How Rankings Go Wrong: Structural Bias in Common Ranking Systems Viewed as Complex Systems
work in progress

Patrick Grim, Natalia Jenuwine, Jared Stolove, Adrian Apaza, Hannah vanWingen, Jaikishan Prasad, Paulina Knoblock, Callum Hutchinson, Chengxi Li, Kyle Fitzpatrick, Chang Xu & Catherine Ming

Center for the Study of Complex Systems, University of Michigan
Ann Arbor, MI 48109
pgrim@umich.edu

Abstract. We introduce agent-based techniques to analyze inherent structural bias in abstract models of common ranking systems such as PageRank, HITs, and Reddit. In the complex dynamics of reputational loops, an element's ranking itself influences factors in terms of which rank is calculated, resulting in the amplification of divergence and the exaggeration of small random and path-dependent differences. Agent-based models of basic algorithms employed in PageRank, HITs, and Reddit are constructed. The hope is that with further development such models will allow comparisons of bias dynamics and effects across a number of parameters.

Keywords: Ranking Systems, Bias, Complex Systems

1 Introduction

In many real world examples, such as college ranking and online search, objects with very similar quality can end up with significantly different rank [1-4]. Sometimes the data that a ranking system relies on may be dubious [5]. But even in the best of conditions and with the cleanest data input, we argue, the very structure of some familiar ranking systems can result in distorted informational output (see also [6]).

The basic idea of all familiar ranking systems is an attempt to read objective quality—what sites, papers, or posts are genuinely worth reading—from social measures of what is read and responded to by whom. In any such system there will be a looping factor: a site, paper, or post will be widely read because it is ranked highly, but will be ranked highly precisely because it is widely read. How severe the loop—and how far rank will come unglued from quality—will depend on the degree to which users base their choice of a site, a paper, or a post on rank as opposed to some independent judgment of quality.

The extent to which looping constitutes a distorting factor, however, will also differ with different ranking algorithms. What we offer here is an initial review of

results regarding approaches with an incomplete sample of results regarding three familiar algorithms—PageRank, HITS and Reddit. This remains a work in progress: a more complete paper with a fuller development of techniques and results will appear elsewhere.

2 PageRank

PageRank analyzes the structure of links between websites in order to determine their ranking [7- 8]. Using a network structure, PageRank treats each website as a node and hyperlinks between websites as directed links between the nodes. Each site is initially assigned a rank value of 1 divided by the total number of nodes. At each time interval, PageRank divides the rank value for each node by the number of outgoing links from that node, and this value is sent as an incoming value to the node at the other end of each outgoing link. Each node's value is then replaced by the sum of its incoming values. PageRank also redistributes an extremely small amount of value equally between the nodes regardless of links, representing the possibility of an individual typing a URL directly into the search bar rather than clicking a link. PageRank then ranks the sites in order of greatest to least node value.

We construct a series of simulations in which pages are assigned an inherent 'quality' between 1 and 100. At each step in the evolution of the model, a small number of pages are added to the network—much as pages are progressively added on the internet, and roughly as nodes are added in a preferential attachment network—and then each page may create additional links to other pages, with probabilities based on the receiving page's inherent quality and current rank.

As noted, the basic idea of all such ranking systems is an attempt to read objective quality from social measures. In the model we can track how well rank corresponds to inherent quality at different settings—in particular, as a function of the degree to which links are formed in terms of a measure of inherent quality or in terms of already-established rank.

The relationship between rank and quality changes significantly as the proportion of link formation determined by rank increases. Figure 1 tracks this relationship at four different levels of the influence given to rank. When links are determined by quality alone, as shown in the upper left panel, the most highly-ranked pages all have relatively high quality. While many high-quality pages end up with low rank, low-quality pages cannot receive rank above a certain quantity. Also, the spread in rank between pages is moderate, with the most successful pages receiving around 3.5 times the average amount of rank. As the influence of rank on link formation is increased, as shown in the following panels, lower quality pages are able to obtain a higher rank; fewer pages have high rank; and the difference between the highest and lowest ranked pages becomes much larger.

Figure 1. Rank vs. quality in PageRank as the probability of link formation is calculated in terms of quality alone, rank 30%, 70% and entirely in terms of rank. Rescaling of the y-axis should be noted.

We introduce 'discrepancy' as a measure for divergence of rank from quality. Discrepancy represents the root mean squared error of rank as a measure for quality, where both are scaled to sum to 1 across all pages. The mean and standard deviation of discrepancy as more weight is given to rank are shown in Figure 2. Discrepancy increases first gradually and then rapidly as the proportion of influence given to rank approaches 1. The standard deviation also increases, suggesting that when users base decisions strongly on rank, the system's ability to approximate quality is subject to large random variation.

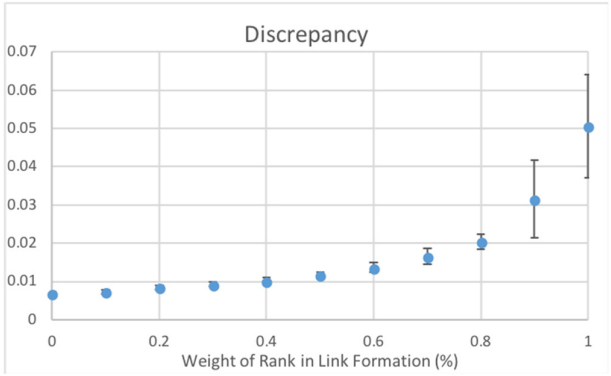

Figure 2. Weight of attention to rank as opposed to quality in link formation and the resultant discrepancy between modelled rank and quality of PageRank sites.

We also track average quality of a page linked to – that is, average quality of the receiving page across all links – as a proxy for the quality that a surfer of the resulting web structure might experience. As shown in Figure 3, this measure is around 65 when links are based mainly on quality, and starts decreasing once weight given to rank exceeds around 50%. With links based only on rank, it reaches 50 – what we would expect if surfers selected pages randomly.

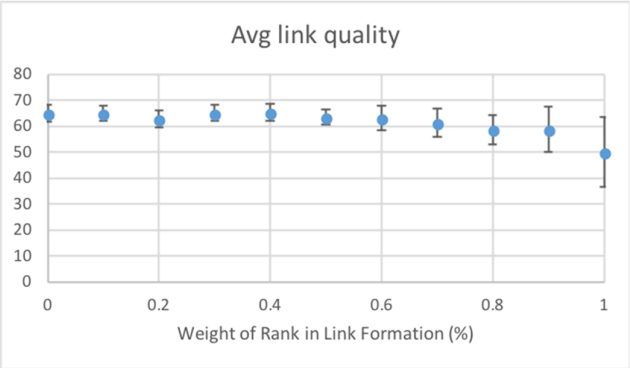

Figure 3. Weight of attention to rank as opposed to quality in link formation and the resultant average quality of the page receiving each link.

As agents respond more to rank—which we target as the more realistic case—the discrepancy between rank and quality is higher, fewer pages dominate the network, and the average quality of a page linked to is lower.

2 HITS

The HITS Algorithm [9-10] is used by Academia.edu, a social networking website that shares papers and monitors their impact. In that academic instantiation, the algorithm assigns values to papers and to authors in mutual recursion. Papers are assigned PaperRanks on the basis of the number of recommendations they receive from authors, weighted by the AuthorRank of those authors. AuthorRank is calculated on the basis of the PaperRanks of that author's own papers.[1]

In our model, both papers are assigned a constant built-in 'quality value.' During setup, authors decide which papers to recommend based solely on paper quality. Subsequently, papers are scored on the prestige and number of authors that have recommended them, while authors are given AuthorRanks on the basis of the PaperRanks of their papers. Each iteration, authors recommend new papers with probabilities based on each paper's inherent quality and current PaperRank.

[1] The instantion of HITS in Academia.edu differs from a wider version of HITS in which the equivalent of AuthorRank is calculated not on the basis of the PaperRank of their own papers but o the basis of the PaperRank of the papers an author recommends.

Here, as in the case of the PageRank model, we can track how well PaperRank corresponds to a paper's inherent quality at different settings. In Figure 4 results are shown at the same levels of influence given to rank.

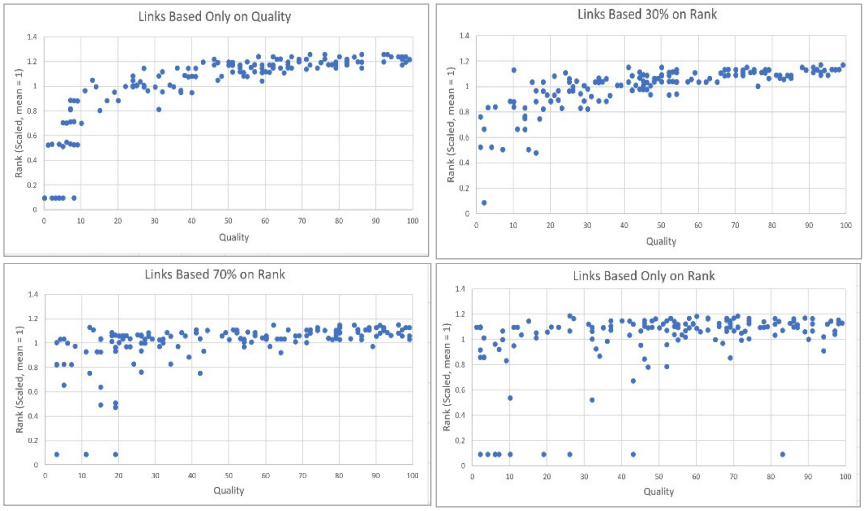

Figure 4. Rank vs. quality in HITS as the probability of link formation is calculated in terms of quality alone, rank 30%, 70% and entirely in terms of rank.

Here again it is clear, as expected, that the relationship between rank and quality decays as modeled users attend increasingly to rank. In other regards comparison of Figures 1 and 4 shows that the algorithms behave quite differently. PageRank allocates a majority of rank to a handful of highly successful pages. The degree of inequality increases as users act based increasingly on rank, indicated by the change in y-axis scaling in Figure 1. HITS distributes rank more evenly between many papers, and the overall distribution does not change when users attend to rank, allowing for the same scaling throughout. However, as the level of rank-influence increases, rank in HITS tends to approach the mean for most papers, regardless of quality, while more low-quality papers rise to the top and more high-quality papers drop below the average.

3 Reddit

Reddit.com is a social news aggregation site where users can discuss and rate posted content. Content is distributed by topic among various ‘subreddits’ where posts with the highest net rankings (upvotes minus downvotes) rise to the top of the page. By ranking based on net score, Reddit is intended to create a preference toward non-controversial content, since a post which receives 50 upvotes and zero downvotes will be ranked the same as a post with 500 upvotes and 450 downvotes. Comments on posts are also voted upon and change order correspondingly. Reddit’s main Front Page shows

posts with the most upvotes across all subreddits, with the order of links changing constantly based both on the time of submission and user votes. A post's score will not decrease as time passes but newer posts will get a higher score.

Our model of Reddit simulates website users reading posts and voting on them. We model users as deciding to read a post and to give it an upvote or downvote based on two factors: its objective 'quality' (a number between 0 and 100) and its rank as determined by the Reddit algorithm. Each modeled user is assigned a threshold that determines how likely they are to read, upvote, or downvote a given post on average. During the simulation, users periodically "leave" the site, replaced by users with different thresholds. New posts with assigned 'qualities' are periodically created as well, then ranked according to Reddit's algorithm.

In order to get a baseline measure to help us evaluate the algorithm, we start off with two highly unrealistic situations: a 'best case scenario' for both reading and voting and a 'worst case scenario.' The best case scenario considers a situation in which users are able to tell the objective quality of a post before even reading it, and make their decisions to read and vote based solely on quality. The worst case scenario describes a situation in which users read and vote on posts based solely in terms of rank. In that case rank is totally unrelated to quality The goal of any ranking algorithm is to offer users posts of high quality. We evaluate how well Reddit performs by an 'average quality score'—the average quality of a post read by a user during the simulation. Figure 5 shows the distribution of the average quality score over 100 instances of the simulation, under both the worst case and best case scenarios. The worst case scenario depicts an average quality score distributed approximately normally around 50, suggesting that user's reading habits are totally unrelated to quality. In contrast, the best case scenario depicts a distribution centered at a quality score of 75, the approximate mean value of user's reading thresholds.

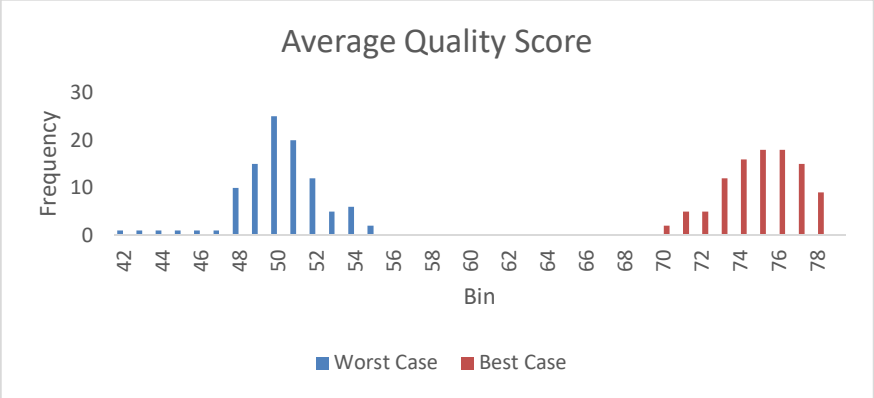

Figure 5. The distribution of average quality of a read post, or "average quality score", in both the best and worst case scenarios over 100 runs of the model.

We then evaluate how Reddit performs given users who base their reading and voting more on quality or more on rank. Figure 6 shows an evaluation of the Reddit algorithm's robustness to rank-bias based on how close the average score is to the best case scenario when we introduce varying levels at which reading and voting are based on rank.

		Voting Bias					
		0	0.2	0.4	0.6	0.8	1
Reading Bias	0	74.9	74.399	74.774	74.26	72.844	72.33
	0.2	73.51	72.84	73.477	74.62	73.94	71.78
	0.4	73.73	72.21	75.16	71.22	72.17	70.82
	0.6	70.97	71.057	72.105	71.033	71.466	68.788
	0.8	69.561	68.34	68.865	67.068	70.044	60.78
	1	68.3827	66.83	65.77	65.98	63.01	49.025

Figure 6. The average quality of a post read in versions of the model with varying levels of Reading Bias and Voting Bias. Cells that are relatively closer to the best case scenario are colored green, those relatively closer to the worst case scenario are colored red.

In our model, performs quite well even at relatively high levels of both kinds of bias. Even with both bias parameters set to .6 rank, the resulting average quality score is 71.033, much closer to the ideal scenario than to the worst case scenario.

Reddit's robustness in the face of rank-related bias appears to be due to the built-in penalization of older posts in the ranking algorithm. Reddit's algorithm provides newer posts with an advantage in two ways. First, newer posts are given a one-time increase in their initial raw score, called the 'time damp.' Second, additional upvotes only impact a post's score logarithmically, so highly ranked posts must acquire exponentially more votes to counteract the advantage given to newer posts. PageRank, we've noted in passing, is dominated by a small number of pages. Reddit's built-in penalization of older posts avoids this consequence. This is a reflection of the differing goals of Reddit and PageRank: the former attempts to show users fresh content while the latter focuses only on quality. Reddit's methodology can reduce the impact of path-dependent behavior by preventing dominant posts from emerging unfairly, but it does so at the cost of pushing high-quality pages down as well.

4 The Looping Effect and Bias in Familiar Ranking Systems

The basic idea of all familiar ranking systems is an attempt to read objective quality—what sites, papers, or posts are genuinely worth reading—from social measures of what is read and responded to by whom. In any such system there will be a looping factor: a site, paper, or post is widely read because it is ranked highly, but it is ranked highly precisely because it is widely read. The existence of loops in ranking systems appears to be ubiquitous and inevitable: its role is demonstrable in even simple agent-based models of PageRank, HITS, and Reddit.

As noted, what we offer here is a review of work in progress. It is clear in all three models that rankings become poorer indicators of quality as users increasingly base their judgments on rank. In PageRank and Reddit we have calibrated this in terms of the decreasing quality of content seen by users. But it is also clear that all three models can withstand moderate amounts of rank-based bias, showing most significant divergence of rank from quality when rank is weighted at 80% or more. Further comparison between the three systems given the current measures is difficult, given differences in both structure and informational goals—Reddit's emphasis on temporally fresh content, for example. We can, however, advance several hypotheses regarding early results that will guide us in further work. Some of the differences between Reddit and the other algorithms may be due to its explicit emphasis on newer elements, dampening an inherent reputational looping bias toward older posts evident in the other two algorithms. Our initial results indicate important differences in the patterns of relationship between rank and quality at different settings. Here a partial explanation may lie in the fact that HITS uses two values in mutual recursion, effectively doubling the effect of reputational loops. Confirmation and expansion of initial results as well as further exploration of our suggested hypotheses remain as tasks for further work.

References

1. Gladwell, Malcolm.: The Order of Things: What College Rankings Really Tell Us. New Yorker Magazine, Feb. 14 & 21 (2011).
2. Bastedo, M., Bowman, N.: The U.S. News and World Report College Rankings: Modeling Institutional Effects on Organizational Reputation. American Journal of Education 116: 163-184 (2010).
3. Bastedo, Michael., Bowman, N.: College Rankings as an Interorganizational Dependency: Establishing the Foundation for Strategic and Institutional Accounts. Research in Higher Education 52, 3-23 (2011).
4. Glenski, M., Weninger, T.: (2016). Rating Effects on Social News Posts and Comments. ACM Transactions on Intelligent Systems and Technology. 8. 10.1145/2963104 (2016).
5. O'Neal, C: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, New York (2016).
6. Nematzadeh, A., Ciampaglia, G., Meneczer, F., Flammini, A.: How algorithmic popularity bias hinders or promotes quality. arXiv:1707.00574v2 [cs.CY] 14 Jul 2017.
7. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertaxtual Web Search Engine. Computer Networks and ISDN Systems 30 (1-7), 107-117 (1998).
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab (1999).
9. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5), 604-632 (1999).
10. Kleinberg, J.: Hubs, Authorities and Communities. ACM Computing Surveys (CSUR) 31 (4es) article 5 (1999).