A Predictive Perspective on Measures of Influence in Networks

Prem Melville, Karthik Subbian, Richard Lawrence, Estepen Meliksetian,
IBM TJ Watson Research Center,
Yorktown Heights, NY, 10598.
{pmelvile,ksubbian,rlawrence,emelikse}@us.ibm.com

Claudia Perlich
Media 6 Degrees,
New York, NY, 10003.
claudia.perlich@gmail.com

1. INTRODUCTION

Identifying the most *important* or *prominent* actors in a network has been an area of much interest in Social Network Analysis dating back to Moreno's work in the 1930's [1]. This interest has spurred the formulation of many graph-based socio-metrics for ranking actors in complex physical, biological and social networks. These sociometrics are usually based on intuitive notions such as access and control over resources, or brokerage of information [2]; and has yielded measures such as Degree Centrality, Closeness Centrality and Betweeness Centrality [3].

In the exploratory analysis of networks, the question of whether these measures of centrality really capture what we mean by "importance" is often not directly addressed. However, when such sociometrics start being used to drive decisions in more quantitative fields, there emerges a need to empirically answer this question. Probably the most popular of these measures in the Computer Science community is PageRank, which is a variant of Eigenvector Centrality [4]. Once its use in Information Retrieval (IR) and Web search in particular became popular, it led to more rigorous evaluation of PageRank and variants on measurable IR tasks [5].

With the rise of Web 2.0, with its focus on user-generated content and social networks, various socio-metrics are being increasingly used to produce ranked lists of "top" bloggers, twitterers, etc. For example, Twitterholic.com and WeFollow.com use degree centrality (number of *followers*), while TunkRank.com uses a variant of PageRank to rank users of the micro-blogging service Twitter. In the domain of blogs, Technorati assigns an authority score to a blogger based on the number of blogs linking to her website in the last six months. Similarly, Blogpulse ranks blogs based on the number of times it's cited by other bloggers over the last 30 days. Do these rankings really identify "influential" authors, and if so, which ranking is better? With the increased demand for Social Media Analytics, with its focus on deriving marketing insight from the analysis of blogs and other social media, there's a growing need to address this question. This paper is a step in that direction.

It is our position, that the question of whether a particular influence measure is *good* is ill-posed, unless it is put in the context of a measurable task or desired outcome. Constructing such predictive tasks of interest, not only guides the choice of relationships we build a network on, but also allows for the quantitative comparison of different socio-metrics. In this paper, we present a case study on data collected for 40 million Twitter accounts. We look at marketing-driven tasks, such as detecting the potential for viral outbreaks of messages (tweets). We build three different graphs based on the network of followers, rebroadcast (retweet) networks, and the network of replies and mentions. We conduct a similar study on detecting the influence of publications, through the analysis of citation networks. Extensive empirical results demonstrate that different measures provide the best ranking for these tasks – underscoring the importance of addressing the question of influence based on a desired objective. Taking a predictive perspective of measures of influence can also suggest alternative socio-metrics, and we show that combining aspects of different measures produces a composite ranking mechanism that is most beneficial for each desired predictive task. We compare several approaches to combining influence measures through rank aggregation methods, such as approximations of Kemeny optimal aggregation [6]. In addition, we introduce novel supervised rank aggregation techniques that leverage the ground truth on a subset of users to further improve ranking. We demonstrate the efficacy of these methods compared to several baseline approaches.

2. TWITTER CASE STUDY DETAILS

Our case study was based on the Twitter discussion around Pepsi. What piqued our interest in Twitter and the role of influencers was the infamous iPhone app called "AMP UP B4 U SCORE". An avalanche of Twitter users slammed the app ultimately leading to an apology from Pepsi.

When constructing a graph of Twitter users an obvious way to define connections is through the *follower* relationship. However, many users have 100K or more *friends* and therefore *following* may not be sufficient indication of influence. For this reason, we consider two alternative, implicitly embedded graphs that reflect the user's current behavior: the Retweet Graph and the Mention Graph. This information is not directly available, but can be extracted from the tweets following the convention of inserting the mentioned or retweeted user preceded by a "@".

Base User List: Over a period of a month during 11/2009 we used keyword searches on "Pepsi" to generate a list of tweets by a total of 9625 users. This was our base set of users for the balance of the study.

Follower Graph: Starting with the 9625 base users, we identified the 2.5 million unique users following them. Repeating this snowball sampling process we get a second iteration of 35 million users.

Retweet and Mention Graph: Starting from our 10K user base, we search for all tweets that contained any of these usernames preceded by "@". We obtained a unique list of users that made these tweets. Using this set of users we pulled a second iteration of tweets. We repeated this one more time to generate a third iteration. From these three iterations, we extracted links between users originating from retweets and links between users that reflect mentions or replies. The syntactic difference between a mentioning instance and a retweet is the leading "RT @k". A link from user *i* to user *k* in this graph means that *i* is retweeting *k*. We generate two versions of both the mention and the retweet graph, one collapsing all repeat connections from the same user *i* to the user *k* into just one edge. The second version uses the number of retweets/mentions as edge weights. Statistics on these graphs are presented in Table 1.

Statistic	Follower Graph	Retweet Graph	Mention Graph
Number of Users	39,855,505	412,744	637,509
Number of Edges	1,098,443,217	975,326	1,473,361

Table 1 - GRAPH STATISTICS

For each of the three graphs and the weighted variants we compute in-degree, out-degree, and PageRank. Table 3 lists the graph measures, along with more intuitive names.

Test Data: Ultimately we want to predict outbursts of retweets. So once these three graphs were generated, we continued obtaining additional data over the following week in the beginning of 12/2009. We gathered the first iteration of the retweet graph to keep track of how often the users in the base list were retweeted.

Measure	Definition
Followers	Follower Graph Indegree
Friends	Follower Graph Outdegree
Follower Pagerank	Follower Graph Pagerank
Distinct Past Retweets	Retweet Graph Indegree
People Retweeted	Retweet Graph Outdegree
Retweet Pagerank	Retweet Graph Pagerank
Past Retweets	Weighted Retweet Graph Indegree
Retweets Made	Weighted Retweet Graph Outdegree
Distinct Mentions Received	Mention Graph Indegree
People Mentioned	Mention Graph Outdegree
Mention Pagerank	Mention Graph Pagerank
Mentions Received	Weighted Mention Graph Indegree
Mentions Made	Weighted Mention Graph Outdegree

Table 2 - MEASURE DEFINITIONS

3. IDENTIFYING VIRAL POTENTIAL

One of the biggest opportunities and threats presented by social media is the viral outbreak of messages, videos, tweets, etc. For marketing and PR organizations this can be a boon or curse based on the sentiment expressed in these messages towards specific brands, products or entities. As such, marketers are constantly looking for ways to influence positive outbreaks or thwart negative ones. Either way, they often base their actions on the perceived importance of authors in the social media space.

In the micro-blogging universe of Twitter, this suggests that a useful task would be to predict which twitterers will be significantly rebroadcasted via retweets. We construct such a task from our data by dividing users in our test phase into two classes – people who have been retweeted 100 or more times within a week, and those who have not. Roughly 1.6% of our population (151 people) fall in the first target class. We treat this as a binary classification problem, where the ranking produced by each measure in Table 2 is used to predict the potential for viral retweeting in the test time period. Since we are primarily concerned with how well these measures perform at ranking users, we compare the area under the ROC curve (AUC) based on using each measure by itself. This gives us a mechanism to objectively compare the different measures. In addition, we devise methods to combine these individual measures, which we describe in the following section.

4. RANK AGGREGATION

One straightforward approach to combining individual measures is to use them as inputs to a classifier, such as logistic regression, which can be trained to predict the target variable on historical or held-out data. However, given that the individual influence measures produce an ordering of elements and not just a point-wise score, we can, instead leverage approaches of aggregating rankings. Methods for rank aggregation or preference aggregation have been used extensively in Social Choice Theory, where there is no ground truth ranking, and as such are unsupervised. Here, we introduce several supervised approaches to rank aggregation that can be trained based on the ground-truth ordering of a subset of elements.

Rank Aggregation Task: We begin by formally defining the general task of rank aggregation. Given a set of entities S, let V be a subset of S and assume that there is a total ordering among entities in V. We are given r individual rankers $\tau_1, \ldots, \tau_r$ who specify their order preferences of the m candidates, where m is size of V, i.e., $\tau_i = [d_1, \ldots, d_m]$, $i = 1, \ldots, r, if\ d_1 > \ldots > d_m$, $d_j \in V$, $\forall j = 1, \ldots, m$. Rank aggregation function ψ takes input orderings from r rankers and gives τ, which is an aggregated ranking order. If V equals S, then τ is called a full list (total ordering), otherwise it is called a partial list (partial ordering). All commonly-used rank aggregation methods satisfy one or more of several desirable *goodness* properties [10], such as *Unanimity*, *Neutrality*, *Consistency* and the *Condorcet Criteria* [7]. Of particular value is the *Extended Condorcet Criterion* (ECC), defined as follows. Let us split the entities into two partitions, Q and R. If for all $d_i \in$ Q and $d_j \in$ R, a majority of rankers prefer d_i to d_j, then the aggregate should prefer d_i to d_j. In this section, we describe rank aggregation methods that satisfy some of these goodness properties, as well as supervised versions of these methods.

Borda Aggregation: In Borda rank aggregation [8] each candidate is assigned a score by each ranker; where the score for a candidate is the number of candidates below him in each ranker's preferences. The Borda aggregation is the descending order arrangement of the average Borda score for each candidate averaged across all ranker preferences. Borda satisfies all goodness characteristics except Condorcet and Extended Condorcet Criteria [6]. In fact, it has been shown that no method that assigns weights to each position and then sorts the results by applying a function to the weights associated with each candidate satisfies the Condorcet criterion [6]. This includes approaches such as logistic regression. This motivates us to consider order-based methods for rank aggregation that satisfies both Condorcet criteria.

Kemeny Aggregation: Kemeny is an order-based aggregation method [6], in which the final rank aggregation reduces the number of pairwise disagreements between all the rankers, i.e., the average Kendall-Tau distance between τ and τ_i is minimum. Kemeny Optimal Aggregation is the only function that is neutral, consistent and satisfies the Condorcet criteria. However, it has been shown that computing Kemeny aggregation for r ≥ 4, is NP-Hard [6]. So, instead we use Local Kemenization [6], which is a relaxation of Kemeny Optimal aggregation that still satisfies the Extended Condorcet Criterion.

Local Kemenization is computationally tractable in practice, as opposed to Kemeny Optimal Aggregation. A full list τ is locally Kemeny optimal, if there is no full list τ^+ that can be obtained by single transposition of adjacent pair of elements, such that,

$$K(\tau^+,\tau_1,...,\tau_r) < K(\tau,\tau_1,...,\tau_r);$$

where, $K(\tau,\tau_1,...,\tau_r) = \frac{1}{r}\sum_{i=1}^{r} k(\tau,\tau_i)$

The function $k(\sigma,\tau)$ is Kendall tau distance which is the number of pairwise disagreements between two lists σ and τ. Every Kemeny optimal aggregation is also locally Kemeny optimal [6], whereas the converse is false. Dwork et al. [6] show that Local Kemenization satisfies the Extended Condorcet Criterion and can be computed in O(rm *log*(m)), where m is the size of V . The local Kemeny procedure can be viewed as a stable sorting algorithm, where given an initial ordering, elements d_i and d_j are only swapped if d_i is preferred to d_j by the majority of rankers (τ_i's). It is important to note that the initial aggregation passed to Local Kemenization may not necessarily satisfy Condorcet criteria. However, the process of Local Kemenization produces a final ranking that is maximally consistent with the initial aggregation, and in which Condorcet winners are at the top of the list [6].

Supervised Rank Aggregation: Borda and Kemeny aggregations, being motivated from social choice theory, strive for fairness and hence treat all rankers as equally important. However, fairness is not a desirable property in our setting, since we know that some individual rankers (measures) are likely to perform better than others in our target tasks. If we knew a priori which rankers are better, we could leverage this information to produce a better aggregate ranking. In fact, given the ordering of a (small) set of candidates, we can estimate the performance of individual rankers and use this to produce a better ranking on a new set of candidates. We use such an approach to produce different supervised rank aggregation methods, which we describe in more detail below.

In order to accommodate supervision, we extend Borda and local Kemeny aggregation to incorporate weights associated with each input ranking. The weights correspond to the relative utility of each ranker, which may depend on the task at hand. In this section, we focus on the task of viral prediction as described in Sec. 3. As such, we weight each ranker based on its (normalized) AUC computed on a validation (training) set of candidates, for which we know the true retweet rates. Incorporating weights in Borda aggregation is relatively straightforward, where instead of simple averages, we take weighted averages of Borda scores. When called with uniform weights we will refer to the algorithm simply as Borda. When used with weights based on training set performance, we will refer to it as Supervised Borda. A similar approach to supervised Borda was used in [9], where weights were based on average precision of each ranker for a meta-search task.

Algorithm 1: Weighted Local Kemeny (LK)

Input: $\tau_i = [\tau_{i1},...,\tau_{im}], \forall i = 1,...,r$, ordered arrangement of m candidates for r rankers.

$\omega = [\omega_1,...,\omega_r]$ – where ω_i is the weight of ranker *i*

$\mu = [\mu_1,...,\mu_r]$ – initial ordered arrangement of m candidates

k – the number of candidates to consider in each rankers preference list (k ≤ m)

Output: τ – rank aggregated arrangement of candidates in decreasing order of importance

1) Initialize $M_{i,j} \Leftarrow 0, \forall i, j = 1,...,m$
2) For each ranker $p = 1,...,r$
3) For each candidate $i = 1,...,k-1$
4) For each candidate $j = i+1,...,k$
5) $M_{\tau_{pi},\tau_{pj}} \Leftarrow M_{\tau_{pi},\tau_{pj}} + \omega_p$
6) Stable sort μ, using M_{μ_i,μ_j}. If $M_{\mu_i,\mu_j} > 0$ then $\mu_i > \mu_j$. If $M_{\mu_i,\mu_j} = 0$ then $\mu_i = \mu_j$
7) Return τ

For supervised Local Kemenization we incorporate weights directly in sorting the initial ordering. So, instead of comparing candidates based on the preference of the simple majority of individual rankers, we use a weighted majority. This can be achieved simply by using weighted votes during the creation of the majority table M which represents the sum of weights of the rankers who prefer the row candidate to the column candidate for each pairwise comparison. This weighted Kemeny procedure is presented in Algo. 1. As with Supervised Borda, we select weights based on each ranker's AUC computed on a training set. When Algo. 1 is invoked with uniform weights we will refer to it as Local Kemenization or LK. When used with the performance-based weights we will refer to it as Supervised LK. Instead of using total orderings provided by each ranker, we can also use partial orderings (for a subset of candidates). Of particular interest, is using the partial ordering only on the top k candidates for each ranker. We refer to this variant as Local Kemeny TopK (LK TopK). In Algo. 1, setting k to |S| (the default) corresponds to Local Kemenization on total orderings. Local Kemenization is sensitive to the initial input ordering μ provided to the algorithm. We experimented with initializing with Borda and Supervised Borda.

To summarize, our weighted Local Kemenization algorithm can run with varying three options, namely (1) with and without supervision, (2) with total orderings or partial (top K) orderings, and (3) with different initial orderings. We experimented with several combinations of these three options. By default, Local Kemenization (LK) refers to unsupervised Local Kemenization with total orderings and initializing with Borda. All other variants are listed in Table 3, where the names list the departures from these defaults, and the initial ordering is mentioned in parentheses, e.g. Supervised LK TopK (Supervised Borda) corresponds to using Supervised Borda for initialization, partial orderings for top K and the supervised version of Local

Kemenization. In our experiments with partial orderings we use the top ranked 15% of candidates for each ranker.

While, supervised versions of Borda appear in prior work, to our knowledge, this is the first supervised version of locally optimal Kemeny aggregation and the top K variant of it.

5. EXPERIMENTAL RESULTS

We compared all individual and aggregate measures of influence on the task of viral prediction described in Sec. 3. Our results were averaged over 20 trials of random stratified train-test splits. In each trial, the true retweet rate of 5% of the users (481 users) was set aside for training the supervised approaches, and all influence measures were tested on the remaining 95%. The results in terms of AUC for all methods are summarized in Table 3. We begin with our observations on the individual influence measures, and then discuss our results on rank aggregation.

We find that 9 of the 13 measures by themselves are quite effective at ranking the top potentially viral twitterers with an AUC > 0.8. Not surprisingly, the total number of times that someone has been retweeted in the recent past, as well the number of distinct people who have retweeted this person, are the most predictive measures. However, just using the number of followers also produces a very good ranking. Note that the Spearman rank correlation between Distinct Past Retweets and Followers is not high (0.43), suggesting that there are multiple forces at work here.

Pageranks on the Retweet Graph and Follower Graph also perform well, but not as well the in-degree measures on the corresponding graphs. This may suggest that for people who are retweeted a lot, it is sufficient to deduce this from their immediate neighbors in the Follower Graph and Retweet Graph. It would appear that the intuition behind PageRank, that links from people with higher ranks are more important, does not apply in this case.

Next, we compared the different supervised and unsupervised rank aggregation techniques; where all 13 individual measures were used as inputs to each aggregation method. We compared 8 rank aggregation methods (see Table 3), as well as logistic regression using the score from each individual measure as a feature.

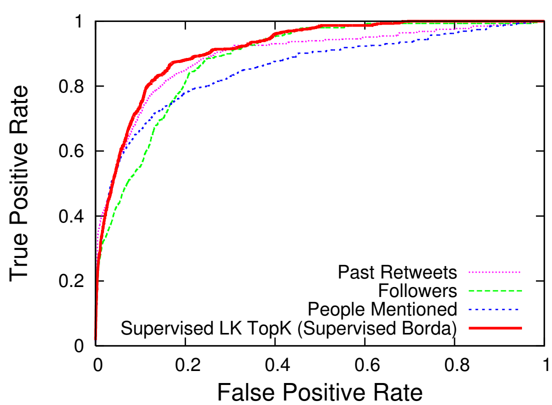

Figure 1 – COMPARING ROC CURVES FOR IDENTIFYING POTENTIAL FOR VIRAL REBROADCASTING

Measure	AUC
Past Retweets	0.8934
Distinct Past Retweets	0.8934
Followers	0.8840
Retweet Pagerank	0.8784
People Retweeted	0.8633
Follower Pagerank	0.8612
Retweets Made	0.8610
People Mentioned	0.8553
Mentions Made	0.8409
Friends	0.7629
Mention Pagerank	0.7032
Distinct Mentions Received	0.6020
Mentions Received	0.6014
Logistic Regression	0.9142
Borda	0.9084
Supervised Borda	0.9119
LK (Borda)	0.8774
Supervised LK (Borda)	0.8780
LK TopK (Borda)	0.9121
Supervised LK TopK (Borda)	0.9124
LK TopK (Supervised Borda)	0.9150
Supervised LK TopK (Supervised Borda)	**0.9156**

Table 3 - COMPARING DIFFERENT RANKING MEASURES FOR IDENTIFYING POTENTIAL FOR VIRAL REBROADCASTING

As expected, the supervised versions of each rank aggregation method performed better than the unsupervised versions. We also observe that all aggregation techniques improve over the individual rank measures. The exception here is Local Kemenization on total orderings, which can often perform worse than Borda and Past Retweets. This is counter to what one might expect from the work of Dwork et al. [6]. However, the real benefit to using Local Kemenization can be seen when it's applied only to the partial ordering of the top k candidates of each component ranking. In fact, this is implicitly the case in the results of Dwork et al. [6], where they are constrained to using partial orderings in their domain of meta-search, since the component search engines only return result on a subset of pages (which are presumably the top ranked by each). When applied to partial orderings, LK TopK performs better than Borda. These results are improved by using the supervised weighted version Supervised LK TopK; which are further improved by using Supervised Borda as the initial ranking.

The results also show that while combining individual scores through logistic regression performs well, most of our supervised rank aggregations perform better. Notably, the best performing approach is Supervised LK TopK (Supervised Borda). This confirms the advantages of supervised locally optimal order-based ranking compared to Borda and unsupervised methods. Fig. 1 demonstrates the improved performance of rank aggregation, by comparing the best supervised approach to the best individual measure from each of our three underlying graphs. These results underscore the fact that each aspect (network of followers, diffusion of past retweets, and interactions through replies and mentions) contributes to one's potential to reach a large audience. By focusing on selecting a single centrality measure to capture influence we would miss out on the opportunity to more precisely detect potentially viral tweets.

6. STUDY ON CITATION NETWORKS

In addition to Twitter data, we also performed a case study on publication citation networks. For this we used a collection of papers with their citations that was used in the KDD Cup contest held in 2003. This data[1] consists of 1,716 papers in the field of High Energy Physics Theory (*hep-th*), published on arXiv.org during a 6 month period. The data set also contains the number of times each paper was downloaded during the 60 day period after it was published on arXiv.org. This download information gives us an extrinsic proxy for the influence of a paper. As such, we define the task of predicting highly influential papers, as measured by downloads, based on the citation data of the papers. If a paper received 600 or more downloads, we consider it as a high-influence paper (77 papers); else we consider it to have little or no influence.

Measure	AUC
Pagerank	0.8109
Indegree	0.8042
Authority	0.8039
Outdegree	0.6433
Hub	0.6107
Logistic Regression	0.7602
Borda	0.7747
Supervised Borda	0.7827
LK (Borda)	0.8012
Supervised LK (Borda)	0.8013
LK TopK (Borda)	0.8154
Supervised LK TopK (Borda)	0.8163
LK TopK (Supervised Borda)	0.8142
Supervised LK TopK (Supervised Borda)	**0.8170**

Table 4 - COMPARING DIFFERENT RANKING MEASURES FOR IDENTIFYING INFLUENTIAL PAPERS

First, we constructed a citation graph based on all publications in *hep-th*, which was also provided as part of KDD Cup 2003. In this citation graph, each node represents a paper and each edge represents a citation. As of May 1, 2003, there were 29,014 papers and 342,427 citations in total in the *hep-th* data. Next, for each of the 1,716 papers with download information, we used this citation graph to compute 5 influence measures – Indegree, Outdegree, Pagerank, Hub and Authority score [10].

We ran experiments as before, using 20% of the data (343 papers) for training the supervised methods. We also set k in Algo. 1 to 1200 for the LK Top K approaches. The results in terms of AUC for each method are presented in Table 4. As expected, the number of papers citing a given paper (in-degree) is a good indicator of how often the paper will be downloaded. Furthermore, having more citations from highly-cited papers, as captured by PageRank is a better indicator of influence in this data. Note that, this was not the case in predicting viral potential in Twitter. The number of papers a paper is citing (out-degree) and Hub-score have some, though weaker, ability to predict influence. This is probably because some survey papers do become influential if they refer to many good papers in that area.

These results also substantiate the fact that supervised rank aggregation algorithms perform better than their unsupervised counterparts. Notably, Supervised LK TopK (Supervised Borda) outperforms all individual measures and rank aggregation techniques.

[1] We thank Jure Leskovec for providing the KDD '03 download estimate task data.

7. CONCLUSION

The analysis of social media is revolutionizing the way we think about actor importance and centralities in social networks. Understanding influence and authority within blog and micro-blog networks has become a crucial technical problem with a strong marketing motivation. In this paper, we have addressed this problem by casting it in the form of predictive tasks, which allows us to assess the effectiveness of different measures of influence in light of standard classification and ranking metrics. We have applied this approach to a case study involving 40 million twitter accounts, and have examined marketing-driven tasks such as determining the potential for viral out-breaks. By taking a predictive perspective of measures of influence, we demonstrated that combining aspects of different measures produces a composite ranking mechanism that is most effective for a desired task. We corroborated these results through a similar study on citation networks. We also demonstrated the merits of supervised locally-optimal order-based rank aggregation. Hopefully, this study will motivate the quantitative comparison and creation of more task-driven socio-metrics.

8. REFERENCES

[1] J.L. Moreno, Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama, Washington D. C.: Nervous and Mental Disease Publishing Co., 1934.

[2] D. Knoke and R.S. Burt, "Applied Network Analysis," Applied Network Analysis, Newbury Park, CA: Sage, 1983.

[3] S. Wasserman and K. Faust, Social Network Analysis: Methods & Applications, Cambridge, UK: Cambridge University Press, 1994.

[4] L. Katz, "A new status index derived from sociomertric analysis," Psychometrika, vol. 18, 1953, pp. 39-43.

[5] M. Richardson, A. Prakash, and E. Brill, "Beyond PageRank: machine learning for static ranking," WWW, 2006.

[6] C. Dwork, R. Kumar, R. Naor, and D. Sivakumar, "Rank Aggregation Methods for the Web," WWW, 2001.

[7] H.P. Young and A. Levenglick, "A Consistent extension of Condorcet's election principle," SIAM Journal on Applied Math, 1978, pp. 285-300.

[8] J.C. Borda, "Memoire sur les elections au scrutin," Histoire de l'Academie Royale des Sciences, 1781.

[9] J.A. Aslam and M. Montague, "Models for metasearch," SIGIR, 2001.

[10] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol. 46, 1999, pp. 604-632.