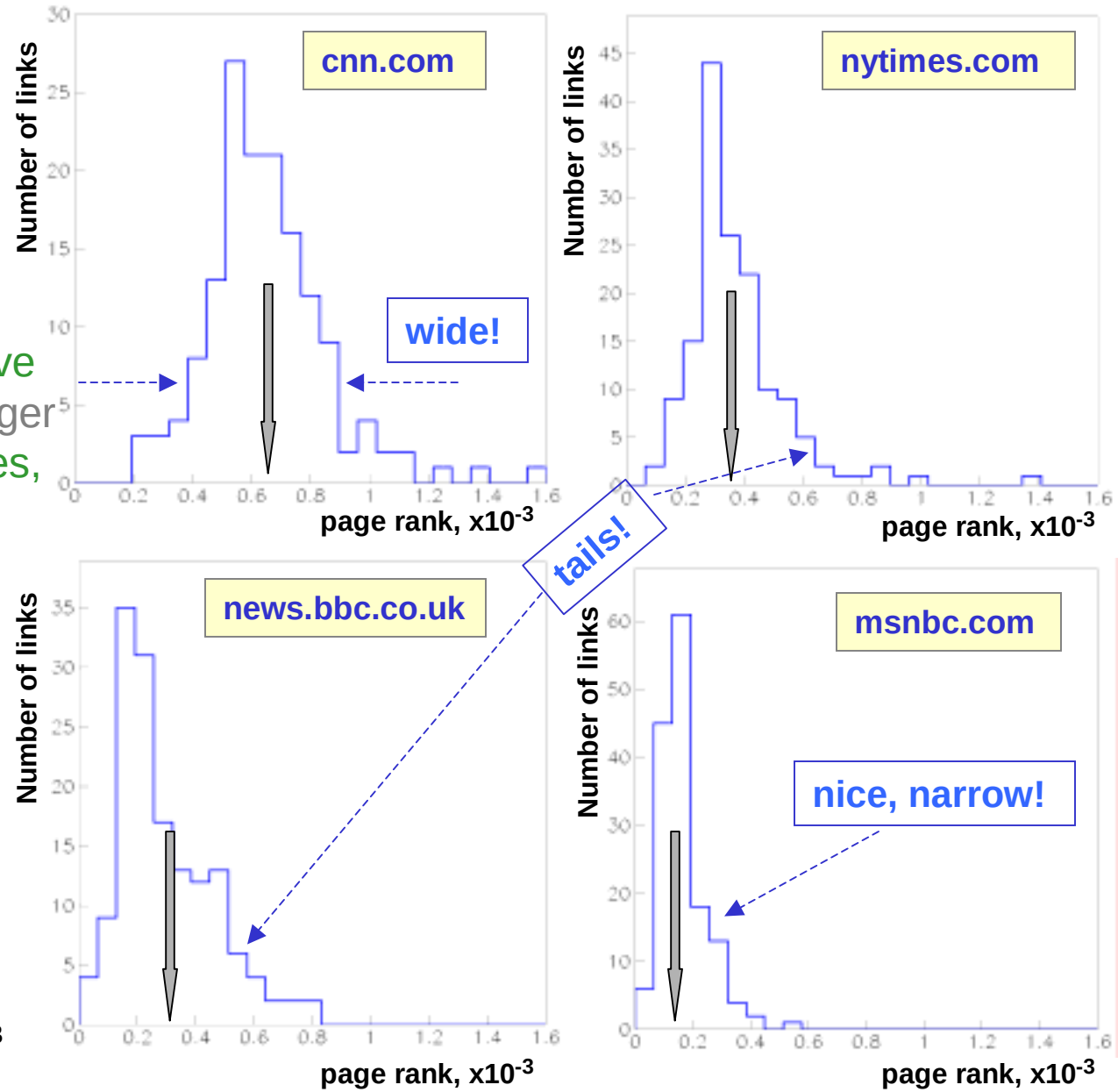
Average number of N(incoming links)/N_{sample}

$$r_{cnn} = 0.653 \times 10^{-3}$$

$$r_{nytimes} = 0.371 \times 10^{-3}$$

$$r_{bbc} = 0.295 \times 10^{-3}$$

$$r_{msnbc} = 0.165 \times 10^{-3}$$



Examples of correlations in page rankings

- If we ask a question: How often do we see cases when a random web page containes links to cnn.com and nytimes.com pages at the same time? In other words, what are the degree of correlations in the number of incoming links among various URLs?

- To answer, let‘s consider our 4 URLs, and formally calculate 4x4 correlation matrix:

$$Cov(x,y) = \sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

$$Cor(x,y) = Cov(x,y) / (\sigma_X \cdot \sigma_Y)$$

$$\sigma_X^2 = \sum_{i=1}^{N} (x_i - \overline{x})^2$$

- Here‘s what we get(factor x10^{-3} removed):

cnn	nytimes	bbc	msnbc

$$Cov = \begin{Vmatrix} 0.043 & 0.011 & 0.011 & 0.007 \\ 0.011 & 0.028 & 0.007 & 0.006 \\ 0.011 & 0.007 & 0.025 & 0.004 \\ 0.007 & 0.006 & 0.004 & 0.006 \end{Vmatrix} \quad Cor = \begin{Vmatrix} 1.000 & 0.307 & 0.334 & 0.422 \\ 0.307 & 1.000 & 0.270 & 0.469 \\ 0.334 & 0.270 & 1.000 & 0.289 \\ 0.422 & 0.469 & 0.289 & 1.000 \end{Vmatrix}$$

cnn
nytimes
bbc
msnbc

Correlations are huge!.

Interpretation of widths, tails

To help with argumentation, I'd like to use orginal terms introduced in **L. Page** and **S. Brin** in the PageRank paper: a given WWW document gets "votes" from other sites which serve as "referees"

Widths and tails:

- the shape of **N(votes)/N(sample)** distributions can be related to the density of the page **referees** across the WWW graph:

- large widths may be caused by clusters of "referees" voting either abnormally high or abnormally low, and thus indicate a "ranking controversy". For two pages with similar ranks, the one with a smaller width has more uniform WWW graph penetration, and thus, should be "rewarded" for this "qualitative edge"(with somewhat higher rank!),

- long tails are more dangerous than symmetric width increases, since they drive the mean rank away from the most probable value(the peak of the distribution), and thus create a biased value of page rank(artificially high or low),

- long tails are expected when a page gets few "votes" from some high rank referees,

- it is quite possible, that long tails can reflect an "unhealthy" activity by some companies to create islands of hyper-high voting activity(e.g. by asking their multiple customers/daughter companies to provide back links(even if they are not needed))

Interpretation of correlations

Rank Correlations:

- Imagine we have three WWW documents(A,B,C), A,B have an identical set of referees(I.e. 100% correlated in rank), and C is uncorrelated with A and B. The following ranks are found: r(A)=0.5, r(B)=0.4, r(C)=0.4. How would we order them? Doesn't it make sense to read the best book recommended by one set of "referees"(A), and then choosing between equally good ones(B,C), read the one recommended by an independent panel of "critics"(C)?

Content Correlations:

- Everywhere in the text, we had in mind page rank correlations between different Web documents, but one could extend this concept to account for the content correlations between web documents(e.g. most of the Web pages serving News Information are highly correlated in content and often even have identical article headings)

Both rank and content correlations can(should) be treated simultaneously!

Impact of Correlations:

Could as important as PageRank™ itself!

$$\begin{Vmatrix} 1.000 & 0.307 & 0.334 & 0.422 \\ 0.307 & 1.000 & 0.270 & 0.469 \\ 0.334 & 0.270 & 1.000 & 0.289 \\ 0.422 & 0.469 & 0.289 & 1.000 \end{Vmatrix} \quad \sum_{On-diagonal} = 4 \quad \sum_{Off-diagonal} = 4.2 \; !$$

How do we account for all these effects quantitatively?

One Simple Illustration!

How do we order links for a given search?

- start with a sample of *N* documents which have a “reasonable” degree of text pattern match,

- assume that the information we a looking for is distributed among these documents and the goal is to find a linear combination of these pages which would minimize this combination’s resulting covariance *S* and maximize its overall pattern-rank weight *D*, where

$$D = \sum_{i=1}^{N} \lambda_i d_i \ , \ \ d = w \cdot \overline{r}^{\,2}, \ \ S = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j s_{ij} \ , \text{where}$$

$$w - \text{ text match weight, } \ \overline{r} \ - \text{ mean page rank,}$$

$$s_{ij} = \sum_{\mathrm{k=1}}^{\mathrm{Nsamples}} (r_i^k - \overline{r}_i)(r_j^k - \overline{r}_j), \quad r_i \ - \text{ page rank of web-page } i$$

$$\lambda_i, \lambda_j - \text{ new ranks of web-page } i \text{ and } j$$

- the coefficients λ_i which enter this linear combination, would give the order in which these links have to be ordered on the search results page

How do we order links for a given search?

- to maximize this ratio we calculate:

$$\frac{\partial}{\partial\lambda}\left(\frac{D}{S}\right)=\frac{D}{S^2}\left(\frac{S}{D}\times\frac{\partial D}{\partial\lambda}-\frac{\partial S}{\partial\lambda}\right)=0,\ \frac{\partial^2}{\partial\lambda^2}\left(\frac{D}{S}\right)=-\frac{D}{S^2}\frac{\partial^2 S}{\partial\lambda^2}<0\ ,\ \text{where}$$

$$\lambda-\{\lambda_1,\lambda_2,...\},\ \frac{S}{D}\ -\ \text{is a constant factor for all } \lambda_i\text{, so}$$

we just need to resolve a charachteristic equation: $\boxed{\frac{\partial S}{\partial\lambda}=\frac{\partial D}{\partial\lambda}}$

Mean page ranks!

- **Example**: for our 4 web-pages we simply need to solve:

Covariance matrix!

$$\begin{bmatrix} 0.043 & 0.011 & 0.011 & 0.007 \\ 0.011 & 0.028 & 0.007 & 0.006 \\ 0.011 & 0.007 & 0.025 & 0.004 \\ 0.007 & 0.006 & 0.004 & 0.006 \end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\\ \lambda_3\\ \lambda_4\end{bmatrix}=\begin{bmatrix}0.65^2\cdot w_1\\ 0.37^2\cdot w_2\\ 0.30^2\cdot w_3\\ 0.17^2\cdot w_4\end{bmatrix}\quad\text{where,}$$

λ_i – unknowns(new ranking), w_i – text match weights

Some examples

Ideal case of Poisson statistics:

- it is important to note that we maximize the linear combination which has a quadratic dependence on the average page rank:

$$D = \sum_{i=1}^{N} \lambda_i d_i \ , \ \ d = w \cdot \overline{r}^2$$

- the reason is clear when we consider an ideal case when:
the probability of finding exactly x links, in a given sample of random pages when the events occur independently of one another and page sample at an average rate μ links per sample, is given by Poisson distribution function:

$$f(x;\mu) = \frac{e^{-\mu}\mu^x}{x!}, \text{ where } \overline{x} = \mu \text{ and Var(x)=}\mu$$

- when substituting variance from Poisson statistics, the solution of the system returns the input mean values of rank(a la PageRank™):

$$\begin{bmatrix} 0.65 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.37 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.30 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.17 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \\ 0.37^2 \\ 0.30^2 \\ 0.17^2 \end{bmatrix} \Rightarrow \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65 \\ 0.37 \\ 0.30 \\ 0.17 \end{bmatrix}$$

Some examples

Measured widths, no correlations:

- now, let's use only diagonal elements of the covariant matrix from our exmaple (assuming for simplicity that all w_i are equal) :

$$\begin{bmatrix} 0.043 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.028 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.026 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.01 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \\ 0.37^2 \\ 0.30^2 \\ 0.17^2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ \lambda_2/\lambda_1 \\ \lambda_3/\lambda_1 \\ \lambda_4/\lambda_1 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 0.49 \\ 0.34 \\ 0.42 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 4 \\ 3 \end{matrix}$$

compared to the ratios of the orignal mean PageRanks: $\begin{bmatrix} 1.00 \\ 0.57 \\ 0.45 \\ 0.25 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix}$

- by accounting for real widths, we change the order between pages 3 and 4!

Some examples

Small correlations – 10%:

- by simply adding just 10% of correlations reduces rank of pages 3 and 4 by factor of 2-3,

$$\begin{bmatrix} 0.043 & 0.003 & 0.003 & 0.002 \\ 0.003 & 0.028 & 0.003 & 0.001 \\ 0.003 & 0.003 & 0.026 & 0.001 \\ 0.002 & 0.001 & 0.001 & 0.006 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \\ 0.37^2 \\ 0.30^2 \\ 0.17^2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ \lambda_2/\lambda_1 \\ \lambda_3/\lambda_1 \\ \lambda_4/\lambda_1 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 0.37 \\ 0.18 \\ 0.07 \end{bmatrix}$$

ranks drop!

compared to the ratios of the orignal mean PageRanks: $\begin{bmatrix} 1.00 \\ 0.57 \\ 0.45 \\ 0.25 \end{bmatrix}$

Some examples

Large correlations – 90%:

- if 90% correlations is observed – only first page survives

$$\begin{bmatrix} 0.043 & 0.031 & 0.030 & 0.015 \\ 0.031 & 0.028 & 0.024 & 0.012 \\ 0.030 & 0.024 & 0.026 & 0.012 \\ 0.015 & 0.012 & 0.012 & 0.006 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \\ 0.37^2 \\ 0.30^2 \\ 0.17^2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ \lambda_2/\lambda_1 \\ \lambda_3/\lambda_1 \\ \lambda_4/\lambda_1 \end{bmatrix} = \begin{bmatrix} 1.00 \\ -0.1 \\ -0.4 \\ -1.0 \end{bmatrix}$$

drop out!

compared to the ratios of the orignal mean PageRanks: $\begin{bmatrix} 1.00 \\ 0.57 \\ 0.45 \\ 0.25 \end{bmatrix}$

-can we recover “hopelessly lost” page 4??

Yes!(see next slide)

Some examples

Pages 1,2,3 have large correlations – 90%, page 4 – 10%

- by making page 4 less correlated – we make it recover as #2!

$$\begin{bmatrix} 0.043 & 0.031 & 0.030 & 0.002 \\ 0.031 & 0.028 & 0.024 & 0.001 \\ 0.030 & 0.024 & 0.026 & 0.001 \\ 0.002 & 0.001 & 0.001 & 0.006 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \\ 0.37^2 \\ 0.30^2 \\ 0.17^2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ \lambda_2/\lambda_1 \\ \lambda_3/\lambda_1 \\ \lambda_4/\lambda_1 \end{bmatrix} = \begin{bmatrix} 1.00 \\ -0.4 \\ -0.8 \\ 0.1 \end{bmatrix}$$

recovery!

compared to the ratios of the orignal mean PageRanks: $\begin{bmatrix} 1.00 \\ 0.57 \\ 0.45 \\ 0.25 \end{bmatrix}$

Some examples

Finally, the nominal case:

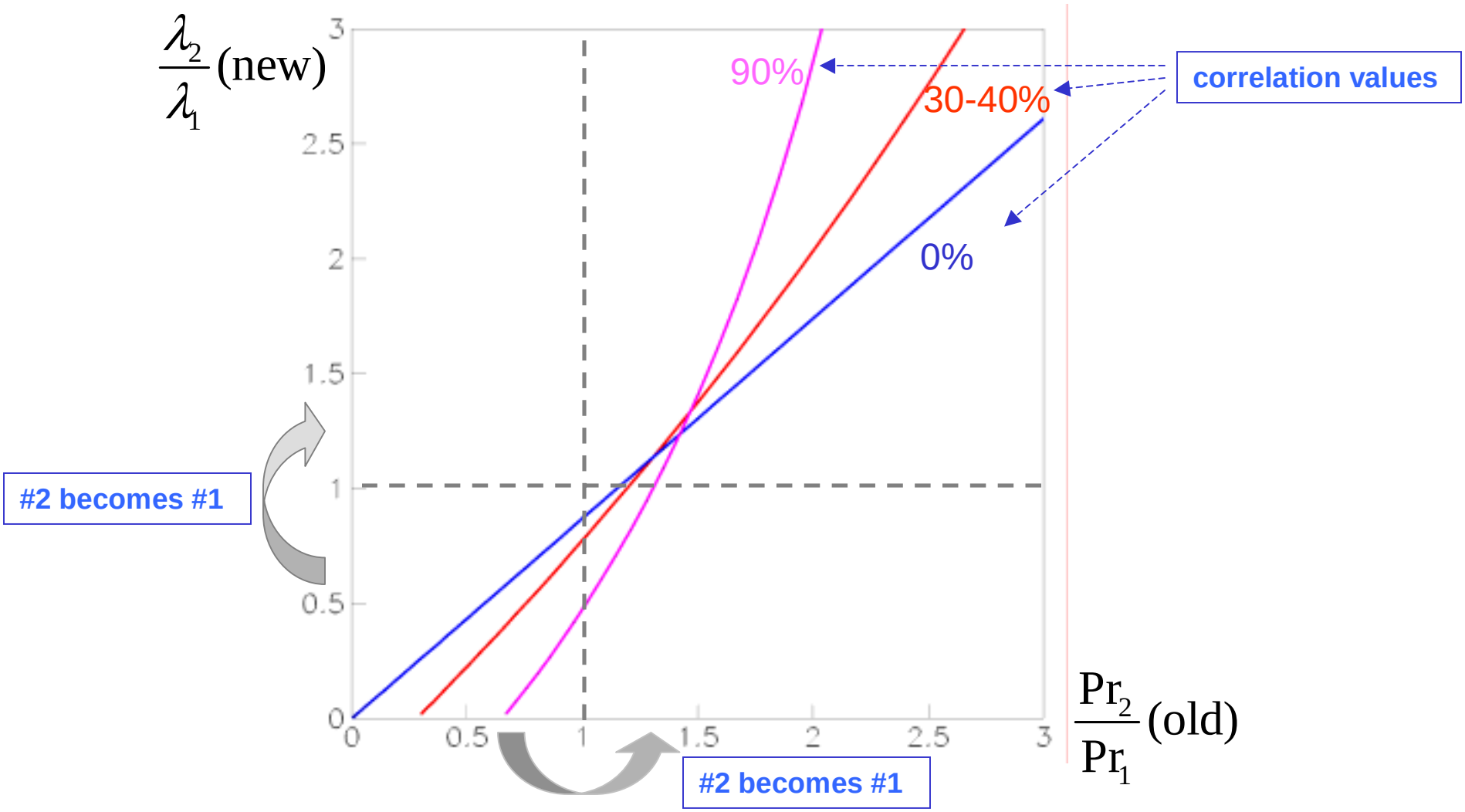
- after plugging the full covariance matrix calculated for our 4 web sites, only first two pages are able to survive:

$$\begin{bmatrix} 0.043 & 0.011 & 0.011 & 0.007 \\ 0.011 & 0.028 & 0.007 & 0.006 \\ 0.011 & 0.007 & 0.026 & 0.004 \\ 0.007 & 0.006 & 0.004 & 0.006 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \\ 0.37^2 \\ 0.30^2 \\ 0.17^2 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ \lambda_2/\lambda_1 \\ \lambda_3/\lambda_1 \\ \lambda_4/\lambda_1 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 0.29 \\ -0.1 \\ -1.0 \end{bmatrix}$$

can't survive!

compared to the ratios of the orignal mean PageRanks: $\begin{bmatrix} 1.00 \\ 0.57 \\ 0.45 \\ 0.25 \end{bmatrix}$

Race for #1



- When correlated, page #2 “follows” page number #1 with a greater rank difference than predicted by PageRank. Once #1 and #2 turn places, their rank difference grows faster than predicted by the ratio of their PageRanks.

Techincal challenges

- Do we have to calculate and store $3x10^9$ by $3x10^9$ rank and content correlation matrices??

No, we don't have to!

can be calculated "on the fly" by some method

for each page, keep page rank values for just 100-300 unique for all WWW samples of 500K-5000K *random* web pages

covariance matrix is calculated, and the characteristic equation is solved "on the fly", and only matched pages(or the top of matched pages)

for pages with small ranks, and not enough statistics, static PageRank™ value can be used

Example 1

Page 1

Sample 1(500k): 65,234 links
Sample 2(500k): 57,567 links
Sample 3(500k): 48,511 links
.
.
Sample 100(500k): 59,092 links

Page 2

Sample 1(500k): 543 links
Sample 2(500k): 573 links
Sample 3(500k): 512 links
.
.
Sample 100(500k): 492 links

Page 3

Sample 1(500k): 5,217 links
Sample 2(500k): 7,614 links
Sample 3(500k): 8,552 links
.
.
Sample 100(500k): 8,280 links

Page 4

Sample 1(500k): 65 links
Sample 2(500k): 56 links
Sample 3(500k): 45 links
.
.
Sample 100(500k): 49 links

$$Cov(x, y) = \sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

$$\begin{bmatrix} 0.043 & 0.011 & 0.011 & 0.007 \\ 0.011 & 0.028 & 0.007 & 0.006 \\ 0.011 & 0.007 & 0.025 & 0.004 \\ 0.007 & 0.006 & 0.004 & 0.006 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \cdot w_1 \\ 0.37^2 \cdot w_2 \\ 0.30^2 \cdot w_3 \\ 0.17^2 \cdot w_4 \end{bmatrix}$$

Example 2

Page 1

	Finance	Travel
Sample 1(500k):	35,234 links	434 links
Sample 2(500k):	27,567 links	567 links
Sample 3(500k):	28,511 links	511 links
.		
.		
Sample 100(500k):	29,092 links	488 links

Page 2

	Finance	Travel
Sample 1(500k):	543 links	43 links
Sample 2(500k):	573 links	73 links
Sample 3(500k):	512 links	51 links
.		
.		
Sample 100(500k):	492 links	65 links

Page 3

	Finance	Travel
Sample 1(500k):	5,217 links	17 links
Sample 2(500k):	7,614 links	14 links
Sample 3(500k):	8,552 links	25 links
.		
.		
Sample 100(500k):	8,280 links	18 links

Page 4

	Finance	Travel
Sample 1(500k):	65 links	5 links
Sample 2(500k):	56 links	4 links
Sample 3(500k):	45 links	7 links
.		
.		
Sample 100(500k):	49 links	9 links

$$Cov(x, y) = \sum_{i=1}^{N} (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

$$\begin{bmatrix} 0.043 & 0.011 & 0.011 & 0.007 \\ 0.011 & 0.028 & 0.007 & 0.006 \\ 0.011 & 0.007 & 0.025 & 0.004 \\ 0.007 & 0.006 & 0.004 & 0.006 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = \begin{bmatrix} 0.65^2 \cdot w_1 \\ 0.37^2 \cdot w_2 \\ 0.30^2 \cdot w_3 \\ 0.17^2 \cdot w_4 \end{bmatrix}$$

Model for topic ranking correction

- What is rank?? - popularity index,
- We saw it's very random over WWW – not much structure out there, quite random,
- WWW – is inefficient, not every body links to pages it need to link to, there is no control "who links to who" – there is much more structure information out there
- URLs make a tiny part of the actual web-page content,
  What's the difference between these two pairs?

Web-Page1	Web-Page2	Web-Page3	Web-Page4
Get travel tips from http://www.travel.com	Get travel tips from http://www.travel.com	Travel tips: ……..	Travel tips: ……..

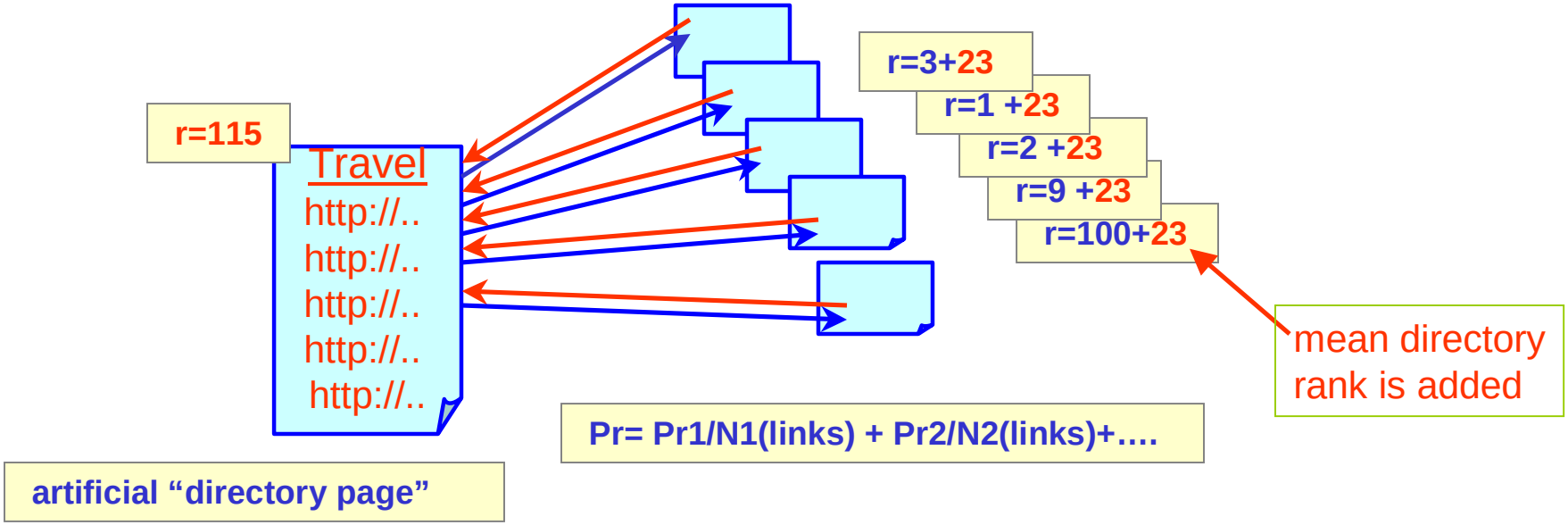
- both refer to travel, but in one case URL explicit(www.travel.com), in another case, It's implicit "Travel tips"

- we need to create new "directory web-pages" which would contain back links to their referees(top of their referees), and use these new pages in the global re-ranking,

- the idea is that different topics(directories) have different popularity index - this is not taken into account at present

- can be thought of as a stock pricing: $price = $index+$sector+$stock

Model for topic ranking correction



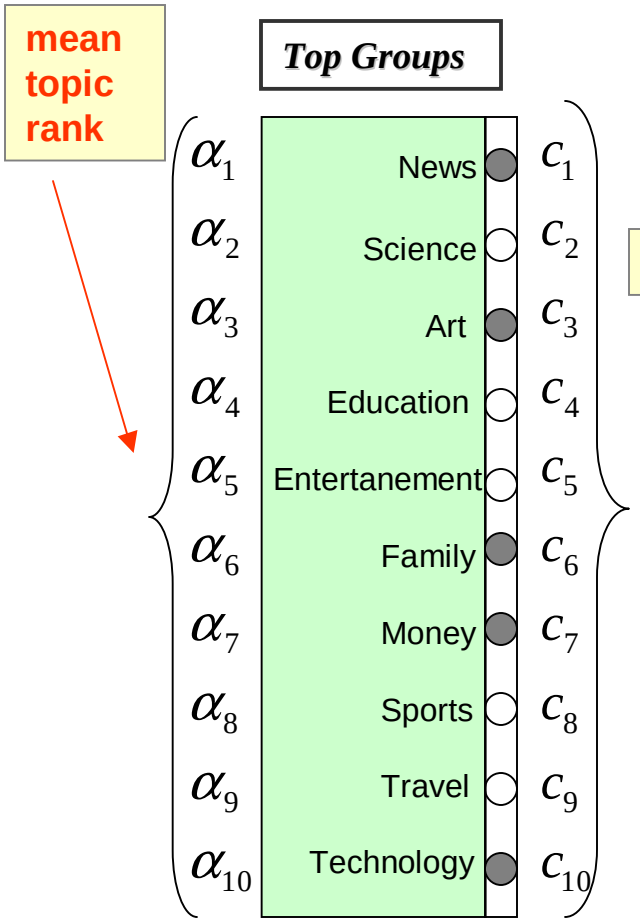
Cool things about directory pages:

1) with 10^9 pages, it won't be difficult to add 1000 new "directory web-pages"(just DB records
2) if search matches directory page topic name – you get an immediate answer
   – just return all the links listed there,
3) if all the content directories have equal popularity, there would be no difference.
4) we will have a way to introduce("1000 control knobs") into the ranking system

Examples:

- one can use "Hot queries list" to dynamically create new "web-directories" and give them higher ranks,
- returned "average directory rank" could be a function of the initial rank (this can be studied and tuned)

Topic enabling/disabling



- instead of worrying about full content-based re-ranking, we could model it analytically, using directory pages model:

$$f_1 = 0$$
$$f_2 = 1$$
$$f_3 = 0$$
$$f_4 = 0$$
$$f_5 = 0$$
$$f_6 = 1$$
$$f_7 = 0$$
$$f_8 = 0$$
$$f_9 = 0$$
$$f_{10} = 1$$

Which topic directory this page belongs

$$r^{page}(\vec{c}) = \left(r^{page}_{mean} + \sum_{i=1}^{Ndir} \alpha_i f_i \Big/ \sum_{i=1}^{Ndir} f_i\right) \cdot \sum_{j=1}^{Ndir} c_j f_j$$

1) scan through all the pages offline, determine which topic(s) it corresponds to(f_i),
2) count total α_i=(directory rank)/n(pages) for each topic directory,
3) for matched pages, calculate their rank on “the fly” using c_i on-off vector state

Monte-Carlo and other ideas

Monte-Carlo simulation of WWW(at more advanced stages!)

- since “rank” doesn’t depend on page content(directly), and has remarkable random behavior, it suggests the possibility of Monte-Carlo simulation of WWW evolution (e.g. instead of calculating ranks for current 10^9 WWW, we could predict future ranks for 10^{10},10^{11} pages sizes, and use them now….) – will allow to always stay one step ahead of competitors(it’s not crazy as it seems at first!)

- reversing all link page-to-page links directions, generation random pages + “mutations”

Many many other ideas(just to name few..):

- we can enforce ranking by counting references to the associated company phone numbers, addresses, product names(Windows == Microsoft),

- using content ranking, create a “joint rank” with Britanica and Library of Congress databases without giving their cites away, and for subscription, give the complete unified index

- …..

Conclusions

- the page relevance is Holy Grail for search engines, and example of Google, and Teoma proves it,

- the content-based page relevance, is even harder problem to address

We address both!

- it was a remarkable, how a simple page ranking idea puts Google's search quality right on the top,

- it's even more remarkable how much they were able to achieve but just "scratching the surface"(no shape information, no second-order effects are considered, only direct links are used for popularity, no content analysis is included into ranking),

- we propose a number of concrete steps, and explain where the new sources of Information are,

- including new information into page relevancy assessment will Inevitably lead(on average) to a better, more comprehensive search results,

Can we make money out of it?

(very subjective view!)

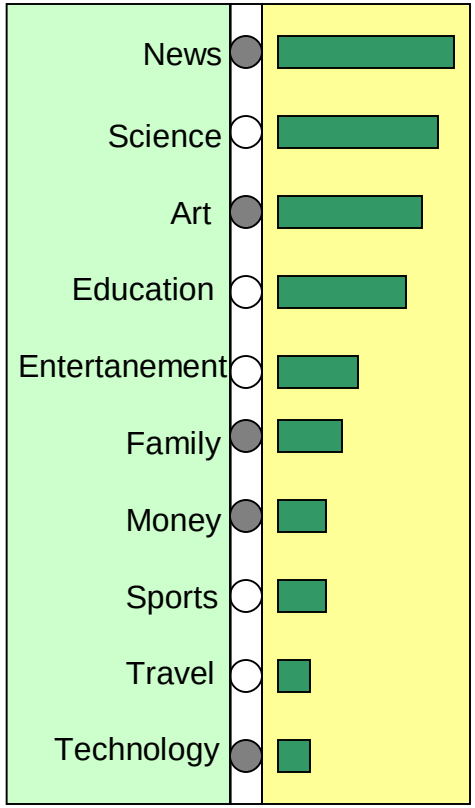
- the whole Google brand-name, the whole business model essentially stands on that single point – PageRank and the quality of the search – and it works!,

- what does takes to challenge them??
  - many good new ideas, top research brains with direct experience in data analyses of huge datasets(I know right people!),
  - expert in parallel computing, few good developers, DBA,
  - disk space is cheap, Linux boxes are cheaper than ever,
  - with statistical and content way of ranking will probably be able to get away without using PageRank™ patent.

- funny, but true – every good idea in page relevancy(ranking) results in a new company: google.com, teoma.com

- there is a new line of “freeware” designed by CERN and SLAC software scientist to handle and analyze huge datasets:
  Root(I/O, compression,object structures), RooFit – most advanced analytical package

Top Groups | **Volume**



On/Off

All Groups

Analyze your cite

Kiiwi™

Best practices

Search

Sponsored links:

Welcome to the **Best Practices** Database

This is a framed site. Please use Browsers that support framed sites.
Description: Searchable database cataloging over 650 good and **best practices** from more than 90 countries on how...
Category: Science > Environment > Sustainability

www.bestpractices.org/ - 2k - Cached - Similar pages

Benton Foundation: Strategic Communications in the Digital Age

... Strategic Communications ... A **best practices** toolkit for achieving
your organization's mission ... in the digital age. Partners **...**

www.benton.org/Practice/Toolkit/ - 8k - Cached - Similar pages

Best Practice Benchmarking, Consulting and Business to Business **...**

Best Practices, LLC, the leading business to business research and consulting firm
specializing in **best practices** benchmarking, provides client organizations **...**
Description: Gateway for executives and specialists seeking to improve their business processes through benchmarking...
Category: Business > Management > Benchmarking and Best Practices

www.best-in-class.com/ - 24k - Cached - Similar pages

Best Practices in Education, a world to learn from

Algebra - Message Board Member login. News from **Best**
Practices in Education Newsletter 2002-2003.
Description: An organization dedicated to working with American teachers to find effective educational **practices**...
Category: Reference > Education > K through 12 > School Improvement

www.bestpraceduc.org/ - 7k - Cached - Similar pages

OIT: BestPractices Home Page

The preceding dropdown menus are not reader accessible, click here
to go the the menu page. Welcome! BestPractices, a program of **...**
Description: An initiative of DOEs Office of Industrial Technologies (OIT), to help manufacturing plants improve...
Category: Science > Technology > Energy > Conservation

www.oit.doe.gov/bestpractices/ - 38k - Cached - Similar pages

Best Manufacturing **Practices**: Your Source for **Best Practices** and **...**

To acheive this goal we: identify **best practices**, document them, and encourage
organizations to share information about them. **Best** Manufacturing **Practices**, **...**
Description: Non-profit portal/source for innovative technologies and **best** manufacturing **practices**. Offers solutions...
Category: Business > Industrial Goods and Services > Consulting

www.bmpcoe.org/ - 43k - Cached - Similar pages

Kiiwi™

Your cite analyzer

fedex.com

Analyze

Overlay average shape

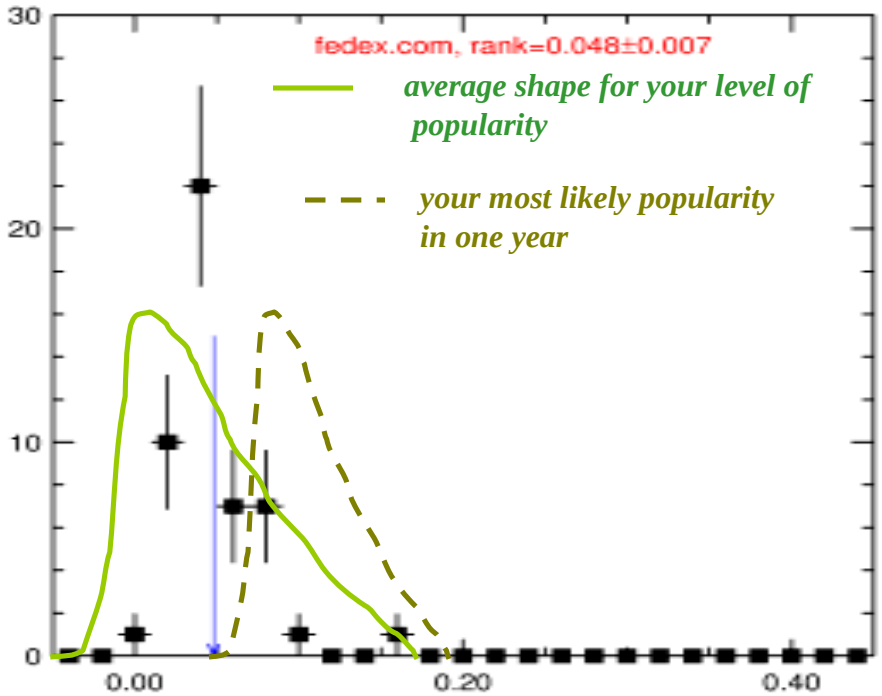
Start history for your site

Study your exposure by regions

Study your exposure by topics

Model the future scenarios for your growth

Sign in



Quick bottom line:

your site acceptance is more uniform than the average