Scoring Procedures for Online Reputation Systems

Dawn C. Porter
McDonough School of Business
Georgetown University
Washington, DC  20057
(202) 687-5405
dcp22@georgetown.edu

J. Keith Ord
McDonough School of Business
Georgetown University
Washington, DC  20057
ordk@georgetown.edu

Bonnie R. Montano
McDonough School of Business
Georgetown University
Washington, DC  20057
brm7@georgetown.edu

Currently under review at Management Science
Scoring Procedures for Online Reputation Systems

Abstract

Reputation scores are important for decision-making in “blind” economic transactions, where the consumer, or purchaser, has little opportunity to gain knowledge about the provider. A number of Internet sites provide scoring systems for their products or sellers, depending on the type of E-business the site conducts. Some sites go further and monitor the reputation of the individuals (the “raters”) who provide the information from which the scores are derived. We point out the limitations of existing methods and provide a systematic way to develop and update scoring systems for both products, or sellers, and raters using the multinomial-Dirichlet distribution. We then consider starting values, which must be selected both to avoid undue barriers to entry and to provide reasonable assessments even when information is limited. We extend these procedures to allow the value of relevant information to decay over time and the weighting of evaluations to differ across raters. A final system allows regular updating of weights across raters. Although the focus is on Internet applications, the methods are generally applicable to any rating system. We provide an example using data on a digital camera, which illustrates the effects of different parameter settings upon the performance of the indices.

Keywords: E-business, Internet, multinomial-Dirichlet distribution, reputation scoring, online reputation system.
1. Introduction

The digital economy has assumed major significance over the last decade, with goods, services, and information being traded, sold (by providers), and purchased (by consumers) at an ever-increasing rate. The increasingly important role of E-business is highlighted by the proliferation of research developments in the area and entire special issues of journals dedicated to the topic (e.g., *Management Science* October and November 2003; *Information Systems Research* September 2002). In turn, its emergence raises a series of interesting research questions, not the least of which is whether E-business involves fundamental changes in the patterns of economic activity, or simply puts a new face on old paradigms. As might be expected, the answer is yes to both options.

Uncertainty is one of the features of E-business that distinguishes it from traditional business settings and is of interest to researchers and practitioners alike (Basu & Muylle 2003). E-business involves a high level of uncertainty about both provider behavior and the quality of the items being traded because traditional authentication mechanisms based on physical inspection are not feasible online (Basu and Muylle 2003; McKnight et. al 2002). Also, E-business is often characterized by provider-consumer pairings that exist for a single transaction, so there is no history between consumers and providers and considerations of future interactions do not constrain behavior as they do in a traditional bricks-and-mortar business setting (Resnick et. al 2000).

In a standard transaction with a bricks-and-mortar establishment, consumers have the ability to gather information about the reputations of the products and prospective providers through many channels. Such sources may include word-of-mouth, visits to the establishments, independent evaluations such as those provided by Consumer Reports (for products) or the Better Business Bureau (for providers), and so on. Consumers making E-transactions may have a huge amount of information available, but much of it has
not been screened or processed and may be of questionable quality. Thus, the set of effective resources is often more limited (Basu and Muylle 2003).

One major source of information about E-business establishments is input from consumers via online feedback mechanisms, a form of electronic word-of-mouth. Feedback serves to develop a reputation for providers. Thus, they are typically called online reputation systems. The details of these reputation systems vary. For example, Resnick and Zeckhauser (2002) describe eBay’s reputation system, in which only a user who has made a transaction or purchase may post an evaluation, whereas those set up like Amazon allow anyone to post an evaluation. Systems have also been set up to take reputation systems one step further by “rating the rater.” In these scenarios, feedback is given on those providing the evaluations to determine their believability (e.g., Epinions) (Resnick et al. 2000). Clearly the type of reputation system will vary with the nature and ownership of the site, and the types of products and services offered. What remains constant is the availability of a set of evaluations that a consumer might use in making a purchasing decision, thereby helping to reduce uncertainty.

1.1 Reputation systems and trust

The uncertainties inherent to E-business inhibit the development of trust between providers and consumers, thus creating a significant barrier to the growth of E-business (Bhattacherjee 2002; Dellarocas 2001). Shneiderman (2000) found that consumers are more likely to complete Web-based transactions when they are confident a trusting relationship exists. To engender a sense of trust, consumers typically seek reliable reports of past behavior (Shneiderman 2000) and online reputation systems are increasingly used to provide such reports. Such systems collect and analyze feedback about transactions between consumers and providers. Reputation systems have been shown to serve as quality indicators (Kollock 1999; Resnick et. al. 2000) and encourage trustworthiness in transactions by making past transactions publicly available for consideration in future transactions (Dellarocas 2001). Since trust is critical for the
success and expansion of E-business (McKnight et. al 2002) and reputation systems enable the development of trust, they too are critical for the success and expansion of E-business.

1.2 Value-added of online reputation systems

Both consumer activity and recent research suggest that reputation systems have value. Keser (2002) has shown that a reputation system significantly increases the level of trust among a group of strangers, relative to transactions among such a group without feedback. Bolton et al. (2004) found that the presence of feedback mechanisms increases trading efficiency, although not by as much as market interactions among partners who trade repeatedly with one another. Thus, in this paper we assume that a reputation system adds value to the consumer decision-making process, and our focus is on providing potential consumers with effective summaries of available evaluations, henceforth rating schemes.

Dellarocas (2003) provides an excellent state-of-the-art review of reputation systems and provides a listing of a number of major systems. He also summarizes the results from a number of studies that seek to measure the impact of such systems on E-business. Although the results are not entirely consistent, they generally show that positive feedback tends to increase the price paid and/or the probability of purchase; negative feedback has more variable consequences, but tends to reduce prices or the probability of sale. In short, rating systems do reflect useful information for consumers.

To illustrate this point, we collected data from eBay for one month of completed transactions of two similar video games that had the same Manufacturer’s Suggested Retail Price (MSRP). We calculated the final selling price as a percentage of the MSRP for the response variable. Since individual observations have less impact when a large number of evaluations are reported, and the sellers’ scores ranged from 0 to 32339, we used a logarithmic transformation of this variable, now \( \ln(\text{Score}) = \ln(\text{Score}+1) \), to create a proportional scale of consumer confidence. Regression results confirm the existence of a positive relationship between the price of an item or service and the reputation of the seller. The fitted model is:
%MSRP = 18.7 + 1.38 lnScore, R^2(adjusted) = 0.073

Although only a small proportion of the variance is explained, the positive coefficient of 1.38 has a p-value of 0.010, implying that sellers with higher reputation scores command a significantly higher price for identical items or services than those with lower scores. Therefore, better reputation scores tend to increase a consumer’s willingness to purchase an item at a given price level.

1.3 Design of a rating system

As noted by Dellarocas (2003) reputation systems are only valuable to the extent that consumers can trust them; he describes several features that might be built into the design of such systems to help ensure their integrity. We seek to answer the question raised by Dellarocas (2001): “What is the best way to design them [reputation systems]?

In this paper we will examine the properties of and develop a model for two different classes of E-business. The first class is a business-to-consumer (B2C) E-business, or storefront (e.g., Amazon), where a product is being rated by consumers. The second type is a consumer-to-consumer (C2C) E-business, typically an online auction site (e.g., Yahoo! Auction and eBay), where the product sellers are being rated. We now discuss certain weaknesses of existing reputation systems so as to determine the key requirements for an effective design.

Lack of an effective summary

Malaga (2001) cites examples of sites that either do not use reputation systems, or provide inadequate summaries of available data. For example, C2C E-businesses typically supply a single aggregate score constructed as the sum of individual scores from the set {-1, 0, +1}. The consumer does not have a clear impression of the seller’s reputation, however, because the distribution of positive, neutral, and negative scores is not readily available. Most B2C E-businesses give an Average Customer Review score and the total number of reviews made for an item but no information about the variability of the score.
Inability to adapt over time

If the reputation score is not updated, the information content is reduced. For providers, the lack of regular updating may either create an unfair barrier to entry (Malaga, 2001) or provide an incentive for cheating (Tullock, 1997). Similarly, as products are on the market longer, their scores should be appropriately updated and adjusted to reflect changes in technology and competition.

Lack of an effective start-up procedure

Malaga (2001) observes that many systems have a standard starting point (e.g. a score of zero). This may serve as a barrier to entry, thereby reducing the level of activity. An effective system should view a new entrant conservatively, but should also enable a new entrant to establish a reputation in a reasonable operational time frame.

An unlimited memory

Many systems continue to record new ratings and to give them equal weight with past evaluations. Yet, Dellarocas (2003, page 1419) observes that the “single most recent rating is just as efficient as a mechanism that summarizes larger numbers of ratings.” Similarly, Keser (2002) found that trading efficiency is only slightly higher if partners have the whole history of each other’s ratings rather than just the most recent rating. These findings imply the need to give greater weight to more recent findings.

Lack of a value-weighting

Some sites encourage feedback on the value of past evaluations (e.g. Amazon). It is evident that certain reviews are identified as useful and new potential purchasers continue to record the value of such reviews. When such information is available, there is potential benefit in providing selective weighting of evaluations. Such a step may also help to protect against filing false evaluations, since these entries would be flagged as being of little value.
In light of the above, our aim in this paper is to provide a reputation scoring system that satisfies the following requirements:

(1) It should provide an effective summary of available data,
(2) The scoring should adapt to new information over time,
(3) New entrants should be able to obtain timely evaluations so there is no unreasonable barrier to entry, and
(4) Other things being equal, more weight is given to recent evaluations.

1.4 Structure of the paper

In section 2 we develop the basic model for rating an object, such as a book, computer printer, or camera (and incorporate requirements 1 and 2). Section 3 then examines the issue of start-up, so that scores do not reflect initial over-optimism nor do they provide barriers to entry for new products or providers (requirement 3). In section 4, we extend the methods to deal with the unequal weighting of assessments to allow some evaluations to be recognized as more valuable than others, including the decay in the value of information over time (requirements 4 and 5). Section 5 presents a detailed example, illustrating the developments of earlier sections. Then, in section 6 we deal with creating a reputation score for the individual raters, also potentially including an element of time decay. This information can then be incorporated in the object rating. Finally, in section 7, we summarize the results and consider directions for further research.

2. The Basic Model

Online reputation systems typically evaluate one or more of the following:
Object Rating: where interest lies in rating a ‘good’ or ‘service’ based upon the ‘votes’ of interested parties;

Reputation Scoring: where we seek to evaluate the performance of the ‘sellers’ or ‘providers’; or

Rater Scoring: where we seek to evaluate the raters of an object to create a weighted object rating, a process sometimes known as ‘rating the rater’.

Since there is a similar structure for the development of both an object rating and a reputation score, we will refer only to an item (object or person) being evaluated by a rater. We now proceed to develop a statistical framework for the study of this phenomenon which will satisfy requirements 1 and 2 from section 1.3.

Suppose there is a K-point rating scale with possible values \( r_1 < r_2 < \cdots < r_K \). Each rater chooses one of \( \{r_i, i = 1, \ldots, K\} \) values and casts one vote per item (or person). In this section we assume each vote to be equally weighted, a requirement we relax later. Many systems [e.g. Amazon] employ a five-point scale with values 1-5 \((r_1 = 1, \ldots r_5 = 5)\), and assume that these values are interval-scaled, so that an average value may be computed. The choice of scale is beyond the scope of this paper, and we suppose the values \( \{r_i\} \) to be pre-specified.

We now develop a formal model. Denote the probability that a rater assigns score \( r_i \) by \( p_i \), written as

\[
\Pr(\text{rater assigns score } r_i) = p_i, \ i = 1, \ldots, K.
\]

We denote the set of probabilities by

\[
\{p\} = \{p_1, p_2, \ldots, p_K \mid p_i \geq 0, \sum p_i = 1\}.
\]

If we observe votes from \( n \) raters, denoted by

\[
x = \{x_1, x_2, \ldots, x_K \mid \sum x_i = n\},
\]

where \( x_i \) represents the total number of votes tallied in the \( i^{th} \) category, we may describe \( \{x\} \) by the multinomial distribution:
\[
\Pr(x_1, \ldots, x_K \mid n, \{p_i\}) = \left(\binom{n}{x_1 \ldots x_K} \prod_{i=1}^{K} p_i^{x_i}\right)
\]

(1)

Since our ultimate interest lies in the proportion of votes falling into each category, we next specify the prior distribution for \(\{p\}\) as:

\[
f\left(\{p\} \mid \{\alpha_i\}\right) \propto \prod_{i=1}^{K} p_i^{\alpha_i - 1}, \quad \sum \alpha_j = A, \quad \alpha_i > 0
\]

(2)

which is a standard Dirichlet distribution with parameters \(\{\alpha\} = (\alpha_1, \ldots, \alpha_K)\). For a discussion of the Dirichlet, see Stuart and Ord (1994, pages 271-72). Equations (1) and (2) lead to the Dirichlet posterior:

\[
f\left(p_1, \ldots, p_K \mid \{x_i\}\right) = \frac{\Gamma(A+n)}{\prod_{i=1}^{K} \Gamma(\alpha_i + x_i)} \prod_{i=1}^{K} p_i^{\alpha_i + x_i - 1}.
\]

(3)

2.1 A Score function

If we knew the population proportions \(\{\pi_i\}\) we would use the simple score function \(S = \sum r_i \pi_i\) where \(\pi_i = \lim E\left(p_i \mid x_i\right)\) as \(n \to \infty\). In the absence of such knowledge, we use the posterior distribution and consider the posterior mean as the estimated score function:

\[M_n = E\left(S_n\right) = \sum r_i E\left(p_i \mid \{x_i\}\right).
\]

(4)

We will often write \(E\left(p_i \mid \{x_i\}\right) = E\left(p_i\right)\) for notational ease, but the dependence on \(\{x_i\}\) should not be overlooked.

2.2 Inter-rater reliability

In order to provide an effective summary (requirement 1), we also need to consider the degree of consistency or reliability among raters. For example, suppose that an item has an average score of 3.0 out of a possible 5.0. We would have a very different impression of the item if we knew that (1) half the
ratings were 1.0 and the other half 5.0, or (2) all ratings were equal to 3.0. In the first case, we observe major disagreement among the raters, whereas in the second case we can safely conclude that the item is of modest quality. Therefore, a measure of inter-rater reliability \( IRR \) would be a useful accompaniment to the score function.

Let \( R_u \) and \( R_v \), where \( R_u \in \{r_1, \ldots, r_K\} \) and \( R_v \in \{r_1, \ldots, r_K\} \), denote the ratings provided by two individuals, \( u \) and \( v \). A measure of \( IRR \) may be based upon \( E[(R_u - R_v)^2] \), the mean squared difference between two ratings. However, if we assume that the two raters assign their scores independently, this measure reduces to \( 2V(R) \), or twice the variance of the rating random variable, \( R \). When the rating scale ranges from \( r_1 \) to \( r_K \), it is easily shown that the maximum value of the variance is \( (r_K - r_1)^2 / 4 \) so a possible measure of \( IRR \) would be:

\[
IRR = 1 - \frac{8V(R)}{(r_K - r_1)^2}
\]  

which ranges over \([-1, 1]\) with \( IRR = 1 \) denoting complete agreement and \( IRR = -1 \) corresponding to maximum disagreement.

For example, suppose that we have a 5-point rating scale with values (1, 2, 3, 4, 5). The measure reduces to \( IRR = 1 - V(R) / 2 \). Now assume that the ratings are equally spread over \( j \), where \( j \) is the cardinality of the set \( J \), where \( J = \{r_1, r_2, \ldots, r_K\} \). For example, \( j = 2 \) when \( J = (1, 2) \). The corresponding values of \( IRR \) are as follows:

<table>
<thead>
<tr>
<th>Subset, ( J )</th>
<th>Type</th>
<th>( j ) (# of scores)</th>
<th>Value of ( IRR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3, 4, 5</td>
<td>All scores equal</td>
<td>5</td>
<td>0.0</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>Any 4 adjacent</td>
<td>4</td>
<td>0.375</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>Any 3 adjacent</td>
<td>3</td>
<td>0.667</td>
</tr>
<tr>
<td>1, 2</td>
<td>Any 2 adjacent</td>
<td>2</td>
<td>0.875</td>
</tr>
<tr>
<td>1</td>
<td>Unanimous score</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>1, 3, 5</td>
<td>Major disagreement</td>
<td>3</td>
<td>-0.333</td>
</tr>
</tbody>
</table>

Table 1. \( IRR \) Values for Different Levels of Agreement
As a guide to users, descriptive labels could readily be assigned to sub-ranges of IRR, as in Table 1.

2.3 Accuracy

Our first requirement is that an object score both provides an effective summary of available data and accurately represents the overall population rating. This suggests that we incorporate a measure of variability or consistency in our summary measures.

We may measure the accuracy of the score function by its sampling variance, given by:

\[ V_n = V(S_n) = E\left( S_n^2 \right) - \left[ E(S_n) \right]^2 \]  \hspace{1cm} (6)

where

\[ E\left( S_n^2 \right) = \sum_i r_i^2 E(p_i^2) + \sum_{i \neq j} r_i r_j E(p_i p_j) \], \hspace{1cm} (7)

again suppressing the dependence on \( \{x_i\} \) in the expectations.

We can use the following properties of the Dirichlet distribution to evaluate the mean and variance:

\[ E(p_i | \{\alpha_i\}) = \frac{\alpha_i}{A} \]

\[ E(p_i^2 | \{\alpha_i\}) = \frac{\alpha_i (\alpha_i + 1)}{A(A+1)} \]

\[ (i \neq j) \ E(p_i p_j | \{\alpha_i\}) = \frac{\alpha_i \alpha_j}{A(A+1)} \]

with an obvious extension for the posterior distribution, when we replace \( \alpha_i \) by \( \alpha_i + x_i \) and \( A \) by \( A + n \).

Evaluation of the expectations in (4) and (6) leads to
\[ M_n = E(S_n) = \frac{\sum r_i (\alpha_i + x_i)}{(A+n)} , \]  

and

\[
E(S_n^2) = \frac{\sum r_i (\alpha_i + x_i)(\alpha_i + x_i + 1)}{(A+n)(A+n+1)} + \frac{\sum \sum r_i r_j (\alpha_i + x_i)(\alpha_j + x_j)}{(A+n)(A+n+1)} 
\]

\[
= \frac{1}{(A+n)(A+n+1)} \left[ \sum r_i^2 (\alpha_i + x_i) + \left[ \sum r_i (\alpha_i + x_i) \right]^2 \right].
\]

Then we can see that

\[
V(S_n) = \frac{1}{(A+n)(A+n+1)} \left[ \sum (\alpha_i + x_i)(r_i - M_n)^2 \right].
\]

Thus the reputation score is given by the posterior mean \( M_n \), and \( V(S_n) \) may be used to measure the closeness of that score to the population mean score.

Clearly \( V_n = V(S_n) \) depends upon both the number of respondents \( n \) and the spread of the scores. A natural statistical measure to represent this spread would be the standard deviation (SD) or some multiple of it. For example, we could use the approximate posterior probability interval:

\[ P(\text{average score lies within } \pm 2 \text{ SD}) \approx 0.95, \]

suitably truncated at the ends of the range. This will be rather inaccurate when \( n \) is small, especially if many of the values are recorded at one end of the range. Nevertheless, reporting some range of mean \( \pm c \cdot \text{(SD)} \), for suitably chosen \( c \), seems the best way to indicate the accuracy of the estimated score function without becoming involved in complex calculations.

### 2.4 Updating the Summary Measures

A strength of our model is that the scores are readily updated, thereby satisfying requirement 2. Thus, if the \((n+1)\)th observation yields value \( r \), the new mean is
\[ M_{n+1} = M_n + \left( \frac{r - M_n}{A + n + 1} \right) \]

and the variance is

\[ V_{n+1} = V_n \left[ 1 - \frac{2}{A + n + 2} \right] + \frac{(A + n)(r - M_n)^2}{(A + n + 1)^2 (A + n + 2)}. \]

Thus the only quantities that need to be stored for each object evaluated are: \( A+n, M_n, V_n \), and the latest score, which is an important issue if scores for a large number of items are to be recorded and updated.

Finally, we observe that \( V(R) \), defined in section 2.2, is given by \( V(R) = (A + n + 1)V(S_n) \), so that only one set of calculations is needed. This result can be seen by demonstrating that \( V(R) \) is so defined, from equation (10).

Thus, in this section we have developed a rating scheme that meets requirements 1 and 2 from an underlying statistical framework that allows for easy updates, provides information on the level of agreement among raters, and also measures the accuracy of the reported rating.

### 3 Initial Conditions

Before any votes are recorded for an item, an initial score must be given (requirement 3). It is imperative that this initial score is somewhat conservative so a “good” reputation has to be earned, but it should not be so low as to be a barrier to entry (Malaga, 2001). For example, if the item is given an initial rating of 5, but the successive actual ratings are all 2s and 3s, the audience is given a false impression of the rating of the item to start. The initial score distribution should be sufficiently diffuse so that actual ratings are reflected in the current score relatively quickly. Conversely, the first one or two ratings should not dominate the prior information. Malaga (2001) reviewed seven electronic markets and four online communities and found that, in all of those systems, the initial rating of an object or person was 0 (or
similarly defined “neutral” score). This leads to a barrier to entry in that system. Although a provider may use a discounted price to offset a lower reputation, that person may be forced out of the system if the required discount is too steep.

Thus, we are led to propose that the updating procedure should:

- Reflect decreasing uncertainty when new ratings are consistent with previous ratings
- Allow for increased uncertainty when new ratings differ sharply from old ratings
- Not allow the final score to be affected by the order of presentation of the ratings (should not be time dependent, a requirement that we relax in section 4.1).

3.1 Initial scoring

One approach is to select the coefficients directly, based on prior considerations. However, we recommend this choice only when the prior information is very weak. Quite often, we will have available sample ratings taken across N comparable earlier items (e.g. movies, books, cameras, etc.) using the same rating scale \( \left( r_1, r_2, \ldots, r_K \right) \) as illustrated in section 2.2. We suggest two possible weighting schemes for creating prior distributions.

*Proportional weighting.* If we observe \( x_{ij} \) observations for rating \( r_i \) on item \( j \), and \( n_j \) observations for item \( j \), we could specify the initial proportions \([a_i = \alpha_i]\) as

\[
a_i = \frac{x_{i1} + \cdots + x_{iN}}{n_1 + \cdots + n_N}.
\]

*Equal weighting.* An alternative that gives equal weight to each comparable item is

\[
a_i = \sum_{j=1}^{N} \frac{x_{ij}}{n_j}.
\]
If we decide that this prior information is “worth \( n_0 \) observations” we could specify \( A = n_0 \) so that the prior coefficients become \( \alpha_i = a_i n_0 \). However, keeping in mind the need to be somewhat conservative we might reduce the weight in the top class(es) to reflect the need to earn a good score. For example, initial scores may be defined for all products in the same way, with a recognized “downward bias” that signals newness but is not so severe as to stifle new entrants. Alternatively, if the product has some antecedents, such as a book by an established author or a new model of digital camera from a major manufacturer, the producer’s distribution may be shrunk towards the common prior by some approved weighting scheme.

It is important to recognize that the choice of prior is important only in the start-up phase, as can be seen from the examples presented in section 5 below. However, as noted earlier, this phase is crucial if we are to avoid both barriers to entry and misleading initial scores. Indeed, without sound initial conditions later observations may not exist. Further, items may have only a few evaluations, so that the prior information continues to have a significant impact.

It may be shown that the mean and variance after \( n \) observations, given in (8) and (10) combine the initial and data based values in the following way:

\[
M_n = E(S_n) = \frac{\sum r_i (x_i + \alpha_i)}{(n + A)} = \frac{1}{(n + A)} \left[ n\bar{r}_n + A\bar{r}_0 \right]
\]

\[
V(S_n) = \frac{AV_0 + nV_n}{(A + n)(A + n + 1)} + \frac{An(\bar{r}_0 - \bar{r}_n)^2}{(A + n)^2 (A + n + 1)}
\]

where \( \bar{r}_n = \sum \frac{r_i x_i}{n}, \bar{r}_0 = \frac{\sum r_i}{K}, V_0 = \frac{\sum (r_i - \bar{r}_0)^2}{A} \) and \( V_n = \frac{\sum x_i (r_i - \bar{r}_n)^2}{n} \). Generally, \( V(S_n) \) will be \( O(n^{-1}) \) and \( V_n \) will dominate as \( n \) increases.
We will also examine the effects of different choices for \( n_0 \) in 2 special cases in section 5.

(1) *Uninformative prior (Classical index)* – all \( \alpha_i \to 0 \), so that \( A = 0 \).

\[
M_n = E(S_n) = \bar{r}_n \text{ and }
\]

(2) *Uniform prior* – all \( \alpha_i = 1 : p_i^{n_0-1} = p_i^0 = 1 \) and \( A = K \).

\[
M_n = E(S_n) = \frac{\sum r_i (x_i + 1)}{(n + K)} = \frac{1}{(n + K)} [n\bar{r}_n + Kn_0]
\]

The advantage of case (2) is that it gives a non-zero variance whereas case (1) always gives \( V(S_i) = 0 \) after one observation, which is clearly undesirable. Indeed, when \( x_j = n, \ x_i = 0, \ i \neq j \) we still obtain \( V(S_n) = 0 \) for case (1), which gives a spurious indication of accuracy for small \( n \).

4. **Extended Weighting Schemes**

Several extensions to the scheme just developed may be considered which will meet requirement 4 from section 1.3. There are cases in which certain raters, or ratings, may be considered more or less valuable than others. It may be useful, therefore, to incorporate a set of weights that will adjust the overall score appropriately. For example, the rate of product development may be sufficiently rapid such that older evaluations become less valuable than newer ones. Such a decline in the value of an assessment is particularly likely for high-tech products, but may apply quite broadly (e.g. a new edition of a textbook).

A second case of interest arises when some raters are more heavily weighted than others (when E.F. Hutton speaks, people listen). We now explore each of these extensions in turn.
4.1 Rating with time decay

Using an entropy-based argument, or a discounted likelihood, an approximation to Bayes Theorem (c.f. Harvey and Fernandes, 1989) enables the following approach. The likelihood after $n$ observations may be written as:

$$L_n \propto \prod_{i=1}^{K} p_i^{z_i(n)}$$

where

$$z_i(n) = wz_i(n-1) + d_i(n)$$

and we define the indicator variable

$$d_i(n) = \begin{cases} 
1 & \text{if } n^{th} \text{ rater chooses class } i \\
0 & \text{otherwise}
\end{cases}$$

We may now specify the prior distribution by

$$z_i(0) = \alpha_i$$

The coefficient $w$ denotes the discount factor, where $0 \leq w \leq 1$. There is no discounting when $w = 1$, and $w = 0$ implies that only the most recent rater is being used. More general weighting schemes are possible, but the use of a discount factor has proved both popular and successful in exponential smoothing, so that is the formulation we pursue here.

Using similar arguments to those applied earlier, we find that the mean becomes:

$$M_n = E(S_n) = \sum r_i E(p_i|z_i(n)) = \frac{\sum r_i z_i(n)}{Z_n}$$

where $Z_n = z_1(n) + \cdots + z_K(n)$. Likewise, the variance is given by:

$$V(S_n) = \frac{1}{Z_n(Z_n+1)} \left[ \sum r_i^2 z_i(n) - Z_n M_n^2 \right] = \frac{1}{Z_n(Z_n+1)} \left[ \sum z_i(n)(r_i - M_n)^2 \right].$$

Finally, $V(R) = (Z_n + 1)V(S_n)$. 

18
4.2 Unequal weighting of raters

We now assume that the \( m \)th rater has weight \( W_m \), which may be assigned objectively by some rating system, or it could be a subjective weight assigned by an individual user of the rating system. The general form of the likelihood may be written in the form of (13) as before, but with

\[
z_i(n) = \sum_{m=1}^{n} d_i(m) W_m
\]

so that the weighted score is collected for each category. The prior is specified as before and the results are given by (15) and (16) with the changed form of the z-function given by (17).

We may also implement a combination of weighted raters and time decay by indexing the order of presentation of the raters and applying the discount factor as in section 4.1.

When the user is concerned that the rating scheme does not represent his or her potential preference, a personal set of weights would be more desirable. Some sites provide this option in a rudimentary way with inputs such as “People who bought A (your most recent purchase) also bought B, C,… . Clearly, if sufficient information was stored about the current user and about each reviewer, personalized rating schemes could be generated using (17).

5. Canon Powershot S50: an example

To illustrate the use of the rating system we have developed, we extracted data for a digital camera from the listings of Amazon. Due to the availability of information and ease of presentation, we chose to create an example using data from a B2C E-business rather than a C2C site. Tests of the effectiveness of the proposed rating system for C2C sites are more effectively pursued in a controlled experimental environment; c.f. Keser (2003) and Bolton et al., (2004).
For each product, Amazon provides a set of reviews contributed by members of the public. Each reviewer rates the item on a five-point scale [1 through 5, where 1 is the least and 5 is the most desirable score], and Amazon records a current mean score with equal weighting across all reviewers. The text and numerical value of each review is displayed and is also eligible for rating by other users. Many other sites have similar scales, some with only two or three points, however, but the methods described are quite general and the choice of example should not be viewed as a restriction on the applicability of the approach.

The camera we selected was the Canon Powershot S50 with 5.0 megapixels. Although any of a large number of items could have been chosen and would provide insights into the operations of the rating schemes developed in earlier sections, this one offers several advantages:

- Digital cameras are very popular items, thereby being frequently researched by consumers;
- Technology changes rapidly, so there is a necessity for continued reviews;
- The duration of time over which the camera has been reviewed, coupled with the fast-changing technology, allows us the opportunity to evaluate time series decay;
- This particular model had a relatively large number of reviews;
- The brand is very well-known.

We recorded the evaluations for the camera over a nine-month period from the beginning of April 2003 (the date of the first review) through to the beginning of January 2004, when the data were assessed for this paper. The raw data appear in the Appendix, yielding 58 reviews in all.

5.1 The unweighted index

In order to analyze the raw data, eight different prior distributions were considered. Each is represented below in Table 2:
(1) The diffuse prior with all initial weights equal to zero, which yields the traditional score function [prior 8];

(2) The uniform prior with all initial weights equal to 1.0 so that the prior distribution is weighted as five observations and has a mean of 3.0 [prior 3];

(3) A series of six other unequal weighting schemes which set the starting scores at various different levels. See table below for actual chosen values. Prior 3 represents the strongest prior information with A = 5, followed by prior 1 with A = 3.5.

Table 2. Various Choices of Prior Distributions

<table>
<thead>
<tr>
<th>priors</th>
<th>r₁</th>
<th>r₂</th>
<th>r₃</th>
<th>r₄</th>
<th>r₅</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>Similar to prior 3</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>Similar to prior 3</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>Similar to prior 3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>Similar to prior 5</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0</td>
<td>Similar to prior 8</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>Similar to prior 8</td>
</tr>
<tr>
<td>7</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Similar to prior 8</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Similar to prior 8</td>
</tr>
</tbody>
</table>

Prior distributions (1) and (6) were selected to reflect that the manufacturer’s previous cameras were well-regarded so that the new one was likely to be favorably received. Priors (2), (3) and (8) are uniform with priors of varying strength, whereas distributions (4) and (5) are structured to center the initial scores close to a value of 3.0 until sufficient evidence has been collected to demonstrate otherwise. Choice (7) reflects some skepticism towards a new product.

Clearly, the posterior distributions will converge as the number of observations increases. The reason for being concerned about the choice of prior is that we need reliable feedback to consumers even when the number of responses is small. This is especially important if the prior distributions are heavily weighted at the low end of the rating scale, as in prior (7), which could result in a barrier to entry issue for the new product. The effects of the different priors in this case are illustrated in Figure 1.
An examination of Figure 1 quickly reveals that prior (7) is too pessimistic for this particular product. After 5 observations have been recorded the mean (4.12) is well below the raw sample mean (4.38). A prior distribution with a low mean is inappropriate when prior information about the product indicates a higher set of scores is likely. By contrast, the diffuse prior (8) is too erratic and reflects the chance event that the first score was a ‘5’. Although all lines appear to stabilize after about 20 or so observations, there is still a slight difference in their means. Overall, priors (1), (3) and (5) seem to combine a slight conservatism with a reasonable speed of adjustment to a stable level.

As an alternative to Table 2, we could utilize the informative prior methods described in section 3.1. Evaluations based upon informative priors are plotted below in Figure 2. We restricted attention to priors with an overall weight \( A \) equal to 1 by construction, and Figure 2 includes the three choices from Table 2 that also have \( A = 1 \). The choice of \( A = 1 \) has no particular significance; we made it to keep the ‘amount’ of prior information constant across the priors. The two informative priors from section 3.1, proportional and equal weighting schemes, were calculated from the Amazon site for the camera data using 10 other 5.0 megapixel digital cameras of similar quality. The data and results can be seen below in
Table 3. Since the two distributions differed by no more than 0.01 on any rating level and the charted results were almost indistinguishable, only one of the lines was plotted in Figure 2. As might be expected, the informative prior produces a stable level more quickly than the others.

Table 3. Informative Prior Computation

<table>
<thead>
<tr>
<th>Data taken 1/17/04 from Amazon.com</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Score</strong></td>
</tr>
<tr>
<td>Sony DSCV1</td>
</tr>
<tr>
<td>Olympus Camedia C-5050</td>
</tr>
<tr>
<td>Canon PowerShot G5</td>
</tr>
<tr>
<td>Sony DSCP10 Cybershot</td>
</tr>
<tr>
<td>Olympus Camedia C-50</td>
</tr>
<tr>
<td>Sony DSCF717</td>
</tr>
<tr>
<td>HP PhotoSmart 935</td>
</tr>
<tr>
<td>Minolta Dimage F300</td>
</tr>
<tr>
<td>Sony DSCP92 Cybershot</td>
</tr>
<tr>
<td>Nikon Coolpix 5400</td>
</tr>
<tr>
<td><strong>totals</strong></td>
</tr>
<tr>
<td><strong>(1) initial proportions</strong></td>
</tr>
<tr>
<td><strong>(2) equal weights</strong></td>
</tr>
</tbody>
</table>

Figure 2. Effects of Informative and Uninformative Priors on Early Means. All Priors Sum to 1.0
The other question we need to examine is the degree of inter-rater reliability (IRR) discussed in section 2.2. Figure 3 shows the changes in IRR for the same choice of priors used in Figure 2. In general, the diffuse prior (8) and single point priors such as (4), (6) or (7) may produce erratic IRR values at the outset. Overall, a prior such as (1) or a slightly scaled-down informative prior appear to provide a good balance of initial conservatism, reasonable adjustment speed and a stable IRR. Again, the informative prior stabilizes more quickly and seems to give rather accurate results as the number of evaluations increases. Finally, we note that an approximate posterior probability statement of the form ‘mean ± c standard deviations’ provides a reasonable summary of the rating accuracy. For example, with c = 1, 68% coverage is provided by the interval 4.65 ± 0.18, or [4.47, 4.83], using prior (4) after 20 observations. An alternate way of expressing the uncertainty would be to cite 0.18 as the ‘margin of error’, as done in sample surveys, particularly in reports for public consumption.

Figure 3: Change in IRR Values for Different Priors

5.2 Discounting over time

As we observed at the beginning of this section, the data occur over a nine-month period. Since the technology used for the object being rated is rapidly changing, it is useful to examine the effect of
discounted ratings over time. Electronic items are especially susceptible to rapid innovation so that a particular model of digital camera, for instance, has the tendency to become obsolete over time.

The results are shown in Figure 4 below. Referring to equation (14), it is obvious that assigning a weight of \( w = 1.0 \) is equivalent to the basic model without any time decay; each observation is worth the same amount. This is seen in the figure as the smoothest line. Any degree of discounting shows the slight decline in ratings in the later part of the period, and a corresponding increase in the standard deviation. Discounting by a factor of 0.5 appears excessive in this case, producing wild swings in the rating, and is not included in the figure. The value of \( w = 0.90 \) appears to strike a reasonable balance between the needs for stability and flexibility.

More generally, when the times between successive evaluations are unequal, and time is thought to be important, we might consider \( w \) to be time-dependent. For example, we could set

\[
w_n = \exp\left[-b\left(t_n - t_{n-1}\right)\right],
\]

(18)
where $t_n = \text{time of } n^{th} \text{ evaluation}$. Then, if an opinion 6 months ago is valued at 50% of a current view and $t_n$ is measured in months, $b$ can be computed; that is, we would use (18) to set $0.5 = \exp(-6b)$, which yields $b = 0.116$. This approach has the advantage that we allow for aging and need to store only $(b, t_{n-1})$ for real-time updating. Clearly, $b$ can vary by product. Also, updating can be performed in batches rather than one at a time to redefine $d_i(n)$.

5.3 Evaluation of the results

The reviews used in this example are unsolicited and they are posted without comment. This procedure is fairly standard but it has two limitations:

(i) reviewers are self-selected, so that they do not form a random sample of an appropriate population;

(ii) review quality is not directly assessed.

These limitations do not overshadow the value provided by online reputation systems. The selection of a random set of reviewers is an almost impossible goal. What should be the population: users, prospective users, all users of the system? The question does not seem capable of a reasonable answer. Thus, users of this and most other rating schemes must ask themselves whether they accept the review process as valid for their needs.

The second limitation can be addressed through an additional feature in the rating system. Amazon has a rudimentary system whereby review quality is assessed by reporting “x out of X people found this review helpful”. Such information may provide the basis for a weighted rating scheme. We now proceed to examine this issue in section 6.

6. Rating the raters
As noted in section 5, the approaches considered so far do not allow for the quality of the raters to be assessed, except by some prior mechanism. A more desirable state of affairs would be to allow users to “rate the rater” and then to incorporate these evaluations into the assessment of the object. Rating the rater is also of direct interest on E-auction sites such as eBay, where seller reputation is critical.

Once a rating scale has been decided for the raters, updating proceeds as in section 3. However, we must then decide how to translate these scores into weights for the purpose of evaluating the object. In the above example using the Canon Powershot S50 digital camera, we have gathered information from the Amazon website as suggested at the end of Section 5. Each rating, 1, …, n, has associated with it two values: a total number of people who read the review and the number of those people who found the review helpful. Below is a chart summarizing the data. It is interesting to note that users appear to find the more positive reviews more helpful.

Figure 5. Reviews per Score
A full discussion about the appropriate choices for modeling the rater weightings is beyond the scope of this paper. However, for purposes of example, we used a simple linear progression to assign weights as follows:

\[
   w_{mn} = f \left( H_m(n) \right) = \left[ 1.0 + B H_m(n) \right] / C_n, \quad m = 1, \ldots, n
\]

where \( B \) is a non-negative constant. At the point when there are \( n \) total evaluations, \( H_m(n) \) represents the number of people who felt the \( m^{th} \) evaluation was helpful. Note that it is possible to comment on the usefulness of an evaluation of an item without having actually purchased the product. Buyers are also not required to rate a purchased item or comment on an evaluation. Therefore, the value of \( H_m(n) \) is not directly dependent on the total number of evaluations, \( n \). \( C_n \) is a scaling constant chosen to make the average weight equal to 1.0; that is, we rescale the weights each time a new rater appears, or even whenever \( H_m(n) \) changes. When \( B=0 \), \( C_n=1 \) for all \((m,n)\).

Figure 6 shows the progression over all \( n \) raters for various choices of \( B \), using the informative prior of section 5.1. It should be noted that this figure uses the weights \( \{w_{mn}\} \) for all evaluations; a real-time application would use \( w_{mq} \) when \( q \) evaluations were available, progressively updated. We see from that the mean level increases with \( B \), reflecting that good reviews are more highly rated in our example.

Figure 6. Weighted Raters Using Informative Priors
7. Summary and directions for further research

E-business has profoundly transformed the way business is conducted. Thus, it is of intense interest to researchers and practitioners alike. Past research has shown that one of the barriers to E-business is the uncertainty inherent in transactions between partners who do not have a shared history. Online reputation systems work to minimize this uncertainty by providing consumers with information about providers’ past behavior.

We have developed a scoring method for use in online reputation systems. An assessment of current approaches led us to identify a number of weaknesses, which we have tried to remedy in this paper. Our first concern was to develop a more complete summary that not only gives average scores but also provides a measure of inter-rater reliability, so that a user may see how consistent the ratings are across raters. We then developed a framework for creating reasonable starting conditions, both to avoid barriers to entry and to ensure that ratings based upon very limited data are recognized as having a wide margin of error. The proposed measures are easy to update, and can readily be maintained to reflect the latest inputs.
to the system. We have also developed more general weighted measures to allow for differing expertise among raters, or a decay in the value of information over time. Finally, we extended the approach to allow for ‘rating the rater’ so that weights can be modified in the light of rater performance. An extended example, based upon B2C E-business data, illustrates the performance of these various summary measures.

The focus of the paper has been the development of a reliable online reputation system. The next question is the extent to which consumers might base their decisions upon such information. Several avenues for future empirical research are being considered:

1. Testing the effectiveness and added value of the proposed measures, relative to existing procedures
2. Using the proposed measures in an experimental C2C E-business framework and testing its value relative to complete alternative summaries of available information
3. Checking to see whether positive reviews do indeed receive more weight and, if so, whether this reflects greater information content in such reviews or that potential buyers are seeking positive reinforcement prior to making a purchase.
4. Examining whether either the B2C or C2C scoring models can apply to B2B E-business.

Acknowledgement

The authors are grateful to Ross Malaga for a thought-provoking seminar on reputation systems, which stimulated this research, and his subsequent helpful comments.
References


Appendix. Ratings on the Canon Powershot S50 Digital Camera

Data taken 1/08/04 from Amazon.com

<table>
<thead>
<tr>
<th>Date</th>
<th>Index</th>
<th>Score</th>
<th># people</th>
<th># helped</th>
<th>Date</th>
<th>Index</th>
<th>Score</th>
<th># people</th>
<th># helped</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/4/2003</td>
<td>1</td>
<td>5</td>
<td>61</td>
<td>60</td>
<td>8/8/2003</td>
<td>30</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4/13/2003</td>
<td>2</td>
<td>5</td>
<td>26</td>
<td>25</td>
<td>8/18/2003</td>
<td>31</td>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>4/19/2003</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>8/25/2003</td>
<td>32</td>
<td>5</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>5/22/2003</td>
<td>5</td>
<td>5</td>
<td>21</td>
<td>19</td>
<td>8/30/2003</td>
<td>34</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5/31/2003</td>
<td>6</td>
<td>5</td>
<td>31</td>
<td>30</td>
<td>9/6/2003</td>
<td>35</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5/31/2003</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>9/15/2003</td>
<td>36</td>
<td>4</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>6/6/2003</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>9/23/2003</td>
<td>37</td>
<td>2</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>6/9/2003</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>9/28/2003</td>
<td>38</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>6/12/2003</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>9/28/2003</td>
<td>39</td>
<td>4</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>6/12/2003</td>
<td>11</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>10/6/2003</td>
<td>40</td>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>6/17/2003</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>10/8/2003</td>
<td>41</td>
<td>5</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>6/20/2003</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>10/17/2003</td>
<td>42</td>
<td>3</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>6/23/2003</td>
<td>14</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>10/21/2003</td>
<td>43</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>6/26/2003</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>10/22/2003</td>
<td>44</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6/29/2003</td>
<td>16</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>10/27/2003</td>
<td>45</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7/2/2003</td>
<td>17</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>10/28/2003</td>
<td>46</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>7/5/2003</td>
<td>18</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>11/9/2003</td>
<td>47</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7/7/2003</td>
<td>19</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>11/11/2003</td>
<td>48</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>7/8/2003</td>
<td>20</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>11/12/2003</td>
<td>49</td>
<td>5</td>
<td>28</td>
<td>26</td>
</tr>
<tr>
<td>7/9/2003</td>
<td>21</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>11/21/2003</td>
<td>50</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>7/11/2003</td>
<td>22</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>11/21/2003</td>
<td>51</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>7/26/2003</td>
<td>23</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>11/23/2003</td>
<td>52</td>
<td>5</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7/29/2003</td>
<td>25</td>
<td>1</td>
<td>17</td>
<td>3</td>
<td>12/9/2003</td>
<td>54</td>
<td>4</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7/29/2003</td>
<td>26</td>
<td>1</td>
<td>17</td>
<td>3</td>
<td>12/15/2003</td>
<td>55</td>
<td>4</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>7/31/2003</td>
<td>27</td>
<td>5</td>
<td>16</td>
<td>15</td>
<td>12/25/2003</td>
<td>56</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>8/1/2003</td>
<td>28</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>12/30/2003</td>
<td>57</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8/4/2003</td>
<td>29</td>
<td>5</td>
<td>12</td>
<td>10</td>
<td>1/2/2004</td>
<td>58</td>
<td>5</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>