Ranking: The reality, illusion and manipulation of objectivity

Péter Érdi
perdi@kzoo.edu

Henry R. Luce Professor
Center for Complex Systems Studies
Kalamazoo College
http://people.kzoo.edu/perdi/

and

Wigner Research Centre, Hungarian Academy of Sciences, Budapest
http://cneuro.rmki.kfki.hu/
Content

• Comparison, ranking, rating
• Ranking everywhere
• From social sciences to algorithms
• Recursive ranking: from webpages to patents
Comparison, ranking, rating

Social comparison theory states that we determine our own social and personal worth based on how we stack up against others. As a result, we are constantly making self and other evaluations across a variety of domains (for example, attractiveness, wealth, intelligence, and success). Most of us have the social skills and impulse control to keep our envy and social comparisons quiet but our true feelings may come out in subtle ways.

Types of Social Comparison

- Upward Social Comparison
  - You > Me
  - “You are better…”
- Downward Social Comparison
  - Me > You
  - “I feel sorry for you…”

comparison is the thief of joy
— theodore roosevelt —
**Ranking** is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. Ranking procedure generates a rank-ordered list of elements. **Rating** gives a list of numerical scores, one for each object. A rating list, when sorted, generates a ranking list.

<table>
<thead>
<tr>
<th>Score</th>
<th>Grading</th>
</tr>
</thead>
<tbody>
<tr>
<td>720 and Above</td>
<td>Excellent</td>
</tr>
<tr>
<td>680 to 719</td>
<td>Good</td>
</tr>
<tr>
<td>620 to 679</td>
<td>Average</td>
</tr>
<tr>
<td>580 to 619</td>
<td>Poor</td>
</tr>
<tr>
<td>500 to 579</td>
<td>Bad</td>
</tr>
<tr>
<td>Less than 500</td>
<td>Miserable</td>
</tr>
</tbody>
</table>

- 1. The Beatles
- 2. The Rolling Stones
- 3. Led Zeppelin
- 4. Pink Floyd
- 5. Queen
Ranking everywhere

- pecking order
- social ranking
- universities
- countries

### 2017 Best Countries Rankings

*See the full rankings [here](#)

#### Overall

1. Switzerland
2. Canada
3. United Kingdom
4. Germany
5. Japan
6. Sweden
7. United States
8. Australia
9. France
10. Norway
• one of the largest databases of opinions with more than 250 million votes gathered on over one million items.
• "wisdom of crowds" lists
• How does ranker work?
• editorial team creates rankings
• visitors can vote and rerank all the items, and add items they think deserve a spot on the list.
From social sciences to algorithms
From social sciences to algorithms

Challenges:

• the algorithms behind individual human ranking procedures
• the algorithms behind collective human ranking procedures
• the algorithms behind computational ranking procedures
From social sciences to algorithms

- Voting Protocols
- plurality protocol
- binary protocol
- Borda protocol
- Condorcet protocol

(from Antonio Moreno:
https://www.slideshare.net/ToniMorenoURV/ mas-course-lect-9)

Basic elements

- Aim of the negotiation: rank feasible social outcomes based on the individual ranking of those outcomes by each agent
  - A - set of $n$ agents
  - O - set of $m$ feasible outcomes
  - Each agent $i$ has a preference relation
    $\prec_i : O \times O$, asymmetric and transitive

Voting mechanisms

- Some voting mechanisms:
  - Plurality protocol (highest number of votes wins)
  - Binary protocol (series of votes of 2 options each)
  - Borda protocol (sum of all the preferences of the agents)
  - Condorcet protocol (pairwise comparison of options, given full preference ordering of each agent)

- All the protocols are problematic in one sense or another
From social sciences to algorithms

- Voting Protocols
- plurality protocol
- binary protocol
- Borda protocol
- Condorcet protocol
- (from Antonio Moreno)

**Plurality protocol**

- Each agent can give 1 vote to 1 of the alternatives
- The alternative with the highest number of votes wins

**Binary protocol**

- All the options are ordered and then evaluated in pairs (options 1 and 2, the winner with option 3, the winner with option 4, etc.)
- The option that wins the last evaluation is the overall winner
  \[
  \text{win}(o_5, \text{win}(o_4, \text{win}(o_3, \text{win}(o_2, o_1))))
  \]
From social sciences to algorithms

- Voting Protocols
- plurality protocol
- binary protocol
- Borda protocol
- Condorcet protocol
- (from Antonio Moreno)

**Borda protocol-1770**
- For each voter, we assign \(|O|\) points for the highest preference, \(|O|-1\) points for the second, and so on
- Example: one agent considers the order ABCD
  - A:4 points, B:3 points, C:2 points, D:1 point
- The points are added across the voters and the alternative with the highest count becomes the social choice [similar to Eurovision song contest]

**Condorcet protocol-s.XVIII**
- Each voter ranks the candidates in order of preference
- Each candidate is compared to each other
- If a candidate wins all the comparisons, it is the winner of the election
- In the event of a tie, use another resolution method (e.g. Borda count)
Reddit: social network aggregation and web content rating website open sourced and the code is freely available

**Hot**

\[ rating = \log_{10}(z) + \frac{yt}{45000} \]  

where \( y \) is a signum function based on the difference of upvotes and downvotes and \( t \) is the time difference between the `post.created` and `Time.now`

\( \log \): The first 10 upvotes have the same weight as the next 100 upvotes

45000sec -> 12 hours

**Best** is the highest upvote to downvote ratio, **top** is the most upvotes regardless of downvotes, and **hot** is the most upvotes recently.
Recursive ranking: from webpages to patents

- Centrality measures
- PageRank
- Rank Reversal
- Citation networks: structure and age
- Ranking Patents: dependence on the damping factor
Centrality measures

Ranking of nodes in a network of diverse number of connections (‘degree’) is an extensively studied field. In the theory of social networks centrality measures were constructed to rank nodes of networks based on their (not unique) topological importance.

- degree
- ”betweenness”
- ”closeness”
- ”eigenvalue centrality’
Centrality measures

Social Networks
Social Network Analysis

Degree: the number of direct connections a node has. In the network above, Diane has the most direct connections in the network. She is a 'connector' or 'hub' in this network.

\[ C_d(v) := \text{deg}(v) \quad (2) \]
Centrality measures

Social Networks

Social Network Analysis

Betweenness: While Diane has many direct ties, Heather has few direct connections – fewer than the average in the network. Yet, in many ways, she has one of the best locations in the network – she is between two important constituencies. She plays a ‘broker’ role in the network.

- For each pair of vertices \((s, t)\), compute the shortest paths between them.
- For each pair of vertices \((s, t)\), determine the fraction of shortest paths that pass through the vertex in question.
- Sum this fraction over all pairs of vertices \((s, t)\).
Closeness: Fernando and Garth have fewer connections than Diane, yet the pattern of their direct and indirect ties allow them to access all the nodes in the network more quickly than anyone else. They have the shortest paths to all others – they are close to everyone else.

\[
C(x) := \frac{1}{\sum_y d(y, x)} \tag{3}
\]
Centrality measures

- "eigen-vector centrality": measure of the influence of a node in a network
- adjacency matrix: $A$ where $A_{ij} = 1$ if there is a connection between node $i$ and node $j$ and $A_{ij} = 0$ if there is no connection
- relative scores to all nodes (high-scoring nodes are "more important", so contribute more)
- Importance can be defined recursively

\[ A \mathbf{x} = \lambda \mathbf{x}, \quad (4) \]

- generally $\lambda$ is multiple
- Perron-Frobenius theorem helps to make the solution of the problem unique
- a real square matrix with positive entries has a unique largest real eigenvalue
- the dominant eigen-vector can be found by the power iteration method
- extension for stochastic matrices: elements of $A$ are real numbers ("connection strengths")
- Brin and Page introduced a specific recursive centrality measure: PageRank
In a network with $n$ nodes, we assign all nodes the same initial PageRank, set to be $1/n$.

We choose a number of steps $k$.

We then perform a sequence of $k$ updates to the PageRank values, using the following rule for each update (redistribution process):

**Redistribution process** The PageRank values of the nodes are calculated iteratively using the following equation:

$$P_i^{(t+1)} = \frac{(1 - d)}{N} + d \sum_{j=1}^{n_i} \frac{P_j^{(t)}}{n_i}$$

where $P_i^{(t)}$ is the PageRank value of node $i$ at iteration $t$, $n_i$ is the number of nodes that link to node $i$, and $d$ is called the damping factor.
PageRank: Algorithm

more about the uniqueness of the convergence:


"Now, depending on the network structure, the set of limiting values may not be the only ones that exhibit this kind of equilibrium. However, one can show that if the network is strongly connected - that is, each node can reach each other node by a directed path, following the definition from Chapter 13 - then there is a unique set of equilibrium values, and so whenever the limiting PageRank values exist, they are the only values that satisfy this equilibrium."
some further readings:

- PageRank Algorithm - The Mathematics of Google Search  
- Linear Algebra - in a Nutshell  
  http://www.math.cornell.edu/ mec/Winter2009/RalucaRemus/Lecture1/lecture
- Directed Graphs - Transition Matrices  

(adjacency matrix, primitive, connected, strongly connected, indegree, outdegree, transition matrix, weighted graph.

Consequences:

- For a strongly connected graph, the transition matrix is column-stochastic
- If the matrix is primitive, column-stochastic, then this process converges to a unique stationary probability distribution vector $p$.
- The PageRank vector for a web graph with transition matrix $A$, and damping factor $p$, is the unique probabilistic eigen-vector of the matrix $M$, corresponding to the eigenvalue $1$
The role of the damping factor


The choice of the PageRank damping factor is not evident. The Google’s choice for the value $c = 0.85$ was a compromise between the true reflection of the Web structure and numerical efficiency. However, the Markov random walk on the original Web Graph does not reflect the importance of the pages because it absorbs in dead ends. Thus, the damping factor is needed not only for speeding up the computations but also for establishing a fair ranking of pages. In this paper, we propose new criteria for choosing the damping factor, based on the ergodic structure of the Web Graph and probability flows. Specifically, we require that the core component receives a fair share of the PageRank mass. Using singular perturbation approach we conclude that the value $c = 0.85$ is too high and suggest that the damping factor should be chosen around $1/2$. As a by-product, we describe the ergodic structure of the OUT component of the Web Graph in detail. Our analytical results are confirmed by experiments on two large samples of the Web Graph.
PageRank: Damping Factor

The role of the damping factor: Paolo Boldi: TotalRank: Ranking Without Damping

PageRank is defined as the stationary state of a Markov chain obtained by perturbing the transition matrix of a web graph with a damping factor $\alpha$ that spreads part of the rank. The choice of $\alpha$ is eminently empirical, but most applications use $\alpha = 0.85$; nonetheless, the selection of $\alpha$ is critical, and some believe that link farms may use this choice adversarially. Recent results prove that the PageRank of a page is a rational function of $\alpha$, and that this function can be approximated quite efficiently: this fact can be used to define a new form of ranking, TotalRank, that averages PageRanks over all possible $\alpha$'s. $T := \int_0^1 r(\alpha) d\alpha$.

Different ranking procedures give different results and there is no "magic bullet"!
PageRank (PR) is an algorithm originally developed by Google to evaluate the importance of web pages. Considering how deeply rooted Google’s PR algorithm is to gathering relevant information or to the success of modern businesses, the question of rank-stability and choice of the damping factor (a parameter in the algorithm) is clearly important. We investigate PR as a function of the damping factor $d$ on a network obtained from a domain of the World Wide Web, finding that rank-reversal happens frequently over a broad range of PR (and of $d$). We use three different correlation measures, Pearson, Spearman, and Kendall, to study rank-reversal as $d$ changes, and show that the correlation of PR vectors drops rapidly as $d$ changes from its frequently cited value, $d_0 = 0.85$. Rank-reversal is also observed by measuring the Spearman and Kendall rank correlation, which evaluate relative ranks rather than absolute PR. Rank-reversal happens not only in directed networks containing ranksinks but also in a single strongly connected component, which by definition does not contain any sinks. We relate rank-reversals to rank-pockets and bottlenecks in the directed network structure. For the network studied, the relative rank is more stable by our measures around $d = 0.65$ than at $d = d_0$. 

Rank Reversal

Seung-Woo Son, Claire Christensen, Peter Grassberger and Maya Paczuski: PageRank and rank-reversal dependence on the damping factor
Phys. Rev. E 86, 066104 - Published 5 December 2012
The number of citations is the most commonly used metric for quantifying the importance of scientific publications. However, we all have anecdotal experiences that citations alone do not characterize the importance of a publication. Some of the shortcomings of using citations as a universal measure of importance include the following.

(1) It ignores the importance of citing papers: a citation from an obscure paper is given the same weight as a citation from a ground-breaking and highly cited work.
(2) The number of citations is ill suited to compare the impact of papers from different scientific fields. Due to factors such as size of a field and disparate citation practices, the average number of citations per paper varies widely between disciplines. An average paper is cited cca 6 times in life sciences, 3 times in physics, and < 1 times in mathematics.
(3) Many groundbreaking older articles are modestly cited due to a smaller scientific community when they were published. Furthermore, publications on significant discoveries often stop accruing citations once their results are incorporated into textbooks. Thus, citations consistently underestimate the importance of influential old papers.

These and related shortcomings of citation numbers are partially obviated by Google's PageRank algorithm (Brin and Page, 1998). As we shall discuss, PageRank gives higher weight to publications that are cited by important papers and also weights citations more highly from papers with few references. Because of these attributes, PageRank readily identifies a large number of scientific "gems": modestly cited articles that contain ground-breaking results.
P. Bruck, I. Réthy, J. Szente, J. Tobochnik and P Érdi:

Recognition of emerging technology trends: class-selective study of citations in the U.S. Patent Citation Network.


The USPTO granted approximately 4.5 million patents in the 1976 – 2012 period; its citation network has 44 million edges,
Ranking Patents: dependence on the damping factor

Patent classes of top ranking patents in the USPTO database
Table displays the number of citations (NCIT) and the corresponding PageRank values of the top 20 patents for specific values of $d$. We also indicate which citations originate from the two most important patent classes: Class 435, related to genetics, is marked by green, while Class 347, corresponding to inkjet and laser printer technology, is marked by yellow.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCIT</td>
<td>iterations:4</td>
<td>iterations:10</td>
</tr>
<tr>
<td>1</td>
<td>2361</td>
<td>5,363E-07</td>
<td>2361</td>
</tr>
<tr>
<td>2</td>
<td>2688</td>
<td>8,073E-07</td>
<td>2688</td>
</tr>
<tr>
<td>3</td>
<td>1705</td>
<td>6,835E-07</td>
<td>1705</td>
</tr>
<tr>
<td>4</td>
<td>442</td>
<td>8,095E-07</td>
<td>442</td>
</tr>
<tr>
<td>5</td>
<td>2016</td>
<td>6,109E-07</td>
<td>2016</td>
</tr>
<tr>
<td>6</td>
<td>891</td>
<td>9,433E-07</td>
<td>891</td>
</tr>
<tr>
<td>7</td>
<td>885</td>
<td>5,410E-07</td>
<td>885</td>
</tr>
<tr>
<td>8</td>
<td>875</td>
<td>5,327E-07</td>
<td>875</td>
</tr>
<tr>
<td>9</td>
<td>180</td>
<td>5,291E-07</td>
<td>180</td>
</tr>
<tr>
<td>10</td>
<td>1714</td>
<td>5,241E-07</td>
<td>1714</td>
</tr>
<tr>
<td>11</td>
<td>1731</td>
<td>5,253E-07</td>
<td>1731</td>
</tr>
<tr>
<td>12</td>
<td>1581</td>
<td>4,949E-07</td>
<td>1581</td>
</tr>
<tr>
<td>13</td>
<td>1606</td>
<td>4,494E-07</td>
<td>1606</td>
</tr>
<tr>
<td>14</td>
<td>1248</td>
<td>4,800E-07</td>
<td>1248</td>
</tr>
<tr>
<td>15</td>
<td>1546</td>
<td>4,812E-07</td>
<td>1546</td>
</tr>
<tr>
<td>16</td>
<td>1507</td>
<td>4,725E-07</td>
<td>1507</td>
</tr>
<tr>
<td>17</td>
<td>982</td>
<td>4,424E-07</td>
<td>982</td>
</tr>
<tr>
<td>18</td>
<td>896</td>
<td>4,360E-07</td>
<td>896</td>
</tr>
<tr>
<td>19</td>
<td>1441</td>
<td>4,288E-07</td>
<td>1441</td>
</tr>
<tr>
<td>20</td>
<td>344</td>
<td>4,268E-07</td>
<td>344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum$</td>
<td>26691</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum$</td>
<td>26210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum$</td>
<td>21782</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum$</td>
<td>7089</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum$</td>
<td>2890</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum$</td>
<td>14758</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Data of the top 20 USPTO patents, ranked by descending PageRank values (d=0.50)
Ranking Patents: dependence on the damping factor

**Rank change** new results
How do the ranks change with the increase of $d$ in the USPTO network?
Thanks to P. Bruck and G. Fenyvesi

First two patents for $d < 0.5$ : Kary Mullis (\#4683195 and \#4683202):
how to prepare large quantities of any DNA sequence from a single template using DNA polymerase of a bacterium found in a Yellowstone spring, which can withstand 94°C temperature; this work was awarded with the Nobel prize in 1993.
Ranking Patents: dependence on the damping factor
<table>
<thead>
<tr>
<th>Level</th>
<th>#4683195</th>
<th>Level</th>
<th>#3988545</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,360</td>
<td>1</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>9,822</td>
<td>2</td>
<td>2,085</td>
</tr>
<tr>
<td>3</td>
<td>17,884</td>
<td>3</td>
<td>29,892</td>
</tr>
<tr>
<td>4</td>
<td>24,474</td>
<td>4</td>
<td>142,936</td>
</tr>
<tr>
<td>5</td>
<td>29,147</td>
<td>5</td>
<td>259,764</td>
</tr>
<tr>
<td>6</td>
<td>27,166</td>
<td>6</td>
<td>252,168</td>
</tr>
<tr>
<td>7</td>
<td>18,936</td>
<td>7</td>
<td>219,504</td>
</tr>
<tr>
<td>8</td>
<td>10,027</td>
<td>8</td>
<td>176,598</td>
</tr>
<tr>
<td>9</td>
<td>4,837</td>
<td>9</td>
<td>126,453</td>
</tr>
<tr>
<td>10</td>
<td>2,361</td>
<td>10</td>
<td>78,932</td>
</tr>
<tr>
<td>11</td>
<td>949</td>
<td>11</td>
<td>43,439</td>
</tr>
<tr>
<td>12</td>
<td>330</td>
<td>12</td>
<td>22,180</td>
</tr>
<tr>
<td>13</td>
<td>106</td>
<td>13</td>
<td>10,542</td>
</tr>
<tr>
<td>14</td>
<td>37</td>
<td>14</td>
<td>4,823</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>15</td>
<td>1,826</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>16</td>
<td>529</td>
</tr>
</tbody>
</table>

**Tree A**

The citation tree of node #4683195

**Tree B**

The citation tree of node #3988545
Though it has an outstanding amount of inlinks (2360), with the increase of $d$ the effect of the distant nodes gradually increases; over $d=0.5$ the intercrossing of the turquoise and the violet lines (R1) indicate a rank change, and after the reversal the rank of the two initially "super-stable" nodes starts a steep descent. It is important to understand that the reason of this descent is not the decrease of the absolute contribution of the super-stable nodes, but the increase of the influence of other, distant nodes as $d$ grows - as shown by the violet and the green lines. Finally at $d = 0.8$ another rank change occurs and node #3988545 going to receive the highest PageRank; this node has only 59 inlinks at the root level.
The value of damping factor \(d\) determines, whether the adjacent neighbors or the remote nodes have larger influence on the PageRank of a node. At low values of \(d\) the influence of the adjacent nodes dominate; here Pagerank is a good measure of centrality and shows good correlation with the number of inlinks. As \(d\) increases, this correlation suddenly comes to an end and the number of inlinks of the rank one node becomes surprisingly low; here the value of PageRank is determined primarily by the contribution of the remote nodes.
Manipulation of objectivity

A Google bomb is an attempt to bias a search result on Google by increasing a Web page’s PageRank algorithm not only considers keywords when determining PageRank, it also considers the number of other sites that link to a page and the words used in the link description. A successful Google bomb requires that a high number of Web sites (often blogs working together) link to a specific Web page using specific words in the hyperlink. Enlisting high-traffic sites is especially important when creating a Google bomb, as their outgoing hyperlinks weigh more in Google’s search algorithm. In contrast, search engine optimization (SEO) is the practice of improving the search engine listings of web pages for relevant search terms.
"Link farming is a questionable search engine optimization (SEO) tactic in which a website enters a relationship of reciprocal linking with another site or pays a provider for a large number of inbound links. Because many search engines use inbound links as a factor in determining a website’s page rank, websites use link farming to boost their rankings in search engine results."

Manipulation of objectivity
Welcome to the site of Ranking

I decided to write a non-fiction book with the title and subtitle RANKING – The reality, illusion and manipulation of objectivity. The book discusses the Hows and Whys of our love and fear of making ranks and being ranked through many real life examples to be viewed from three different angles (reality, illusion and manipulation) of objectivity. Ranking converts scientific theories to everyday's experience by raising and answering such question as:

- Are college ranking lists objective?
- How to rank and rate states based on their fragility, corruption or even happiness?
- How to find the most relevant web pages?
- How to rank employees?

Life and society is really complex, consequently our message is not so simple such as "Ranking is good!" or "Ranking is bad!". Since we permanently rank ourselves and others and are also being ranked, the message is twofold: how to prepare the possible most objective ranking and how to accept that ranking does not necessarily reflects our real values and achievements. The reader will understand our difficulties to