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

Patrick Grim, Jared Stolove, Natalia Jenuwine, Adrian Apaza, Hannah vanWingen, Jaikishan Prasad, Paulina Knoblock,
Callum Hutchinson, Chengxi Li, Kyle Fitzpatrick, Chang Xu & Catherine Ming
University of Michigan Center for the Study of Complex Systems

In many real world examples, such as college ranking and online search, objects with very similar quality can be ranked significantly different. Although sometimes the data that a ranking system rely on can be dubious, even the best of conditions and data input, we argue, the very structure of some familiar ranking systems can result in distorted informational output. This research project uses agent-based techniques to analyze inherent structural bias in abstract models of common ranking systems such as PageRank, HITS, and Reddit. Using the simpler example of University rankings such as U.S. News and World Report as an example, the project outlines the existence of reputational loops as a core source of ranking distortion across a variety of ranking systems. 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, HIT, and Reddit are constructed. These models allow comparisons of bias dynamics and effects across a number of parameters.

PageRank

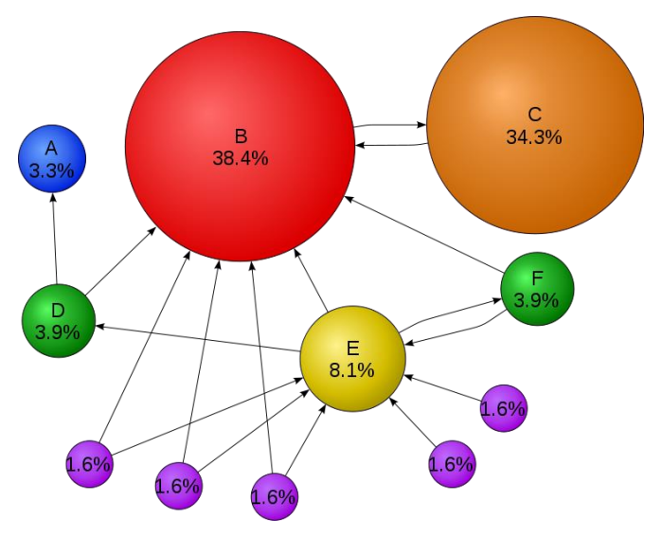


Figure 1. A simple PageRank network

PageRank analyzes the traffic between sites in order to determine the ranking of the websites. 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 node value of 1 divided by the total number of nodes. At each time interval, PageRank divides the node 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 out 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 node. PageRank then ranks the sites in order of greatest to least node value (Figure 1).

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.

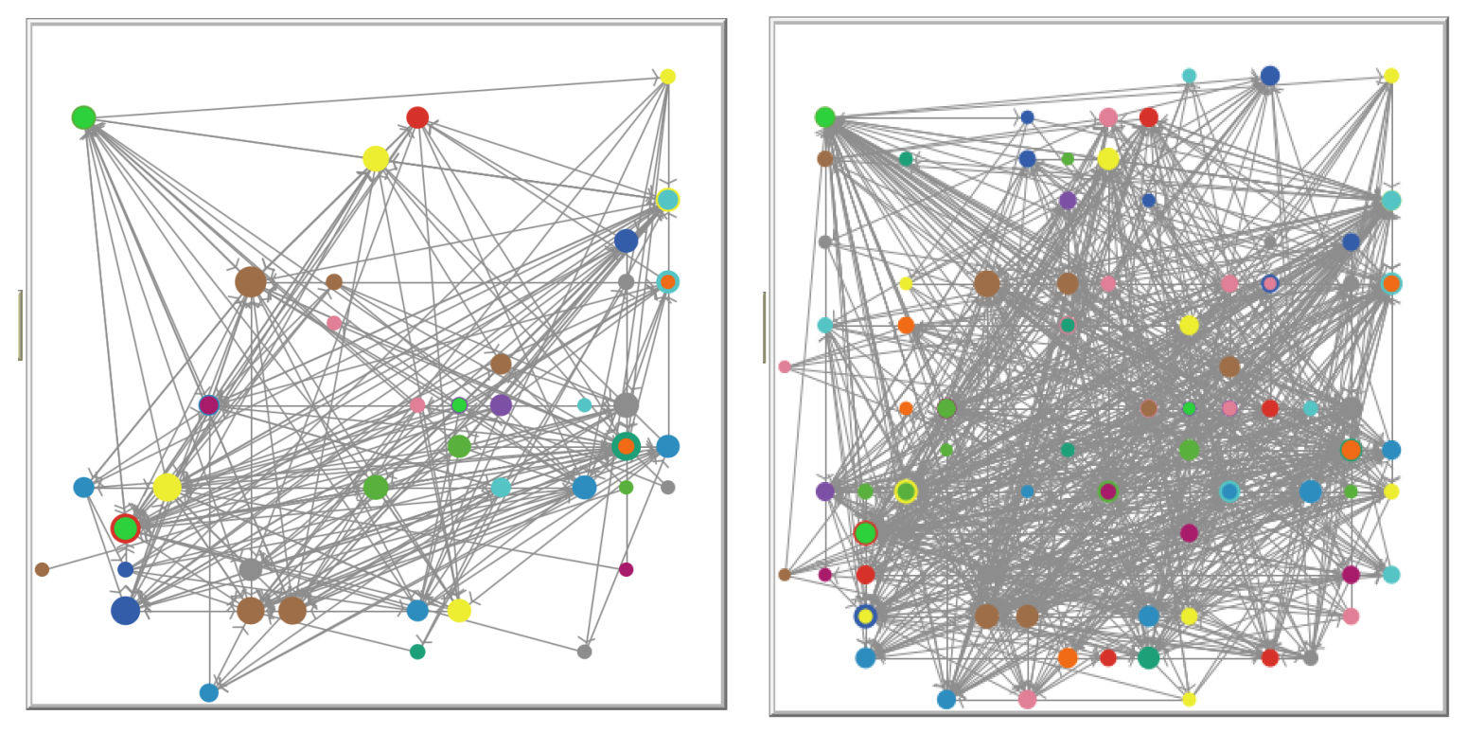


Figure 2. An example of the general progression over time of a model in which links from a node x to y are established probabilistically in terms of the quality of y alone. Stages 50 and 100 in a typical model evolution shown.

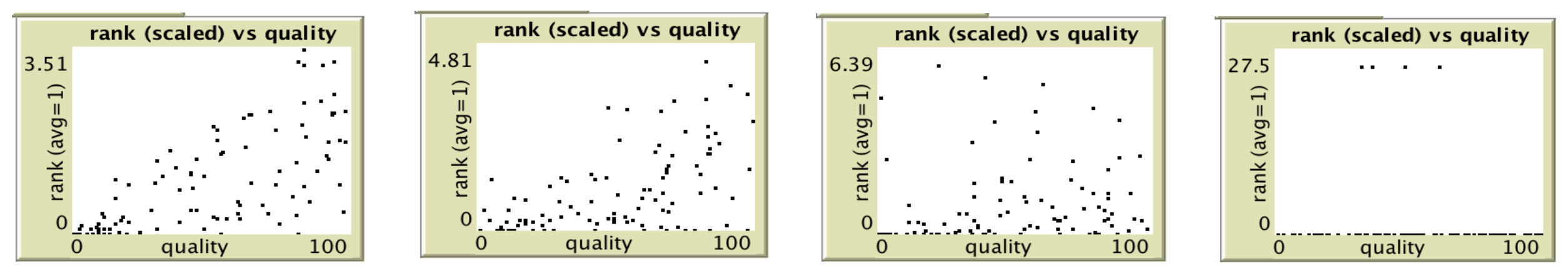
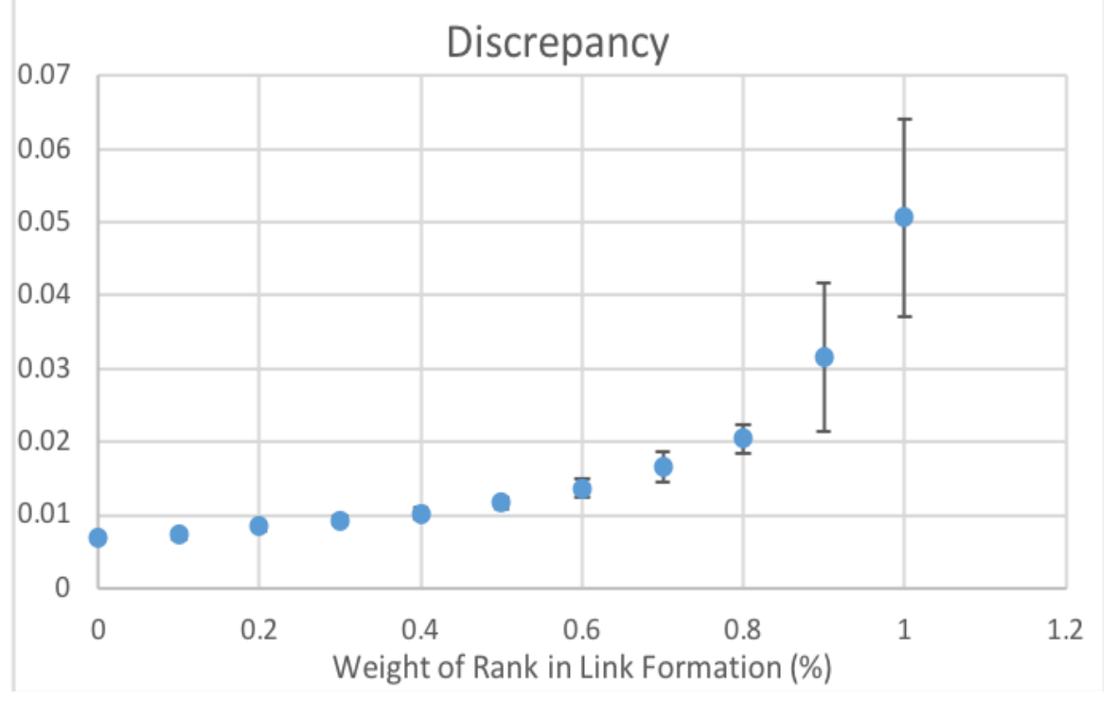


Figure 3. Rank vs. quality as the probability of link formation is calculated in terms of quality alone (left), rank .3, .7 and entirely in terms of rank (right). (num-to-add = 1, iteration 100).

The relationship between rank and quality changes significantly as the proportion of link formation determined by rank increases. When links are determined by quality alone, 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. Lower quality pages are able to obtain a higher rank; and the difference between the highest and lowest ranked pages becomes much larger.



We introduce 'discrepancy' as a measure for divergence of rank from quality. Over a wider range of cases, increasing discrepancy with increasing weight given to rank are shown in Figure 4.

Figure 4 - rank-quality vs discrepancy. Measurement taken at step 100, averaged over 20 runs.

As agents respond more to rank—what we have targeted as the more realistic case—the discrepancy between rank and quality is higher, the average quality of a page linked to is lower, and fewer pages can dominate. If links are formed entirely in terms of rank, not surprisingly, the quality of pages linked to is essentially random. As rank is considered more, pages which are established early have a decisive advantage.

HITS

The HITS Algorithm is used by Academia.edu, a social networking website that shares papers and monitors their impact. . In general, for each web page, the scheme assigns two values recursively. One is Hub Score, the sum of authority scores of all the nodes that the site points to. The other is Authority Score, the sum of the hub scores of all nodes pointing to this specific web page. A page that suggests more widely recognized authorities would obtain a higher Hub Score, whereas a page that is recommended by more quality hubs would achieve higher Authority Score. Hence, the results produced by HITS Algorithm has an interaction between these two parameters.

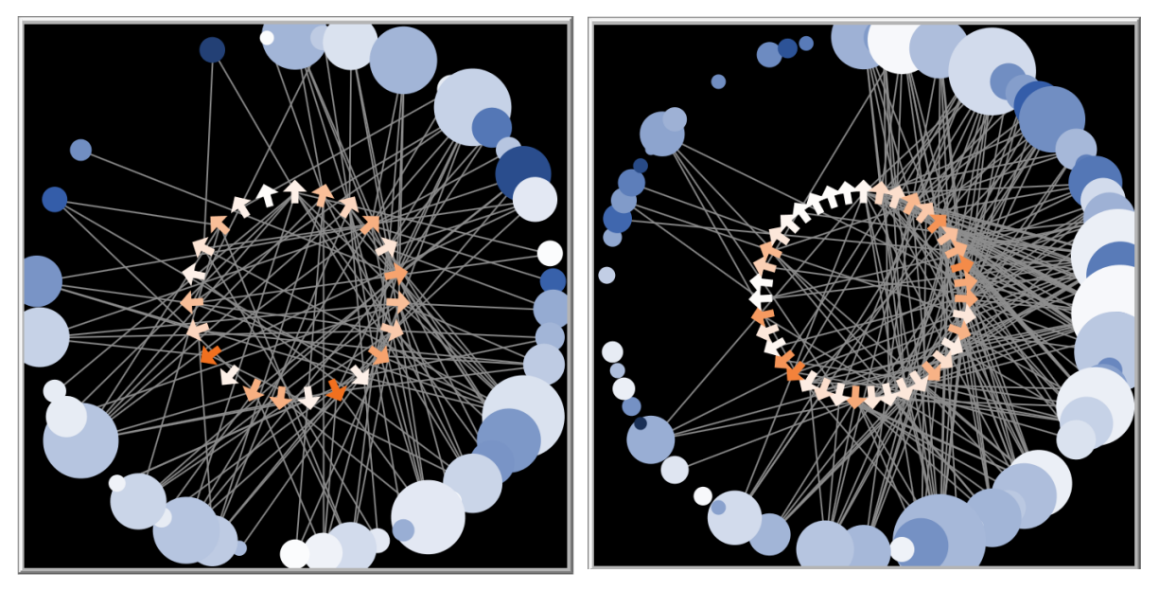
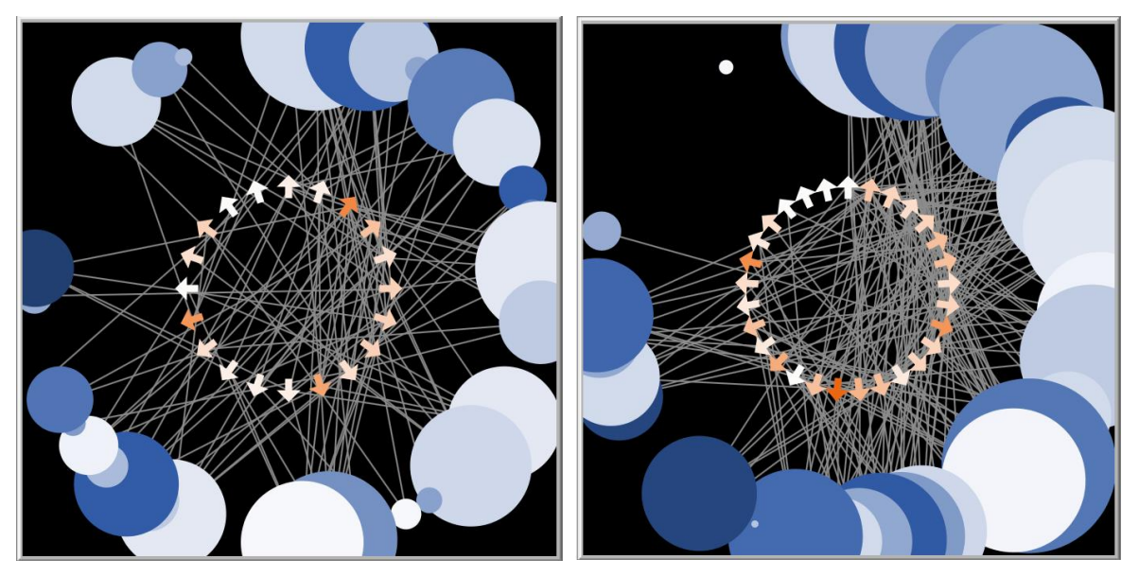



Figure 5. (left) A typical run of the model based solely on quality at 12 and 24 ticks.
(right) A typical run of the model with links made independent of quality at 12 and 24 ticks.

Orange arrows represent authors and the saturation of color represents the degree of difference between their hub score and average paper quality.

Blue circles represent papers with saturation representing the degree of difference between the authority score and absolute quality and their size representing authority relative to the other papers.

The model we built simulates an implementation of the HITS ranking algorithm used to rank academic papers. In this application, paper and author objects represent authorities and hubs respectively, each with a constant built-in quality value. During the setup, authors decide which papers to recommend solely based on the paper quality. Subsequently, papers are given authority scores based on the authoritativeness and number of authors that have recommended them, while authors are given hub scores calculated by the popularity and subjective quality of their papers. A potential problem arises in that the recommendation of an author, whose papers are popular in the academic community but of average quality, may carry more weight than that of another author with less popular but higher quality articles.

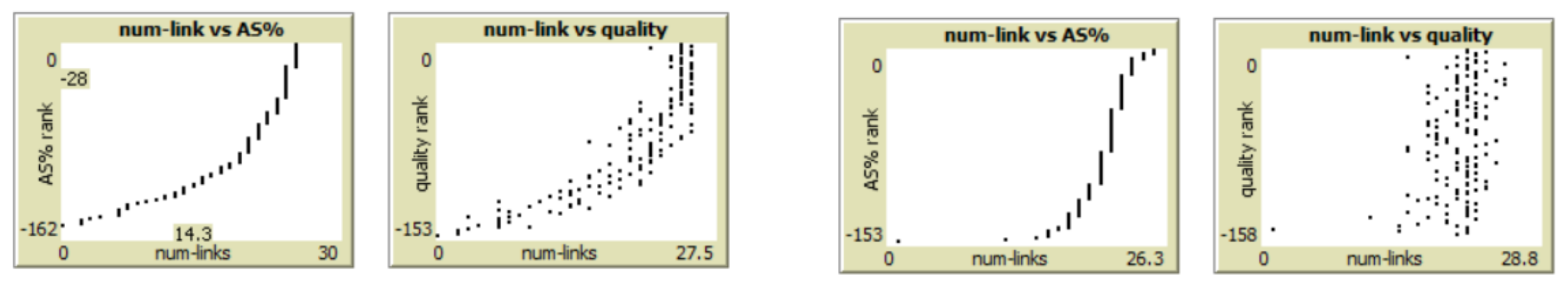


Figure 6. The relationship between quality and popularity in terms of the number of recommendations.

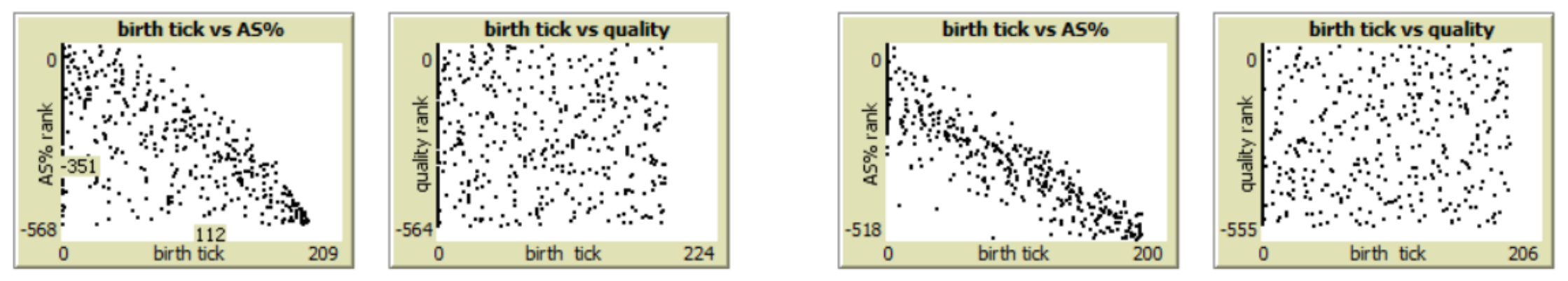
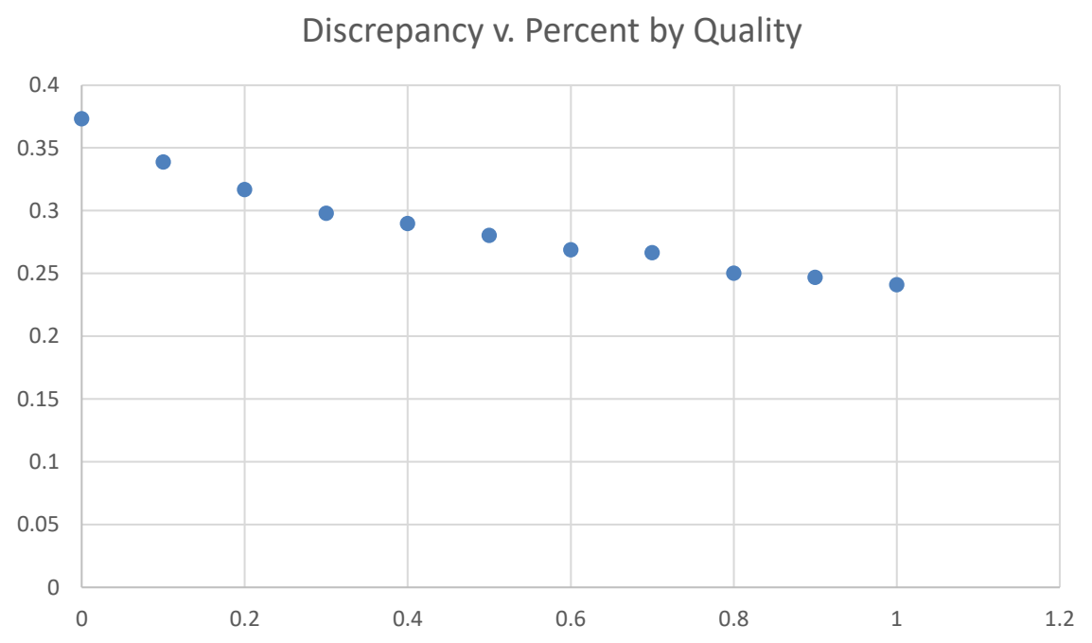


Figure 7. the influence of age on rankings produced by the HITS ranking.

*Based on intrinsic quality (25 authors, 150 papers) *Based on subjective quality (25 authors, 150 papers)



The HITS algorithm has a discrepancy that is an order of magnitude larger than that of PageRank; however, this is not unexpected. As a recursive function with two reinforcing variables a higher degree of error propagation is anticipated.

Figure 8 – Discrepancy versus Percent-by-Quality.

Since quality should be independent of age, graphs with random correlations as illustrated on the right are expected. However, the ranking based on Authority Score calculated in HITS Algorithm clearly favors papers that has been around for a longer period of time. It is difficult for high quality papers to achieve an appropriate ranking if they join the network late, indicating that there exists a bias towards older papers. We have also found that there is a bias towards older authors but that it may be justified in similar academic applications.

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 creates 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. On every post are comments, which 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.

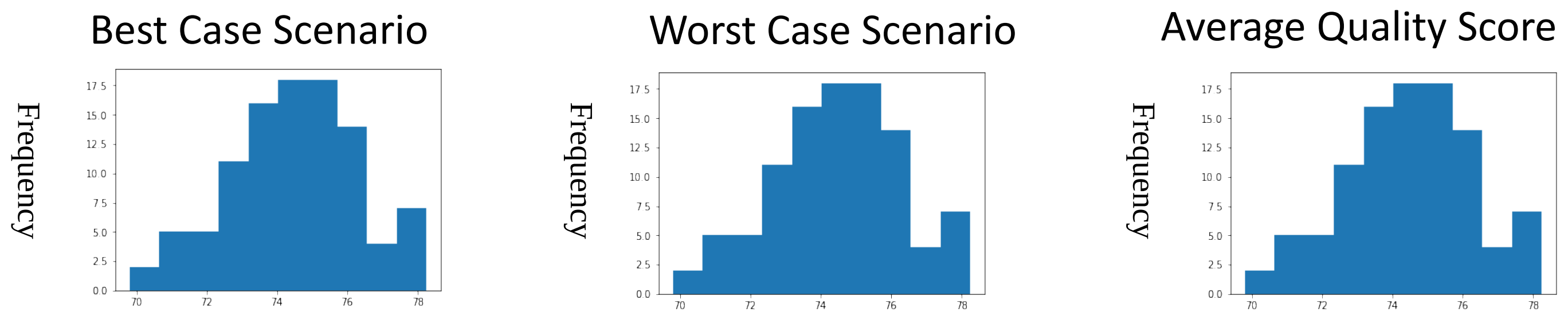


Figure 9. gives the distribution of average quality of a read post, or "average quality score", in both the best and worst case scenarios over 100 runs of the model.

Our model of Reddit simulates the website's users reading posts and voting on them. For users, the decision to read a post and to upvote or downvote a post is based on both the post's objective quality score (a number between 0 and 100) and its rank, as determined by Reddit's algorithm. Each user has personal thresholds that determine 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. Further, new posts are periodically created, which are assigned objective quality scores and then ranked according to Reddit's algorithm.

		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 10. describes 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.

		Voting Bias					
		0	0.2	0.4	0.6	0.8	1
Reading Bias	0	74.11	74.312	75.16	74.11	73.399	74.11
	0.2	72.608	70.25	71.58	74.08	71.076	73.74
	0.4	69.7	71.57	68.1166	72.358	69.18	69.757
	0.6	67.734	67.8977	66.8	66.756	67.494	64.43
	0.8	63.863	65.07	67.208	65.68	65.905	60.13
	1	63.398	66.33	62.195	63.166	61.771	48.93

		Voting Bias					
		0	0.2	0.4	0.6	0.8	1
Reading Bias	0	-0.79	-0.087	0.386	-0.15	0.555	1.78
	0.2	-0.902	-2.59	-1.897	-0.54	-2.864	1.96
	0.4	-4.03	-0.64	-7.0434	1.138	-2.99	-1.063
	0.6	-3.236	-3.1593	-5.305	-4.277	-3.972	-4.358
	0.8	-5.698	-3.27	-1.657	-1.388	-4.139	-0.65
	1	-4.9847	-0.5	-3.575	-2.814	-1.239	-0.095

Figure 11. Average Quality Score of the No-Time-Damp Model (above) and Difference between the No-Time-Damp Model and the Original Model (below)

		Voting Bias					
		0	0.2	0.4	0.6	0.8	1
Reading Bias	0	75.066	75.454	74.024	73.67	74.04	75.32
	0.2	74.16	73.83	73.383	72.78	73.19	71.813
	0.4	73.58	72.252	73.081	72.83	72.16	68.79
	0.6	72.64	73.746	71.664	72.03	69.88	65.27
	0.8	71.58	70.215	71.99	69.26	69.16	58.14
	1	67.14	68.992	69.11	67.95	64.55	51.18

		Voting Bias					
		0	0.2	0.4	0.6	0.8	1
Reading Bias	0	0.166	1.055	-0.75	-0.59	1.196	2.99
	0.2	0.65	0.99	-0.094	-1.84	-0.75	0.033
	0.4	-0.15	0.042	-2.079	1.61	-0.01	-2.03
	0.6	1.67	2.689	-0.441	0.997	-1.586	-3.518
	0.8	2.019	1.875	3.125	2.192	-0.884	-2.64
	1	-1.2427	2.162	3.34	1.97	1.54	2.155

Figure 12. Average Quality Score of the No-Time-Damp Model (above) and Difference between the No-Time-Damp Model and the Original Model (below)

Although we are still working on this section, we have noticed that Reddit's algorithm has robustness to rank-related bias, and this is because of 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, this is called the "time damp." Second, additional upvotes only impact a posts' score logarithmically, so highly ranked posts must acquire exponentially more votes to counteract the advantage given to newer posts. Also unlike PageRank, where dominate pages exist, in Reddit, the distribution of votes and quality is actually linear, so there are no dominant pages