The “Top N” News Recommender: Count Distortion and Manipulation Resistance

Shankar Prawesh and Balaji Padmanabhan
Information Systems and Decision Sciences
College of Business, University of South Florida
4202 E. Fowler Avenue, Tampa, FL 33620

{shankar1,bp}@usf.edu

ABSTRACT

The broad motivation for our research is to build manipulation resistant news recommender systems. However, there can be several different algorithms that are used to generate news recommendations, and the strategies for manipulation resistance are likely specific to the algorithm (or class of algorithm) employed. In this paper we will focus on a common method used by many media sites of recommending the N most read (or popular) articles (e.g. New York Times, BBC, Wall Street Journal all prominently use this). Through simulation results we show that whereas recommendation of the N most read articles is easily susceptible to manipulation, a simple probabilistic variant is more robust to common manipulation strategies. Further, for the “N most read” recommender, probabilistic selection has other desirable properties. Specifically, the $(N+1)^{th}$ article, which may have “just” missed making the cutoff, is unduly penalized under common user models. Small differences initially are easily amplified – an observation that can be used by manipulators. Probabilistic selection, on the other hand, creates no such artificial penalty. We also use classical results from urn models to derive theoretical results for special cases.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Algorithms, Reliability

Keywords

News recommender systems, probabilistic sampling, manipulation

1. INTRODUCTION

Historically, mass media has played an important role in creating and sustaining mass opinion and behavior in society on issues ranging from policy, violence, new product adoption, family and health related issues [8, 10]. Traditionally, editorial perspectives have driven the decisions of what news to present to readers, and media editors have therefore been in positions to form and shape opinion. However, that trend is changing somewhat with technology-driven decisions that are being used instead, or in conjunction. Online news websites now use news recommendations for users that are generated automatically.

The focus of this research is to investigate the phenomena emerging through reader interaction with News Recommendation Systems (NRS hereafter) and to address the issue of *manipulation* in NRS. While there is little work that has addressed the issue of recommender manipulation for news, this topic is important because significant public opinion in society is known to be influenced by user exposure to news. For example, Phillips [9] studied the effect of publicity given to suicide stories and found that there were immediate increase in suicide cases after such news were publicized specific to time and geographic locations. To distort opinion, recommender systems are an easy target for manipulators. For example, Lerman [6] describes a Digg controversy in which a user posted an analysis proving that top 30 users of Digg were responsible for a disproportionate fraction of the front page. The allegation was that the top users conspired to promote their own articles at the expense of other articles, leading to such an increased concentration. In response, Digg modified the algorithm to devalue votes from friends. While it is not clear if the algorithmic response was appropriate – this incident highlights the context of our research agenda. Further, NRS in comparison to other recommender systems operate in a fundamentally different environment due to a constant stream of news. Such an environment places a greater need for effective recommender systems, yet suffers from potentially easier manipulation due to several factors such as the greater use of implicit feedback mechanisms where clicks are counted as votes, sparseness in various topic categories and incentive mechanisms currently in place that encourage greater clicks for higher advertising revenue.

In this paper, we examine the phenomena of count distortion created by Top-N NRS – perhaps the most widely used NRS by news websites. We also demonstrate the susceptibility of count-based NRS towards manipulation[1]. We introduce the probabilistic variant of Top-N NRS and present our findings in two ways. First using simulation, we show that the Top-N recommender is prone to artificially amplifying small differences. The $(N+1)^{th}$ article, which may have “just” missed the cutoff, is often unduly penalized in terms of readership counts in the long run. A simple probabilistic variant is shown instead to be robust. This weakness of the Top-N recommender can be exploited by manipulators who seek to gain popularity for their articles. In this context we also show that the probabilistic mechanism is again more robust. Finally, building on statistical results on classical urn models [2] we derive some theoretical insights for special cases. To our knowledge, these are all unique contributions of this paper.

1.1 Related Work

In one of the earliest research in online manipulation, Dellarocas [1] presented theoretical analysis of manipulation

RecSys’11, October 23–27, 2011, Chicago, Illinois, USA.

[1] Top-N and $count\ based$ are used interchangeably.

strategies and its impact on the firm and consumer assuming that the main source of quality information for consumers is an online product review forum. This work has established various results on effects of online forum manipulation in a simple monopoly setting. The analysis of results shows the existence of a setting where forum manipulation is equivalent to a form of quality signaling that benefits consumers. Also, if consumers expect that firms will manipulate, as the volume and quality of user-generated online content increases, then there will be a certain threshold beyond which firms will have to engage in profit-reducing online manipulation practices. The findings from closed-form solutions have been also generalized in a wide range of multi-firm settings and for a broad class of consumer utilities, firm payoff functions and signal distributions. Finally, author has proposed an idea of filtering technologies that make it costlier for firms to manipulate. We take a similar approach to study NRS through simulation and develop analytical results.

Manipulation resistant recommender systems proposed by Resnick et al. [12, 13] are also related to our work. Resnick et al. introduced the $Influence\,Limiter$ algorithm for items recommendation [12], controlling rater's influence on recommender systems through reputation acquired over time. The authors show that the optimal strategy of a rater is to induce predictions that accurately reveal the rater's information about the item. Using an information-theoretic measure the authors establish that the negative impact of any rater is bounded by a given limit. In their subsequent work Resnick et al. [13] establishes the tradeoff between resistance to manipulation by an attacker and optimal use of genuine ratings in recommender systems. A lower bound on how much information must be discarded is also provided.

Roy et al. [11] have studied linear collaborative filtering (CF) algorithms and have shown it is robust in comparison to nearest neighbor algorithms widely used in commercial systems. This analysis of linear CF algorithms shows that as a user rates an increasing number of products, the average accuracy becomes insensitive to manipulated data. The authors have established bounds on distortion as a function of percentage of manipulated data and number of products rated by a user whose future rating will be predicted. In particular for NRS, Lergillier et al. [4] have discussed a robust voting system for social news websites based on SpotRank. Considering voting as a recommendation, Lergillier et al. present a set of heuristics that demotes the effects of manipulation. SpotRank is built over $ad-hoc$ statistical filters, a collusion detection mechanism and also the reputation of users and proposed news. In their work they discuss several issues of social NRS such as the existence of cabals (collusion of large group of users that vote for each other), those who try to manipulate the system using daily mailing lists, some users posting many links to flood the system, and using several IP addresses to vote for themselves. Finally, Lerman has discussed analytical model for the news aggregation process by Digg for news recommendation and ratings [5].

2. MODEL

We present the main findings of our study using the approach of a thought experiment implemented as a simulation. This has been a powerful tool to address various issues related with social sciences and public policy [7, 14]. For instance, using a thought experiment Schelling [14], showed that a small preference for one's neighbors to be of the same color could lead to total segregation of society, and Maroulis et al. [7] studied the survival of public schools based on individual choices.

2.1 Model Description

We set up a simulation model as follows. We maintain a $comprehensive\ list\ (CL)$ of articles and their corresponding counts (or clicks). From CL, N articles are selected for display as "recommendations". Before the simulation starts articles are assigned random counts in some range (e.g. between 0 and 1000). Articles are sorted in decreasing order of their counts and the articles with high counts are selected for the $Display\ List\ (DL)$. Further, the $(N+1)^{th}$ article was deliberately assigned a count of exactly one less than the count of N^{th} article.

The selection of articles in the DL is updated at a pre-selected time step, and this selection of articles is based on two different selection mechanism namely, $count\ based$ and $probabilistic\ selection$. Count based selection is a "hard cutoff", which selects N articles for display corresponding to the highest counts. This is typically how most online news sites display the most popular or viewed articles, typically in a prominent box or sidebar. Probabilistic selection on the other hand, is a mechanism proposed here, where articles are selected probabilistically based on their counts thus far. In this mechanism, every article in CL will have some probability, based on its count, to appear in DL. Pseudo code for the implementation of these selection processes is discussed later in this section.

Two different reader models were also implemented. In both models a user is assumed to select an article either from DL with some probability p_1 or from the remaining list $RL\ (=\ CL-DL)$ with probability $1-p_1$. In the first model a reader selects an article from DL randomly. Whereas, in the second model the top-most article in the DL has the highest probability of being selected and the bottom-most has the lowest probability, with a linear decrease in the selection probability between top-most and bottom-most articles. Hence, for the second reader model the probability of a particular article with rank $i, i \in \{1,2,..N\}$ in DL being read (selected) is given by $r_i = \frac{N+1-i}{\sum_{i=1}^{N} i}$. Here, we define rank as the order in which articles are displayed in the recommended list. For ease of exposition the present model intentionally leaves out other complicated factors of news arrival and reader behavior.

2.1.1 Count Based NRS

Pseudo code for count-based selection update is presented below. The update for the second user model has same pseudo-code except the selections of articles from DL are performed based on the probability r_i.

```
For each reader
    Sort the updated count and select $N$ articles for $DL$
    If selected article is from $DL$(i.e with probability $p_1$)
        Randomly choose an article from $DL$
        Increase its count by 1
    Else
        Randomly choose an article from $RL$;$(RL = CL - DL)$.
        Increase its count by 1
End for.
```

2.1.2 Probabilistic NRS

Probabilistic selection of articles is based on probabilistic sampling without replacement for N articles. Here, probability that an article will be selected in DL is given by $prob(a) = \frac{count_a}{\sum_j count_j}$, where $count_a$ represents the count of an article 'a' at a given time step and $\sum_j count_j$ represents the total counts of articles those are not yet selected for DL. This sampling process is repeated N times to generate the N recommendations in DL.

Pseudo code for the implementation of probabilistic selection update and probabilistic selection is presented below.

```
For each reader
    Perform probabilistic selection for $N$ articles in $DL$
    If selected article is from $DL$
        Randomly choose an article from $DL$
        Increase its count by 1
    Else
        Randomly choose an article from $RL$
        Increase its count by 1
End for
```

2.1.2.1 Probabilistic selection

```
1. The count of articles are $c[1], c[2], \ldots \ldots c[n]$
2. $count[1] = 0$
3. $count[2] = c[1]$
4. for x = 3 to $n + 1$
    a. $temp \leftarrow c[x - 1]$
    b. $count[x] = count[\mathrm{x} - 1] + temp$
5. end for
6. for y = 1 to $N$
    a. generate a random integer (R) between 0 and $count[\mathrm{n} + 1]$
    b. determine the indices between which R lies, as $(i, i + 1)$
    c. select article corresponding to the count $c[i]$ for $DL$
    d. Remove $count[i + 1]$ and $i$ th article
    e. j ← $c[i]$
    f. While ($i$ is less than $n + 1$)
           $count[i + 1]$= $count[i + 1]$-$j$
    g. end while
7. end for
```

2.2 Measures

In order to compare different user models and selection mechanisms we introduce two specific measures here. Both of these measures are based on the counts of N^{th} and $(N + 1)^{th}$ articles over the complete simulation. Both N^{th} and $(N + 1)^{th}$ articles selected here are based on the initial counts of articles before the simulation starts.

Measure M1. This is defined as the logarithmic-ratio of the counts of N^{th} and $(N + 1)^{th}$ articles at each time-step as follows: $M1(i) = \ln(count_{Ni}) - \ln(count_{(N+1)i}) = \ln \frac{count_{Ni}}{count_{(N+1)i}}$ at the i^{th} iteration of the simulation. This measures the relative change in counts of N^{th} and $(N + 1)^{th}$ article. At the start of the simulation, $count(N) \sim count(N + 1)$, hence $M1(0) \sim 0$.

Measure M2. The count (hits) of the j^{th} article divided by total number of count (hits) at a given time. We denote it as M2 and at i^{th} iteration it will be $M2(i) = \frac{count_{ji}}{\sum_{p=1}^{n} count_{pi}}$. It represents the *share* of the counts for any particular article j in the NRS over iterations.

2.3 Update Rule

At each time period the model proceeds as follows. One reader arrives at each time step. Upon arrival reader selects probabilistically to read an article either from displayed list (DL) or the remaining list (RL) of articles. The probability of selection of an article either from DL or RL is controlled in the simulation. If a reader selects an article from RL then random selection of an article is performed. The count of the selected article is increased by 1.

If a reader selects an article from DL then random selection of an article is performed for Reader Model 1 and selection of an article is performed according to probability r_i for the Reader Model 2. The count of selected article is increased by 1.

For two different NRS count-based and probabilistic, the selection of N articles is made for DL, and DL is updated at each time step.

2.4 Manipulation

To study manipulation, we assume that a manipulator can create artificial clicks to raise the counts of a selected article (such as by creating fake IDs for instance). These fake counts are randomly distributed over the given interval. These fake counts are created by malicious readers who upon arrival increase the count of a particular article by 1. The particular article selected for manipulation in the present model is $(N + 1)^{th}$, mentioned earlier in the section 2.1, since this is the article that would have just missed a hard “top N” cutoff. Also, we study two types of manipulation – early and uniform – to examine what impact each might have. In “early” manipulation, the fake clicks are assumed to be distributed in some early part of the time period; in “uniform” manipulation the fake clicks are uniformly distributed over the entire time interval. We also examine the extent of manipulation (high and low, based on how many fake counts are generated) and the impact it can have.

3. SIMULATION RESULTS

The analyses of our results are based on two sections: (1) without manipulation and (2) with manipulation. The simulation results for “without manipulation”, explains the phenomenon that emerges using different NRS based on different selection mechanisms. In, particular we compare the two measures M1 and M2 for N^{th} and $(N + 1)^{th}$ articles and discuss findings based on them. Manipulation has been introduced to demonstrate the susceptibility of count-based NRS and the robustness of the proposed probabilistic NRS as an alternative. Manipulation has been introduced in two stages to study the effects of early manipulation and manipulation over large interval of time. In the first case the manipulated counts are distributed uniformly between 0 and 100 and in the second case manipulated counts are distributed uniformly between 0 and 1500. We consider different scenarios based on (a) the reader models (two), (b) the existence of manipulation (two) and (c) the selection mechanism (two – count-based and probabilistic) as described in the tree in Figure 1 of the next page. The leaves of the tree correspond to specific simulation scenarios. As Figure 1 shows, there are 12 leaves for some specific choice of global simulation parameters.

Table 1: The model parameters used in the simulation

Parameter	Value
No. of Readers	1500
No. of articles in DL	10
No. of articles in CL	200
Initial counts of articles[2]	Random Integer between 0 and1000
Manipulation Counts	10 and 50
Probability of selection of an article from DL ($P1$)	0.9, 0.5, 0.25, 0.1

[2] Except N^{th} and $(N + 1)^{th}$ articles, counts for these articles were assigned such that $count(N) - count(N + 1) = 1$. This makes it possible for N^{th} article to get selected in DL initially, whereas $(N + 1)^{th}$ just misses to get place in the DL.

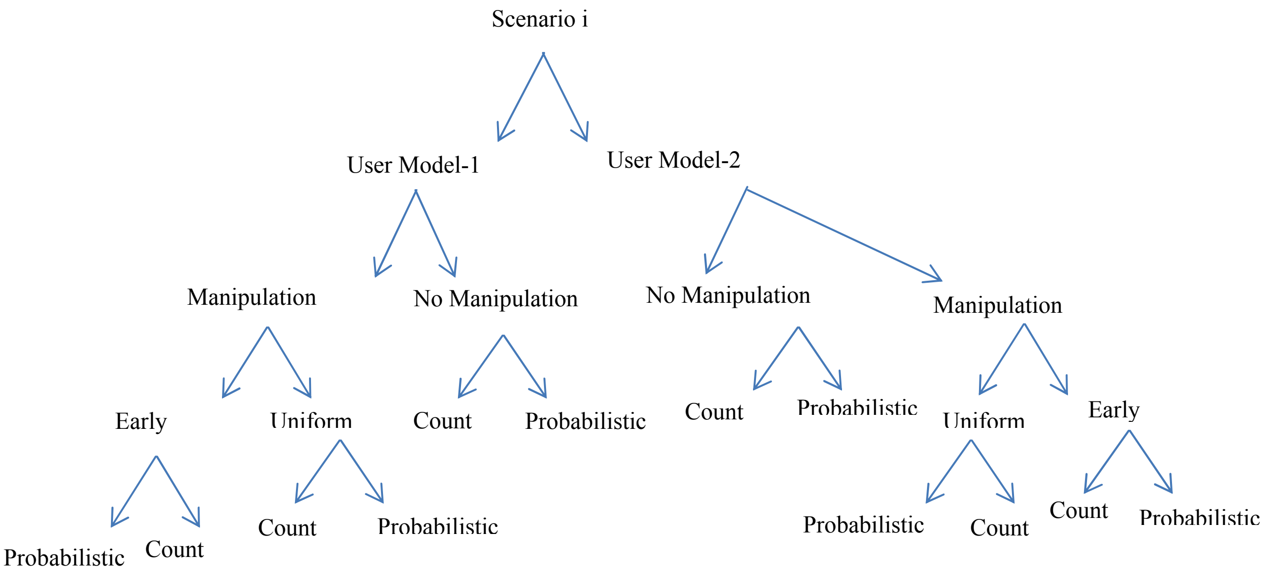

Figure 1: Graph for specific selection of global parameters

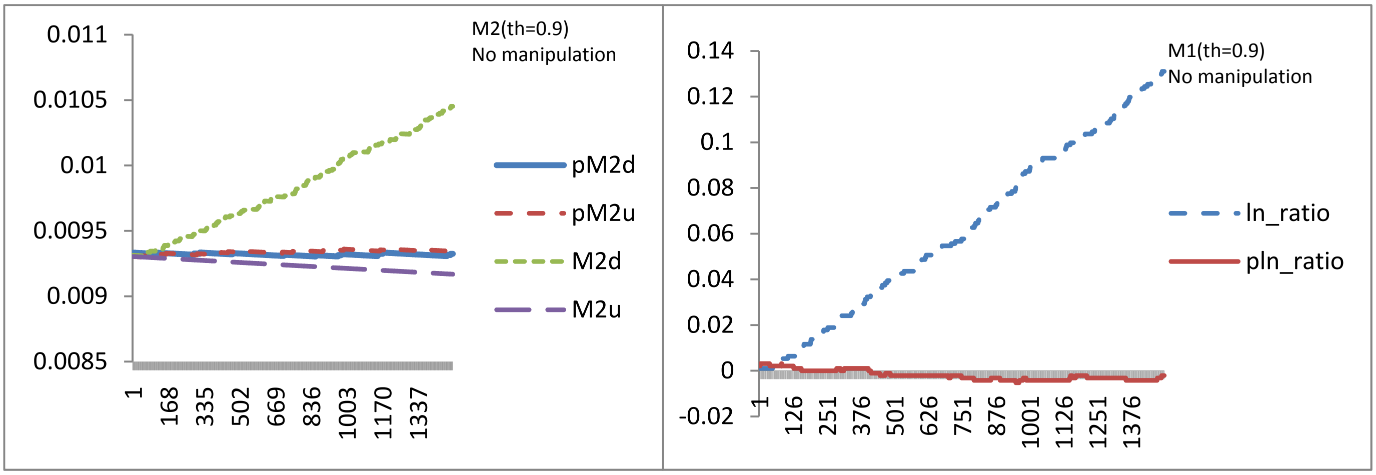

Figure 2: Simulation results for the user-model 1without manipulation (P1=0.9)

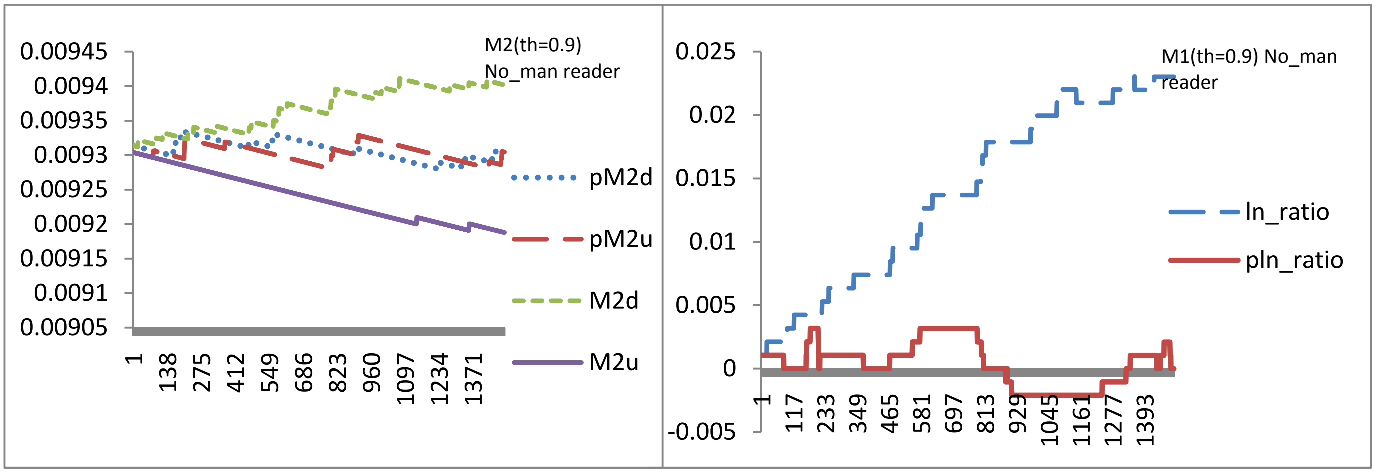

Figure 3: Simulation results for the user-model 2 without manipulation (P1=0.9)

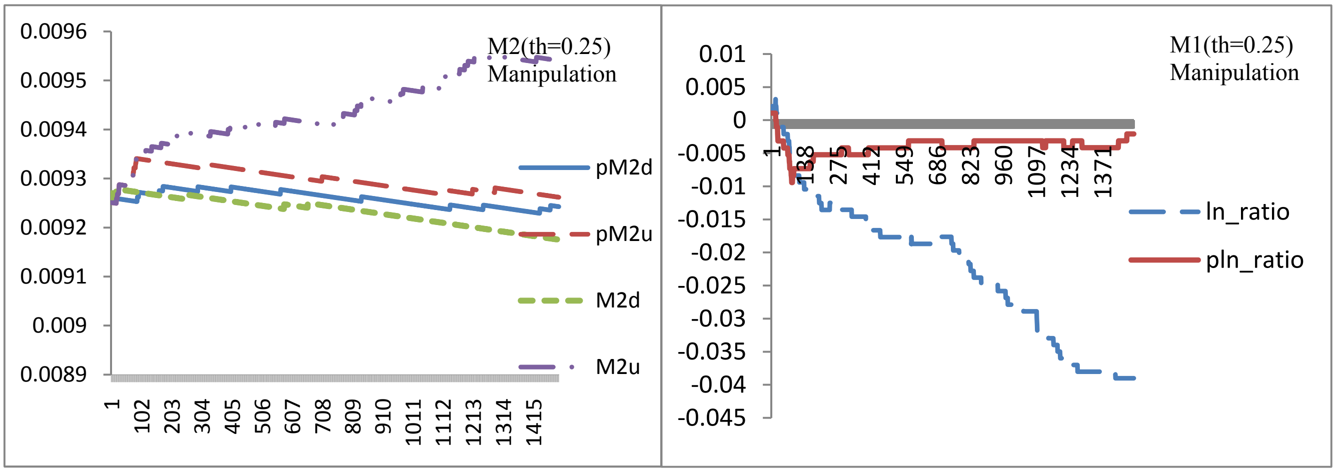

Figure 4: Simulation results for the user-model 1with little early manipulation (P1=0.25)

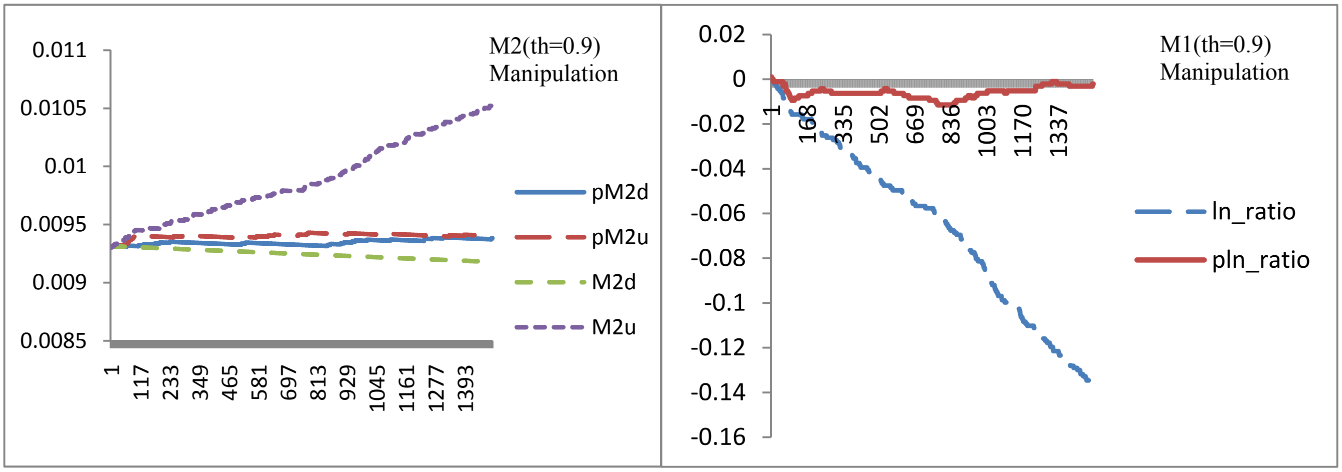

Figure 5: Simulation results for the user-model 1 with little early manipulation (P1=0.9)

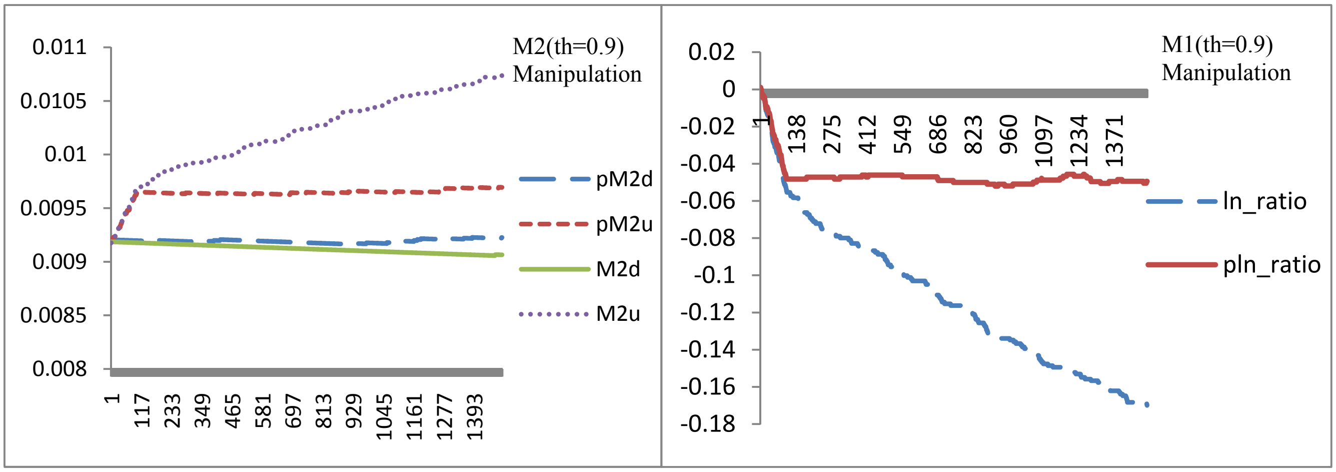

Figure 6: Simulation results for the user-model 1 with heavy early manipulation (P1=0.9)

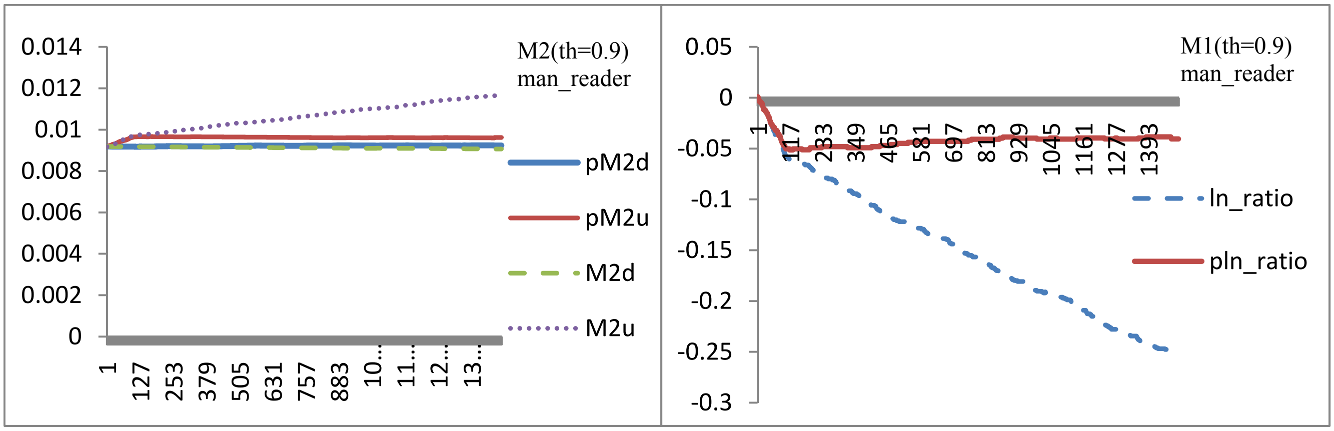

Figure 7: Simulation results for the user-model 2 with heavy early manipulation (P1=0.9)

Two of the global simulation parameters are (1) probability of a reader selecting an article from DL instead of from RL (varied as 0.9, 0.5, 0.25, 0.1), and (2) the extent of manipulation (high or low, implemented in the simulation as manipulated counts). For any specific choice of these two parameters we have 12 graphs in the results (corresponding to the 12 leaves of the tree). For lack of space we present only a few of these graphs here, but discuss our main findings based on the entire simulation results below (the different simulation paths in the graphs are better seen in color).

Table 2: Abbreviations used in the figures

M2d	M2 for the N^{th} article in count-based NRS
M2u	M2 for the $(N+1)^{th}$ article in count-based NRS
pM2d	M2 for the N^{th} article in probabilistic NRS
pM2u	M2 for the $(N+1)^{th}$ article in probabilistic NRS
ln_ratio	M1 for N^{th} and $(N+1)^{th}$ article in count-based NRS
pln_ratio	M1 for N^{th} and $(N+1)^{th}$ article in probabilistic NRS
th(or $p1$)	Represents the probability that an article will be read from the DL

3.1 Results without Manipulation

We summarize our findings based on the measures M1 and M2; through selected simulation scenarios. When there is a high probability that a reader will click on the article recommended by NRS (or DL), even negligible initial difference between the counts of N^{th} and $(N+1)^{th}$ article gets amplified heavily in the count-based NRS, as it is evident from the figures 2 and 3 for both reader models. We find that a similar pattern continues for the probabilities as low as 0.5 for both user–models. However, for the probabilistic NRS the values of M1 and M2 remains almost constant throughout the simulation.

M1 and M2 exhibit a pattern similar to random-walk for very low probability of selection of an article from DL in both count-based and probabilistic NRS. However, this is unlikely as the articles displayed in the DL typically have significantly higher chances of being read in comparison to the other articles.

After all articles have achieved their “natural counts”, in a natural system we expect that share of counts for each article will not vary much. However, for the count-based NRS the difference of M2 between the displayed and non-displayed article shows a consistent increasing pattern even though initial difference between displayed and non-displayed article was negligible. M1 also increases consistently over time for the count-based NRS.

These findings suggest that popular mechanisms using hard cutoffs may be susceptible to fundamentally creating, or amplifying, differences that may not be desirable. Probabilistic selection on the other hand is a more robust mechanism.

3.2 Results with Manipulation

In this section we will discuss the effect of different manipulations on both NRS. Manipulation counts are uniformly distributed over initial 100 (“early manipulation”) and over the entire 1500 article counts (“uniform manipulation”). Two manipulation counts considered are 10 (“low”) and 50 (“high”) when the system is slightly and heavily manipulated. In total we have four different scenarios of manipulation.

- Low fake counts uniformly distributed early.
- Low fake counts uniformly distributed over the entire process.
- High fake counts uniformly distributed early.
- High fake counts uniformly distributed over the entire process.

First we will discuss the findings of low manipulated counts. For the same un-displayed (i.e. $(N+1)^{th}$)article in RL its count was increased by 10 randomly but early in the process. However, findings in this case were reversed from the findings in non-manipulated systems. For the user-model with random selection even with probability as low as 0.25 both measures M1 and M2 (figure 4) suggest that the difference in count for the manipulated ($(N+1)^{th}$) and the non-manipulated article (N^{th}) gets amplified even if genuine readers arrive in the system. For the second user-model, in which selection of an article is based on r_i; similar phenomena is observed for reading probability up to 0.5. This suggests that once a manipulator is successful to make his article appear in the DL the implicit feedback mechanism of count-based NRS will help the manipulated article to gain more counts as more readers arrive. This is a key finding as this characteristic of count-based NRS invites manipulators to put little investment initially to increase the counts of a particular article, to make it appear in the DL, after which no further manipulation may be required.

However, for the probabilistic NRS manipulation seems to have little or no effect (figure 5) even when the selection probability from DL is almost 1. For 10 fake counts uniformly distributed over 1500 the findings are similar to case of non-manipulated count-based and probabilistic NRS. This is because the counts are sparsely distributed over the large interval. Hence, it suggests that a manipulation strategy may not be successful if the effort of a manipulator is distributed over large period of time.

We used the second manipulation strategy with 50 fake counts to compare the performance of both NRS when the system is heavily attacked by manipulators. In the first case when 50 counts are randomly distributed over first 100 counts, i.e. system is heavily manipulated in the early stage. The major benefit of probabilistic NRS appears. In all cases probabilistic NRS produced stable results in which M1 and M2 are not amplified after the manipulation, whereas the performance of count-based NRS is highly distorted for high probability of selection of articles from DL (figure 6, 7). For low selection probability from DL performance of both NRS is similar. Finally for the 50 fake counts distributed over 1500 counts no clear pattern emerges, however both NRS are similar for low selection probability from DL.

4. ANALYTICAL RESULTS

To understand how easily amplification can happen for hard cutoff NRS and robustness of probabilistic NRS toward amplification, we present insights of processes generated through both NRS in a simple setting of a two article case. The discussion that follows in Section 4.1 provides an intuitive explanation of the phenomenon for a single time step and it is just for illustrative purposes. Section 4.2 and the appendix extends this idea and presents theoretical results for any n time steps.

4.1 Illustration

Consider a special case in which both count-based (i.e. Top-N) and probabilistic NRS are implemented for two articles (article-a and article-b). A reader upon arrival reads the recommended article with probability p or reads the other with probability $1-p$. The natural counts of these articles at time $t=0$ are given by n_0 and m_0 respectively. The "natural counts" can be interpreted as the overall preferences of readers for these two articles before any recommender was put in place. Further without loss of generality we assume $n_0 > m_0$.

Let us denote the initial share of article-a and article-b by p_a and p_b respectively and it is given by $\frac{n_0}{n_0+m_0}$ and $\frac{m_0}{n_0+m_0}$.

In this simple one time period model the NRS results in amplification of the count of recommended article if at the next step due to recommendation $E(p_a) > \frac{n_0}{n_0+m_0}$.

4.1.1 Count Based NRS

In hard cutoff NRS the probability of recommended article being read is given by p. Hence, any reading probability $p > \frac{n_0}{n_0+m_0}$ will result in amplification of the counts for the recommended article. Consider a case when $n_0 \sim m_0$ $e.g.$ $n_0 = m_0 + 1$, then hard cutoff NRS will be susceptible to amplification if $p > 0.5$. Given that this is a two article case we expect p to be greater than 0.5 and hence this always creates amplification when $n_0 > m_0$.

4.1.2 Probabilistic NRS

The total probability that an article-$'a'$ will be read is given by

$$p(read) = p * p_a + (1-p) * (1-p_a)$$

So, in case of probabilistic NRS the amplification will happen for the recommended article if

$$p\left(\frac{n_0}{n_0+m_0}\right) + (1-p)\left(\frac{m_0}{n_0+m_0}\right) > \frac{n_0}{n_0+m_0}$$

$$\Leftrightarrow\ m_0 + p(n_0 - m_0) > n_0$$

$$\Leftrightarrow p(n_0 - m_0) > n_0 - m_0$$

$$\Leftrightarrow p > 1\ \nLeftarrow contradiction!$$

The above condition will never be true as $0 \le p \le 1$. It is easy to see that when the counts are similar, probabilistic NRS does not create amplification (reading probabilities will both be 0.5).

Building on this, below we present results for the more general case where we examine counts at the end of n iterations.

4.2 NRS Properties

PROPOSITION 1. *Given reading probability of recommended article is p, in Top-N NRS total expected count (denoted as $E(A_n^h)$) for the article-a with initial count n_0, such that $n_0 > m_0$ after $'n'$ iterations is given by $(n_0 + np)$.*

PROPOSITION 2. *Given reading probability of recommended article is p, in probabilistic NRS total expected count (denoted as $E(A_n^p)$) for the article-a with initial count n_0, such that $n_0 > m_0$ after $'n'$ iterations is bounded by the interval (I_1, I_2). Where $I_1 = \left(\frac{n_0+m_0-1}{n_0+m_0+n-1}\right)\left(\frac{n_0-m_0}{2}\right) + \frac{n_0+m_0+n}{2}$ and*

$I_2 = \frac{n_0}{n_0+m_0}(n_0 + m_0 + n)$.

Proof of both propositions is omitted due to lack of space (will be made available online). The result for proposition 1 is obtained through a simple binomial process. Whereas, for proposition 2 we establish bounds for the expected counts in probabilistic NRS through two different urn processes (1) Bernard Friedman's urn and (2) Pólya's urn processes [2, 3] that start with same initial condition.

Pólya proposed the urn problem to model contagion [2]. This problem is defined as an urn containing two balls of different color with initial counts say, n_0 and m_0. Each time a ball is drawn from the urn randomly and the ball is replaced in the urn with another ball of same color. Bernard Friedman's urn problem [3] is a simple variation of Pólya's urn problem in which upon random draw of a ball, the ball is replaced in the urn with another ball of different color. In both of the urn problem, we calculated expected count of the ball that has initial count of n_0 $(n_0 > m_0)$ and these values are given by I_1 and I_2 for Bernard Friedman's urn and Pólya's urn processes respectively.

4.2.1 Implications

From propositions 1 and 2 we have $E(A_n^h) = n_0 + n * p$ and $\left(\frac{n_0+m_0-1}{n_0+m_0+n-1}\right)\left(\frac{n_0-m_0}{2}\right) + \frac{n_0+m_0+n}{2} \le E(A_n^p) \le n_0 + \frac{n_0}{n_0+m_0}n$

Now consider a case where NRS has fairly strong influence on reader's reading behavior i.e. $p \sim 1$ and the difference in sufficiently large natural counts after which articles 'a' and $'b'$ make into NRS is negligible i.e. $n_0 - m_0 \sim 0$, in particular let us assume $n_0 = m_0 + 1$. So, the approximate value of expected count of article-a in hard cutoff NRS and probabilistic NRS is given by

$$E(A_n^h) = m_0 + 1 + n \qquad (1)$$

and

$$\frac{m_0}{2m_0+n} + \frac{2m_0+n+1}{2} \le E(A_n^p) \le m_0 + 1 + \left(\frac{m_0+1}{2m_0+1}\right)n \qquad (2)$$

Increase in counts of count-based selection and probabilistic selection NRS due to recommendation can be obtained through subtracting the initial count of article-a in expressions (1) and (2). So, we have

$$E(A_n^h) - (m_0 + 1) = n \qquad (3) \text{ and}$$

$$\frac{m_0}{2m_0 + n} + \frac{n - 1}{2} \le E(A_n^p) - (m_0 + 1) \le \left(\frac{m_0 + 1}{2m_0 + 1}\right) n \quad (4)$$

Using approximation $\frac{m_0+1}{2m_0+1} \sim \frac{1}{2}$ in expression (4) gives us following condition

$$\frac{m_0}{2m_0 + n} + \frac{n}{2} - \frac{1}{2} \le E(A_n^p) - (m_0 + 1) \le \frac{n}{2} \qquad (5)$$

For large n, from (3) and (5)

$$E(A_n^h) - (m_0 + 1) \to n \text{ and } E(A_n^p) - (m_0 + 1) \to \frac{n}{2}$$

So, from the findings of above expressions we conclude that for two equally good articles probabilistic NRS is not susceptible to artificial amplification in counts for the recommended article, whereas hard cutoff NRS generates processes that leads to highly amplified counts for the recommended article when the NRS is fairly influential (p is very high). This is the case since two articles with the same counts initially should increase their respective counts by ~ *n*/2 at the end of *n* iterations, which happens with the probabilistic mechanism only.

5. CONCLUSION AND FUTURE WORK

There has been growing evidence of the influence of NRS on users. A recent article in the Wall Street Journal [15] had noted that the influence of NRS is sparking a new form of "payola" as marketers try to get more votes and allow users to vote for their favorite submissions. This phenomenon has been further propelled by social networking applications such as Facebook and Twitter. As per this article, the aggregation process of news through NRS is also giving rise to an "obsessive sub-culture of a few active users who just purely for the thrill of it, are trolling the web-space for news and ideas to share with others". For example, a Reddit user known for "scoping" drove about 100,000 visitors to one amateur photographer's website [15]. There are also some marketing companies in existence who promise clients that they can get a client front-page exposure in exchange for a fee [15].

In light of all this, news recommender systems should be particularly careful to avoid common manipulative strategies. At present, articles with highest count or popularity are displayed on the front page and can be seen by millions of people. It is evident from the findings we present in this paper that the practice of using a "hard cutoff" is in particular a potentially troublesome one. In addition to perhaps unduly penalizing the marginal "next" article that missed this cutoff, this system is vulnerable to manipulation. A simple probabilistic mechanism can instead be used to present popular articles and has more desirable properties as we show.

A practical issue in implementation is that sites may be reluctant to implementing probabilistic "top" lists, since this may be likely to confuse repeat visitors. For instance, a user on a different visit may be shown a different list of recommendations, rather than the one that was previously generated. Also, in such a mechanism the term "most popular" is no longer true, since these are not the actual "top" articles. Instead, this may have to be replaced by a more appropriate term (e.g. "Popular articles") that might make it less appealing to users.

In ongoing work we are working on theoretical results for manipulation as well as additional properties and user models in the simulations. We are also examining other forms of manipulation in recommender systems and solutions.

As a last thought, we also note that a similar argument can be made not just for "news" recommendations, but for any recommender that uses a hard cutoff. However we leave the treatment of this for future work since other types of products (e.g. movies, consumer products) may have other unique characteristics or constraints.

6. REFERENCES

[1] C. Dellarocas. Strategic Manipulation of Internet Opinion Forums: Implication for Consumers and Firms. *Management Science,* volume 52(10): 1577-1593, 2006.

[2] F. Eggenberger and G. Pólya. Über die statistik verketteter vorgäge. *Zeitschriftfür Angewandte Mathematik und Mechanik*, volume 3: 279–289, 1923.

[3] D. A. Freedman. Bernard Friedman's Urn. *The Annals of Mathematical Statistics,* volume 36(3): 956-970, 1965.

[4] T. Largillier, G. Peyronnet, and S. Peyronnet. SpotRank: A robust Voting System for Social News Websites. In *Proceedings of the 4th workshop on Information Credibility,* pages 59-66, 2010.

[5] K. Lerman. Social Information Processing in News Aggregation. *IEEE Internet Computing,* volume 11(6): 16-28, 2007.

[6] K. Lerman. User participation in social media: Digg study. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology,* pages 255–258, 2007.

[7] S. Maroulis, E. Bakshy, L. Gomez and U. Wilensky. Complex Systems View of Educational Policy Research. *Science*, volume 330, 2010 (1st October).

[8] D. J. Myers. The Diffusion of Collective Violence: Infectiousness, Susceptibility, and Mass Media Networks. *American Journal of Sociology*, volume 106(1):173-208, 2000.

[9] D. P. Philips. The Influence of Suggestion on Suicide: Substantive and Theoretical Implications of the Werther Effect. *American Sociological Review*, volume 39(3):340-354, 1974.

[10] E. M. Rogers. New Product Adoption and Diffusion. *Journal of Consumer Research*, volume 2(4):290-301, 1976.

[11] B. V. Roy and X.Yan. Manipulation Robustness of Collaborative Filtering. *Management Science*, volume 56(11):1911-1929, 2010.

[12] P. Resnick and R. Sami. The Influence Limiter: Provably Manipulation-Resistant Recommender Systems. In *Proceedings of ACM Conference on Recommender Systems (RecSys07),* 2007.

[13] P. Resnick and R. Sami. The Information Cost of Manipulation-Resistance in Recommender Systems. In *Proceedings of ACM Conference on Recommender Systems (RecSys08),* 2008.

[14] T. C. Schelling. Dynamic Models of Segregation. *Journal of Mathematical Sociology*, volume 1:143-186, 1971.

[15] J. Warren and J. Jurgensen. The Wizards of Buzz. *Wall Street Journal*, February 10, 2007.