A multi source product reputation model

Umar Farooq, Antoine Nongaillard, Yacine Ouzrout, Muhammad Abdul Qadir

A Multi Source Product Reputation Model

Umar Farooq[a,c,d], Antoine Nongaillard[b], Yacine Ouzrout[a], Muhammad Abdul Qadir[c]

[a]DISP Laboratory, University Lumiere Lyon 2, 160 bd de l'Université, 69676 Bron cedex, France,
Umar.Farooq@univ-lyon2.fr,yacine.ouzrout@univ-lyon2.fr
[b]Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL (SMAC), F-59000 Lille, France,
antoine.nongaillard@univ-lille1.fr
[c]Computer Science Department, Capital University of Science & Technology, Islamabad, Pakistan
aqadir@cust.edu.pk
[d]Computer Science Department, Abdul Wali Khan University Mardan, Pakistan,

Abstract

Product reputation model is very important for customers and manufactures in order to make decisions. Several product reputation models are proposed in literature which use customer reviews in order to compute reputation values. However, the aggregation methods used are not able to estimate a true reputation value when some ratings are false. Some of these aggregation methods are not robust to false and biased ratings because a single false rating is enough to change the result. Others are robust to false ratings but not able to reflect the recent opinions about product quickly. In addition, most of the product reputation models are based on single source, therefore suffer from availability and vulnerability issues. In this paper, we propose a multi-source product reputation model where robust and strategy proof aggregation methods are used. A source credibility measure method is proposed, which uses four factors to determine malicious sources. Furthermore, a suitable decay principle for product reputation is also introduced in order to reflect the newest opinions quickly. The results show that proposed model is robust, strategy proof and able to estimates a true reputation value even if some ratings are false.

Keywords: Product Reputation Model, Reputation System, Rating Aggregation, Product Evaluation

1. Introduction

With the rapid growth of Internet, people have the opportunity to express their opinions about products and services on the Web. Several websites such as e-commerce and review sites allow users to post evaluation information about different products, which can be used to compute product reputation. Product reputation is a perception about product quality and future behaviors. Several reputation values such as aggregated star value (also called five star or simply star value) (Garcin et al., 2009a), feature reputation (Hu and Liu, 2004; Abdel-Hafez et al., 2012) and product reputation based on features (Abdel-Hafez et al., 2012) are computed in literature in order to assess the reputation of a product. These reputation values are useful for both customers and organizations to make decisions. The customers can use the reputation values to compare different products in order to make purchase decision. Similarly, the manufacturers can use the reputation values to know the customers opinions in order to improve their products and to launch different marketing strategies. The product reputation systems can be divided into two types based on the type of rating which is considered to compute reputation value. Two types of ratings (i.e. numeric and textual ratings) are aggregated by the existing reputation models. Several online product reputation systems are available such as Amazon, Ebay and Cnet etc., where both numeric and textual ratings are posted. However, these rating sites aggregate numeric ratings using simple arithmetic mean to determine a single reputation value (Sun, 2012; Abdel-Hafez et al., 2012; Garcin et al., 2009a). Users can also read textual ratings to know the customers opinions and to make purchase decision. However, reading all the reviews about a product is time consuming. On the other hand, some authors also proposed reputation systems based on sentiment analysis to analyze textual ratings in order to produce a summary of opinions either about product or product features (Hu and Liu, 2004; Popescu and Etzioni, 2007; Morinaga et al., 2002; Turney, 2002).

All these reputation systems use summation, arithmetic mean or weighted mean as an aggregation method in order to aggregate different types of ratings. These aggregation methods assume that all users give honest ratings. However, this is not always the case especially in industrial products, some users may post false reviews in order to promote their own product or to damage the reputation of competitor product. Therefore, the aggregation methods are not able to estimate a true reputation value when some ratings

are false. In addition, these methods are not robust to biased and false ratings because a single false rating is enough to move the reputation value up or down. Mean and weighted mean are also not strategy proof because they provide sufficient incentive to opportunistic users to change the reputation value substantially by posting false ratings (Garcin et al., 2009a). An aggregation method is said to be strategy proof if it does not provide incentive to a reviewer in order to obtain their preferred reputation value by lying or hiding the actual evaluation. Some authors suggested to use robust aggregation methods such as mode and median. However, these methods are not able to reflect the recent opinions about product quickly and also treat all the ratings equally regardless of their importance.

Most of these reputation systems are based on single web source. One of the significant issue in single source based reputation systems is the vulnerability to falsifying information (Josang et al., 2007). Since the ratings are obtained and stored on a single location, therefore the malicious users can easily post false ratings. In addition, sometimes single source based reputation systems lack evaluation information because these systems have only one source to obtain ratings. Furthermore, the credibility of the source or central server is another issue (Josang et al., 2007). When the source itself is a buyer, seller or intermediate then how we can trust the reputation system? Does the source adopt sophisticated security measures so that the credibility of evaluation information can be trusted?

The existing product reputation systems are also not able to reflect the recent opinions about product because the same weight is assigned to recent and old ratings. Several decay principles are proposed for e-commerce and peer to peer network (Xiong and Liu, 2004; Kinateder et al., 2005; Jsang and Ismail, 2002; Jøsang et al., 2003; ElSalamouny et al., 2009; Selcuk et al., 2004; Kinateder et al., 2005; Zacharia and Maes, 2000), however these methods are not suitable in product reputation context because of three main reasons. First, these decay principles are not robust to false and biased ratings, second these methods are not strategy proof because they provide sufficient incentive to bring reputation value from highest to lowest or from lowest to highest with few recent ratings. Third, these methods are also not able to estimate a good reputation value. On the other hand, rating parameters which increase accuracy and reliability of reputation values such as ratings trustworthiness and source credibility are also ignored by the existing reputation models. Rating trustworthiness determines that whether the rating or review posted on e-commerce or review site is genuine or fake (i.e. false, biased and spam).

Source credibility determines the extent to which a web source from where ratings are extracted is credible to be considered for product reputation.

Considering these issues, in this paper we propose a multi-source product reputation model which offers several benefits over single source based approach. A method which uses four factors to rank the credibility of a web source is proposed. In addition, robust and strategy proof aggregation methods are introduced which provide a good approximation of the true reputation value even if some ratings are false or biased. Furthermore, two decay principle methods, which are more suitable in product reputation context, are proposed in order to reflect the recent opinions about the product. Besides decay principle and source credibility, rating trustworthiness is also considered, which increase the reliability of the determined reputation value.

This paper is organized as follows. In Section 2, related works are presented and analyzed. Section 3 gives an overview of the proposed model. Section 4 explains different rating parameters on the bases of which rating credibility is computed. Section 5 describes the aggregation method proposed to compute aggregated star value. Finally, Section 6 describes the experimental settings and results.

2. Related Work

The literature review is divided into three subsections. The first subsection gives an overview of the reputation models, the second subsection discusses the existing aggregation methods and the last subsection summarizes the literature about decay principle.

2.1. Product Reputation Model

Many authors have investigated reputation models in last few years, most of them devoted their efforts to user trust and reputation in e-commerce environment, very few of them focused on product reputation. The most common product reputation model is based on opinion mining, which analyze textual ratings in order to form a summary of opinions either about product and/or product features. Several supervised and unsupervised methods have been proposed in literature (Hu and Liu, 2004; Popescu and Etzioni, 2007; Morinaga et al., 2002; Turney, 2002). Some of these methods determine the global opinions about products and others are more refined which summarize the opinions about different features of the product. In all product reputation

models based on opinion mining the researchers focused on sentiment analysis instead of mathematical modeling, therefore simple summation is used as an aggregation method. In (Abdel-Hafez et al., 2012) a mathematical model based on opinion mining has been proposed. The reputations of all product features are calculated based on the ratio of positive and negative opinions. The important features are given more weight while computing reputation. However, the overall product reputation is computed using simple weighted mean, which has several issues (Garcin et al., 2009a,b). In addition, this model is based on manual hierarchy of product features and sub-features, while determining feature reputation positive opinions about sub features are also considered. However, developing a manual hierarchy for every product is difficult and time consuming. Moreover, no proper experiment is performed to validate the results. On the other hand, several online product reputation systems are also available such as Amazon, Ebay and Cnet which aggregate numeric ratings using simple arithmetic mean to determine product reputation value (Sun, 2012; Abdel-Hafez et al., 2012; Garcin et al., 2009a). All these reputation models considered that the users give honest ratings, therefore do not take into account the malicious users which give false ratings in order to promote their own product or to damage the reputation of competitor's product. Other rating parameters such as source credibility and decay principle are also ignored. On the other hand, simple summation, arithmetic mean or weighted mean are used as an aggregation method, which have several issues. Moreover, most of these models are single source, hence vulnerable to falsified information and suffer from availability issue.

2.2. Aggregation Methods

Several aggregation methods such as summation (Hu and Liu, 2004; Popescu and Etzioni, 2007), simple mean, weighted mean (Abdel-Hafez et al., 2012), mode and median (Garcin et al., 2009a) are proposed in literature to aggregate user ratings. All these aggregation methods have some issues. Before discussing the issues we first define three different concepts (i.e. robustness, strategy proofness , sensitivity and estimation accuracy) which are used to evaluate the performance of these aggregation methods. The robustness actually measures the resilience of an aggregation method to false and biased ratings. An aggregation method is said to be robust if the false and biased ratings may not easily affect the reputation value. An aggregation method is said to be strategy proof if it does not provide incentive to a reviewer in

order to obtain their preferred reputation value by lying or hiding the actual evaluation. If the reviewers are aware of strategy proofness of an aggregation method then they may rate the product according to actual perception instead of hiding or lying in order to obtain a preferred reputation value. On the other hand, sensitivity measures the ability of an aggregation to reflect the recent ratings about a product quickly. For example, if a product is rated with highest ratings by most reviewers. However, due to some reasons the opinion of reviewers changed and they started to rate the product with lowest ratings. Now sensitivity will measure that how quickly an aggregation method is able to reflect the change in reviewers opinions. Estimation accuracy measures the ability of an aggregation method to provide a good estimation of the actual reputation value.

Summation, mean and weighted mean are not able to estimate a true reputation value when some ratings are false or biased (Garcin et al., 2009a). They are also not robust to false rating because a single false rating is enough to change the reputation value. Moreover, mean and weighted mean are not sensitive to recent ratings and hence allow the manufacture to cheat the system in a smart way. The manufacture can change the quality of the product after gaining a good reputation and will still maintain a good reputation for a long time.

Some authors suggested that median is better aggregation method than mean, weighted mean and mode (Garcin et al., 2009a) because it is strategy-proof as well as robust to false ratings. Mode is also robust but not strategy-proof because sometimes it provides sufficient incentive to change the reputation value substantially (i.e. from 5 to 2.5 or 1 to 2 etc.). For example, if three reviewers rated the product with 5 star and two reviewers rated it with 1 star. Now another reviewer who wants to bring the reputation value to 2/2.5, can do it by hiding the actual perception and rating the product with star 1. Conversely, a strategy proof aggregation method such as median allows to change reputation value but only along the real line (i.e. 1 to 1.5, 3 to 2.5, 4 to 5 etc.) instead of substantial change. The classical mode and median can only aggregate similar kind of information and treat all ratings equally regardless of their importance. Through classical mode and median we cannot aggregate information such as aging factor, rating trustworthiness and source credibility with ratings. In addition, both mode and median need tie breaking rules when we get more than one result. However, an optimal tie breaking rule is difficult to establish, especially when the difference is high. Furthermore, mode and median are not sensitive to recent ratings and hence

do not able to reflect the recent opinion quickly.

2.3. Decay Principle

In e-commerce and peer to peer (P2P) environment the decay principle is implemented in three ways: window based system, exponential decay and using recessive algorithm. The simplest way to implement the decay principle is to select a window of time and compute reputation only based on the ratings posted in that time (Xiong and Liu, 2004; Kinateder et al., 2005). In this method the old ratings are completely ignored, which is suitable in P2P and e-commerce settings because the old ratings are not important to consider. However, in product reputation the old ratings would also be useful and can be considered with less weight. In product context it is also difficult to select an appropriate window only based on time because it is possible to not have enough number of ratings to compute a good reputation value within that duration. Some other parameters should be considered while selecting an optimal window in product reputation context.

The most well know method in both e-commerce and P2P setting is exponential decay. The basic idea of exponential decay is to scale down the past ratings by a constant factor each time a new rating is posted (Jsang and Ismail, 2002; Jøsang et al., 2003; ElSalamouny et al., 2009; Selcuk et al., 2004; Kinateder et al., 2005). Different authors given different names to the same things such as forgetting factors (Jsang and Ismail, 2002), ageing factors (Selcuk et al., 2004) and longevity factor (Jøsang et al., 2003). In exponential decay principle, few recent ratings contribute to the overall reputation and rest have very less weight. For example, let us consider the reputation value computed using 50 ratings, aggregated using weighted mean with an exponential factor of 0.5. The contribution of the 8 most recent ratings represent 99.6% of the reputation value, whereas the contribution of older 42 ratings is only 0.4%. In product context the exponential decay principle is not suitable because of the following reasons. a) The exponential decay is neither robust nor strategy proof because not only few false ratings are needed to change the reputation value but they are also enough to change the reputation value substantially. For example, in exponential decay with factor 0.3 and 0.5 the malicious user need either one or two false ratings with 1 star to change the reputation value from 5 to 1. This is because the contribution of the most recent rating to the reputation value is almost 70 and 50 % for exponential decay with factor of 0.3 and 0.5 respectively, which is too much for a single rating in product reputation context. b) The exponential decay is not able

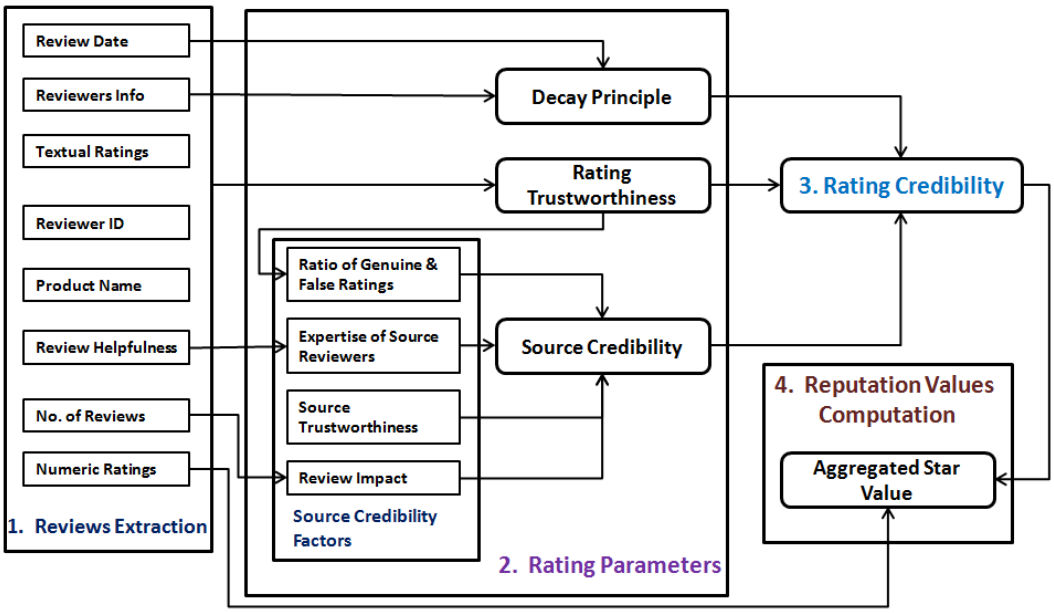
to estimate a good reputation value. As the recent ratings are given more weight which can be false or biased hence the reputation value can be easily misled. Even if the recent ratings are genuine still there is a chance to mislead if some recent raters provide wrong estimation because in product context the ratings are more or less estimation. Therefore, a large sample of ratings is needed to compute a true reputation value. c) In addition, the behavior of the product can be static, and the exponential decay is not suitable for applications where the behavior is static (ElSalamouny et al., 2009). Product associated with static behavior does not change in term of quality and hence customers opinions almost remain the same. In such products, it is more appropriate to compute reputation value based on relative large data sample instead of considering few most recent ratings as in exponential decay.

In (Zacharia and Maes, 2000) a different method is used for discounting the old ratings in e-commerce environment. After each interaction the reputation of the involved parties (i.e. merchants, sellers or buyers) in on-line transaction are re-computed by using the most recent ratings and the previous reputation value. However, this method is also a biased estimation of only few most recent ratings, hence has the same issues as exponential decay. All these authors focused on increasing sensitivity to reflect recent opinions by assigning more weights to recent ratings. However, assigning more weights to recent ratings can also compromise the robustness because the recent ratings can be false or biased. This shows that there is a trade-off between robustness and sensitivity (Liu and Munro, 2012). Therefore, a good reputation model should establish a fair balance between sensitivity and robustness so that to make the model not only resilient to the false ratings but also able to reflect recent opinions. Unfortunately, most of the existing work focused either on sensitivity to recent ratings or robustness to false ratings.

3. The Proposed Model

Based on the need for a specific product reputation evaluation approach, in this section we propose a global architecture and different models answering the research issues highlighted in the state of the art. Figure 1 shows the overall process of the proposed reputation model which is divided into four phases. The arrows represent the direction of information flow. The figure also depicts that how different factors and parameters are related and on the bases of which product evaluation information they are computed. In first

Figure 1: The Overall Process of Proposed Model



phase, reviews and other product evaluation information are extracted from Amazon, Ebay and Cnet Web sources. In second phase several rating parameters such as decay principle, source credibility and rating trustworthiness are computed using reviews and other product evaluation information. In third phase rating parameters are aggregated to determine rating credibility for each rating. In last phase, reputation value is computed by aggregating rating credibility and numeric ratings. The first phase is discussed in Section 3.1, second and third phase are explained in Section 4 and the last phase is presented in Section 5.

3.1. Reviews Extraction and Notation

Huge amount of product reputation data is available on the Web. Several e-commerce and review sites allow users to post reviews about products. In order to compute reputation, reviews about a product are extracted from Web. For this purpose, a wrapper is developed which extracts reviews automatically from *Ebay*, *Cnet* and *Amazon* reviews sites. The wrapper is fully automatic, the user simply need to select a product and the wrapper automatically locates the web pages (containing reviews about that prod-

uct) and extracts reviews from all those web pages. A review r is a tuple of seven things $r(Nr, Tr, rt, t, id, \alpha, \beta)$ and belongs to a set of n reviews $R = \{r_1, r_2, r_3, \cdots, r_n\}$. Nr is the numeric rating given to a product, the numeric rating is in the form of 1 to 5 stars (i.e. $Nr \in \{1, 2, 3, 4, 5\}$) and is a general subjective measure. Tr is textual rating which contains opinions O about a product g as well as about features F of the product, where F is a set of features $F = \{f_1, f_2, f_3 \cdot \cdots \cdot f_q\}$. The rt represents the review title and the t is the date on which review is posted. Similarly, id represents the reviewer id, α and β represent the number of positive and negative helpfulness votes that a review received from other users. Some sources also provide information about reviewers such as age ra, sex rs and location rl etc., which can be used to evaluate product based on demographic information. All these information (i.e. reviews which consist of seven tuples and information about reviewers (age, sex and location)) are extracted by the wrapper. The existing product reputation models considered either only numeric rating or textual rating to compute reputation and ignored other evaluation information. Our model considers all these information in order to compute different rating parameters and reputation value. The web source on which the user post review is represented by s, which belongs to a set of sources $S = \{s_1, s_2, s_3 \cdot \cdots \cdot s_z\}$, where z is the total number of sources. The total number of reviews that a source s has about product g is represented by $m(s, g)$ while n is the total number of reviews that all sources have about product g.

4. Rating Parameters and Rating Credibility

This section first describes different rating parameters and then explains how they are aggregated in order to compute rating credibility. The main characteristic of our reputation model is that it computes several rating parameters and incorporates them into reputation model. These parameters include: (a) Decay Principle, (b) Source Credibility and (c) Rating Trustworthiness. These parameters are derived from existing studies and they are also found useful in product reputation. Decay principle is used both in e-commerce (Jsang and Ismail, 2002) and peer to peer network (Xiong and Liu, 2004) to favor recent ratings. Similarly, source credibility is used in order to determine credible information sources in general survey context (Scott et al., 1984) and in e-commerce network (Cho et al., 2009). On the other hand, rating trustworthiness is used in almost every field where the

users rate a subject, person or an object and some malicious users try to manipulate reputation value (Jindal and Liu, 2008; Lim et al., 2010). A justification is given for each rating parameter in their respective section which answers questions such as why we are using a parameter and which benefits it provides in product reputation.

4.1. Decay Principle

The decay principle is used to favor recent ratings over older ones in order to reflect the recent opinions. The opinions of people about product may change with the passage of time. This change may be caused by several reasons such as rapid change in technology, change in customer needs, market trend and change in the quality of the product. For example, the opinions about a cell phone some days ago may not be the same as today, because several new products with sophisticated features and technologies came to the market. Product reputation models must react to the change in opinions caused by these factors in order to reflect the recent opinions. Based on the issues discussed in the literature review, we propose two decay principle methods, i.e. linear decay and hybrid decay. These methods are more suitable in product reputation context because the reputation value is based on a large sample data and hence more likely to estimate a true reputation value. They are also robust to false and biased ratings because a relative large number of biased ratings are needed to change the result. For example, if we have ten numeric ratings for a product with five star scale, i.e. 2,1,3,3,3,2,3,4,3,2, and computed the aggregated star value (i.e. 2) using the proposed linear decay and aggregation method. Now if malicious users want to increase the aggregated star value from 2 to 3 then they will need four biased ratings with 5 stars. Conversely, the exponential decay with factor 0.3, 0.5 and 0.9 will need 1, 1 and 4 false and biased ratings respectively even to bring the aggregated star value to 5. In addition, these methods also allow the users to either increase or decrease the sensitivity or robustness in order to produce a balance according to the nature of the product and user requirements.

4.1.1. Linear Decay

The main idea of linear decay is to scale down the past ratings by repeatedly subtracting the same value. The linear decay principle distributes the weights fairly and the reputation is not based only on few recent ratings. For example, if we have forty ratings and computed the weights associated

with each rating based on time using linear decay, then the contribution of recent twenty ratings to the reputation value is 75% and the contribution of remaining twenty ratings is 25%. Similarly, if we divide this window of forty ratings into four parts (most recent, recent, old and oldest) each of which includes ten ratings, then the contribution of most recent ten ratings is 44%, recent ten is 31%, old ten is 19% and oldest ten is 6%. This provides a good estimation of reputation because it is based on a relatively large sample of ratings, and it is more suitable distribution in product context because the old ratings also have some contribution in reputation value.

In order to implement the linear decay principle, weights associated with different times on which ratings about a product are posted are computed using Equation 1. The weights are based on the time t on which rating is posted.

$$w_i = w_1 + (i-1)d_w \tag{1}$$

Where w_i is the weight associated with time i, $w_i \in W = \{w_1, w_2, w_3 \cdots\cdots w_n\}$, w_1 is weight assigned to the most recent rating which is equal to 1, w_n is the weight assigned to oldest rating which is computed using Equation 1, d_w is the difference between two consecutive weights which is determined by using Equation 2 only when $i >= 2$. Where l is the total number of unique times on which ratings are posted.

$$d_w = (-w_1)/l \tag{2}$$

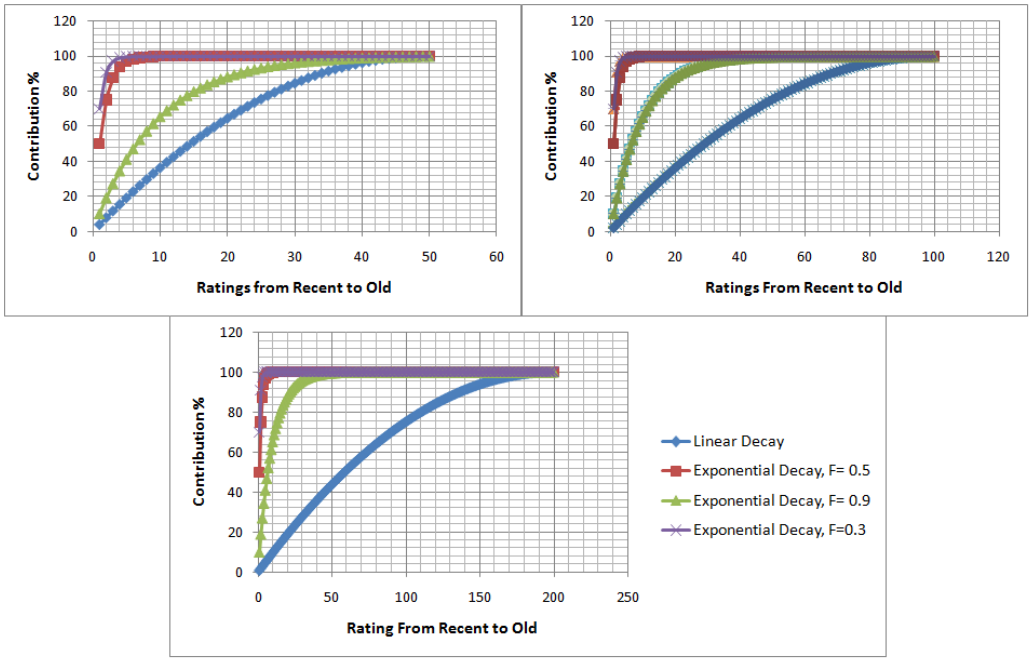
The weights computed in Equation 1 are not normalized because, $\sum_{i=1}^{n} w_i \neq 1$, therefore the Equation 3 is used to make the weights normalized, where WT_i is the normalized weight for rating posted on time i, such that $\sum_{i=1}^{n} WT_i = 1$.

$$WT_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \tag{3}$$

For product with static behaviour, the decay principle is kept constant by using value one ($WT_i = 1$) , which is equivalent to not having an aging factor at all.

Figure 2 shows a comparison between exponential decay and the proposed linear decay. The rating contribution to the reputation value of recent and old ratings is shown through a graph. The x-axis represents the number of ratings

Figure 2: Comparison of Linear Decay and Exponential Decay



from recent to old and the y-axis represents the contribution of these ratings to the reputation value in percentage. The first graph is for fifty ratings and the second and third is for hundred and two hundreds ratings respectively. The graphs show that in all three cases the contribution to the reputation value of few recent ratings in exponential decay (with different exponential factors) is very high and the rest have very low. In addition, the contribution of the most recent rating in exponential decay is almost 70% with exponential factor F=0.3, and almost 50% and 10% with F=0.5 and F=0.9 respectively, which is very high for a single rating in product reputation context. If the most recent rating is false or biased then the reputation value can easily be misled. Conversely, in linear decay the contribution of recent and old ratings is appropriate because in all cases the contribution of recent half part of overall ratings is 75% and the remaining half part is 25%. This shows that the reputation value in linear decay is based on a large sample data and hence more likely to estimate a true reputation value.

4.1.2. Hybrid Decay

The hybrid decay is the combination of window based system and linear decay. In this method a window of recent ratings is selected based on some parameters and linear decay is used to give weights to the ratings inside that window. In product reputation context, an appropriate window cannot be selected only based on time because we may not have sufficient number of ratings to compute a true reputation value. Therefore, in hybrid decay we propose Algorithm 1 which uses three parameters simultaneously which are provided by users in order to select an appropriate window. In order to reflect recent opinions within this window liner decay is used to assign weights. The parameters used to select a window include time span, sufficient number of ratings and rating validity. In this research work, the values for these parameters for each product are taken from users (who want to compute reputation value for a product) through a user interface. However, future work is needed to automatically decide these parameters based on user local context, nature of product, product behavior, ratings behavior, variations in ratings and total number of ratings. The users can select a time span by entering a range of dates (i.e. from and to, e.g. 1/05/2016 to 1/06/2016). The sufficient number of ratings actually represents the size of ratings window that user want to consider to compute the reputation value, which can be 50, 100 or 150 ratings etc. Similarly, as the validity of ratings is based on reviewer age, sex and location, therefore the user can select a specific gender (i.e. male or female), age rang in years (e.g. from 15 to 35 years) and location by choosing a specific country. The algorithm 1 selects the ratings window according to user preferences by comparing the user entered values with information (i.e. review date, reviewers age, sex and location) extracted from reviews sites.

The hybrid decay is only used when the user is interested in one of the following options. a) The user wants to increase sensitivity in order to reflect the recent opinions about product more quickly (e.g. in services). b) When there is large number of ratings and the user want to make selection based on some criteria. b) When the user want to evaluate product based on other criteria such as reviewer age, sex or location.

Time Span of Ratings:. In product reputation context we are not able to select an appropriate window only based on time span because of irregularity in ratings posting. For a product you will find a large number of ratings in one month and few or even no ratings in the next month. We are not able to

guess window that contains sufficient number of ratings based on time span only. Therefore, we also need to consider the sufficient number of ratings as well as rating validity while deciding the window. However, selecting a window based on time is very important for historical study of the product. Several rating windows can be selected in order to know the opinions of product in different time durations, and how and when the opinion changed. These windows can also be used to determine trend in the market and change in customer preferences.

Sufficient Number of Rating:. In product evaluation context the ratings provided by users are more or less an estimation, so while selecting a window of recent ratings we need to make sure that the window contains sufficient number of ratings to compute a good reputation value. The sufficient number of ratings varies from product to product because it depends on the nature of product, product behavior and total number of ratings.

Ratings Validity:. While selecting appropriate window we also need to identify the rating validity in order to build a window of only valid ratings. Rating validity can be based on the information about reviewers, such as sex, age and location. However, only some sources such as Amazon make these information public, which can be used to determine rating validity. For example, a young boy will only be keen to know that how young people rated the product. On the other hand, the people living in a cold area will be interested to know the opinions of only those people who are living in the same sort of environment because the performance of a product may not be the same in different climates. Some products even do not have valid ratings, in such case the reputation model must need to identify this situation so that reputation value cannot be misled.

Description of the Proposed Algorithm for Selecting Rating Window:. The Algorithm 1 is proposed to select the appropriate window ϖ of recent ratings. The inputs to the algorithm are: numeric ratings NR, corresponding time T on which these ratings are posted, value of sufficient number of rating δ, set of valid V and invalid $\bar{V}$ ratings (which are build by using user entered criteria) and time span $t_1 - t_k$ of ratings which is appropriate for user to consider. The first line initializes the time span $t_1 - t_i$ to $t_1 - t_k$ so that the initial window $\varpi_{t_1-t_i}$ can be selected based on appropriate time span $t_1 - t_k$ (Line 2), where i and $k \leqslant n$. The window $\varpi_{t_1-t_i}$ is checked to determine that whether it contains sufficient number of ratings or not (Line-3). If the

Algorithm 1 Algorithm for Selecting Rating Window

Inputs:$NR = \{Nr_1, Nr_2, Nr_3 \ldots\ldots Nr_n\}$, $T = \{t_1, t_2, t_3 \ldots\ldots t_n\}$, δ, V, $\bar{V}$, $t_1 - t_k$

Outputs: ϖ

1: $t_1 - t_i = t_1 - t_k$.
2: $\varpi_{t_1-t_i} = \{Nr_1, Nr_2, Nr_3 \ldots\ldots Nr_i\}$
3: IF $(|\varpi_{t_1-t_i}| \geq \delta)$ Then
4: Go to Step 7
5: Else $i = i + 1$ Until $(\delta \subseteq \varpi_{t_1-t_i}) \vee (i = n)$
6: End IF
7: IF $(\varpi_{t_1-t_i} \subseteq V)$ Then
8: Go to Step 13
9: Else IF $(\varpi_{t_1-t_i} \subseteq \bar{V})$ Then
10: $\varpi_{t_1-t_i} = \varphi$ and go to Step 14
11: Else $i = i - 1$ Until $(\varpi_{t_1-t_i} \subseteq V)$
12: End IF
13: $\varpi = \varpi_{t_1-t_i}$
14: End

window does not contain sufficient number of ratings then it is extended until either to obtain the sufficient number of ratings or all the ratings are included (L-5). The validity of window is also checked in order to make sure that it contains only valid ratings (L-7). If all the ratings in the window are invalid then an empty window is return (L-9-10). On the other hand, if the window contains some invalid ratings then they are excluded from window (L-11). The computed window $\varpi_{t_1-t_i}$ is assigned to appropriate window ϖ and this window is returned (L-13).

4.2. Source Credibility

Our reputation model considers the evaluation information obtained from multiple sources. The most important advantage of a multi-source reputation model is that it has the ability to resist the attempt of falsifying evaluation information. In addition, multiple sources also increase reliability (Sabater and Sierra, 2002) as well as the feasibility of computed reputation value. Sometimes, a single source lacks ratings and is not enough to compute a reputation value (Josang et al., 2007), in this case multi-source reputation model is the only solution. Furthermore, this also addresses the issue of

availability of single source based reputation model because at least some sources will be available to compute reputation value. In a multi-source reputation model it is important to consider source credibility because there are some malicious sources which have their own incentive as a merchant, seller or intermediate and may promote certain products. Indeed, the result of a credible source is more reliable and trustworthy (Wiener and Mowen, 1986).

Source credibility measures the degree to which the evaluation information on a source is credible to be considered for reputation. Here, by source we mean the web sources from which evaluation information is extracted and we are concerned with credibility of user generated evaluation information on a source instead of original contents. In literature, the source credibility refers to the credibility of the evaluator/reviewer who posted the rating. According to source credibility theory, the credibility of a source depends on the expertise, trustworthiness, co-orientation and attraction (Hawkins et al., 2004; Scott et al., 1984). Expertise and trustworthiness are two different concepts. Expertise is the degree to which an information source is capable to provide valuable and correct information. However, trustworthiness is the extent to which a source provides information that reflects the source actual feelings or opinions. Co-orientation is the degree to which a source is similar to a target group, such as having similar problems or other characteristics related to the use of a product. Attraction is the degree to which an information source elicits positive feeling from audience members, such as to emulate the source in some ways. The attraction is not appropriate in online reputation because the information source is not as much revealed as source in a physical survey. The first three factors are generally considered in order to measure the credibility of an evaluator (Cho et al., 2009) in e-commerce. However, we are concerned with the web source credibility instead of the evaluator credibility. We have identified four factors that contribute to the source credibility: source trustworthiness, cumulative expertise of source reviewers, ratio of trustworthy and untrustworthy ratings and review impact. The first three factors are derived from source credibility theory (Hawkins et al., 2004; Scott et al., 1984). The last factor which is review impact is based on number of reviews a source has is used because the source with few reviews about a product are more likely to be spam (Jindal and Liu, 2008). Furthermore, a justification for each factor has been provided in their respective section which answers the question such as why we are using a factor. The user can use some are all of these factors in order to compute source

credibility. Equation 4 is proposed in order to determine source credibility $SC(s, g)$ of source s for product g.

$$SC(s,g) = \frac{ST(s) \times ES(s,g) \times G(s,g) \times RI(s,g)}{\sum_{i=1}^{z} (ST(i) \times ES(i,g) \times G(i,g) \times RI(i,g))} \tag{4}$$

Where $ST(s)$ is source trustworthiness of source s, $ES(s, g)$ is the cumulative expertise of source s about product g, $G(s, g)$ is the ratio of trustworthy and untrustworthy ratings of source s about product g, $RI(s, g)$ is the review impact of source s about product g and z is the total number of sources. Simple product is used in Equation 4 in order to aggregate different factors and this formulation is adopted because it offers the following benefits. (1) A normalized value between 0 and 1 for $SC(s, g)$ is obtained. (2) It also allows to obtain a higher value for source which is consistent across all factors. (3) This equation also allows to make some factor constant in its absence or when the user don't want to consider. For example, if a manufacturer do not want to consider the $ES(s, g)$ while computing source credibility then the value of ES for all sources is set to 1 (which is equivalent to not having this factor at all). (4) Except $ES(s, g)$, other three factors are necessary to consider, hence we want that when $ST(s)$, $G(s, g)$ or $RI(s, g)$ is zero for a source then we must have value 0 of $SC(s, g)$ for that source, which is offered only by this equation.

4.2.1. Source Trustworthiness

Source trustworthiness measures the degree to which evaluation information available on a source can be trusted. We consider three qualitative factors to determine the source trustworthiness: impartialness of source, adopted security measure and rating posting policies and regulations. All these factors are subjective and were ranked between 0 and 1 by three experts who analyzed the web sources. The policies and the regulations of posting ratings on a source which encourage or discourage biased behavior are measured by investigating that how does a source allow a user to post rating. Some relative grading guidelines were used, which are list down in the sequence from lower to higher grading. a) Some sources allow a single user with the same identity to post multiple reviews. b) Some sources allow all their visitors to post reviews. c) Others only allow their registered users or the users who bought the product to post reviews. The adopted security measures by a source to protect the system from malicious users is also

rated by the experts while considering some criteria (i.e. data privacy, data modification, consumer information protection, encryption, authentication and third party verification). The impartialness of a source was examined by the experts while considering two criteria, i.e. whether the source has some incentives as a seller or intermediate and whether they are favoring that product to keep the reputation up or not. The intra-class correlation coefficient (ICC) (Koch, 1982) was used to measure the level of agreement between experts. The statistic application SPSS (MacLennan, 1993) with two-way mixed model was used to compute level of agreement which was 77 %, interpreted as substantial.

After ranking these parameters for all source by experts the Equation 5 (which aggregated all these factors using product) is used to determine source trustworthiness $ST(s)$. In order to obtain a normalized value the product of these factors for a source is divided by the summation of product of these factors for all sources.

$$ST(s) = \frac{I(s) \times P(s) \times SM(s)}{\sum_{i=1}^{z} (I_i \times P_i \times SM_i)} \tag{5}$$

Where $I(s)$ represents impartialness of source s, $P(s)$ is the policies of posting ratings and $SM(s)$ security measures adopted by source s respectively. The source trustworthiness is difficult to measure because it is based on subjective measures, however it is computed only once for each source. For example, if we have five sources then the source trustworthiness need to be computed five times, i.e. once for each source. Conversely, the other source credibility factors are computed for every product.

4.2.2. Cumulative Expertise of Source Reviewers

Another factor that contributes to the credibility of a source is the cumulative expertise of source reviewers which actually measures the collective expertise level of reviewers who posted ratings on a web source. The expertise of a individual reviewer measures the extent to which he is capable to provide valuable information. The higher the expertise of a reviewer, the higher will be the valuableness of the evaluation he would provide (Cho et al., 2009). Similarly, a web source with more expert reviewers is more persuasive than a source with less expertise. In literature, several methods are proposed in order to determine review helpfulness, some of which also make use of review helpfulness votes posted by users (Korfiatis et al., 2012; Connors

et al., 2011). Review helpfulness and reviewer expertise are slightly similar concepts because the helpful reviews are written by expert users (Connors et al., 2011). Some other factors, such as writing style, details and pros and cons also contributes to helpful review. Therefore, the expertise of a reviewer can be measured by using the helpfulness vote that his review received. Similarly, the cumulative expertise of source reviewers is measured by using the cumulative ratio of positive and negative helpfulness votes of all reviews of a web source. In addition, the number of reviews which received these helpfulness votes is also considered and hence the cumulative ratio of positive and negative helpfulness votes is divided by the number of reviews that a source has. Based on this, Equation 6 is proposed which ranks the cumulative expertise $ES(s, g)$ of a source s for a product g between 0 and 1.

$$ES(s,g) = \frac{\sum_{i=1}^{m(s,g)} (\alpha_i / (\alpha_i + \beta_i))}{m(s,g)} \div \sum_{k=1}^{z} (\frac{\sum_{i=1}^{m(s,g)} (\alpha_i / (\alpha_i + \beta_i))}{m(s,g)})_k \tag{6}$$

Where α and β are the number of positive and negative helpfulness votes respectively received by a review, $m(s, g)$ is the total number of reviews on source s about the same product g.

4.2.3. Ratio of Trustworthy and Untrustworthy Ratings

A source is more credible if it has either no or very few number of untrustworthy ratings or spam reviews. Some methods are proposed in literature in order to determine rating trustworthiness as discussed in Subsection 4.4. However, some untrustworthy ratings have similar characteristics like trustworthy ratings and may bypass the spam detection method. The ratio of this type of untrustworthy ratings is likely to be higher if the ratio of identified untrustworthy ratings is high. Therefore, in order to tackle this situation the source with higher ratio of trustworthy ratings is given more weight over sources with a less ratio of trustworthy ratings. We propose Equation 7 in order to compute a normalized ratio of trustworthy and untrustworthy ratings. For this purpose, the ratio of trustworthy overall ratings for a source s is divided by the summation of this ratio for all sources.

$$G(s,g) = \frac{r(s,g)}{r(s,g) + \bar{r}(s,g)} \div \sum_{i=1}^{z} (\frac{r(i,g)}{r(i,g) + \bar{r}(i,g)}) \tag{7}$$

Where $G(s,g)$ is the ratio of trustworthy ratings, and $r(s,g)$ and $\bar{r}(s,g)$ are the total number of trustworthy and untrustworthy ratings respectively on source s about product g. The $r(s,g)$ and $\bar{r}(s,g)$ are determined using the results produced by Section 4.4 (i.e. rating trustworthiness of all ratings/reviews).

4.2.4. Review Impact

Review impact is based on the number of reviews that a source has about a product. Higher the number of reviews for a product a source has, higher is the review impact of that source. Similarly, we know that a reputation value based on large sample data is more reliable and trustworthy. A threshold value of review impact can also be used to ignore the sources which have very few number of reviews about product because a source with only few number of reviews about a product are more likely to be spam (Jindal and Liu, 2008). We proposed Equation 8 to compute review impact $RI(s,g)$ for a source s about product g.

$$RI(s,g) = \frac{m(s,g)}{m(\max,g)} \div \sum_{i=1}^{z} \left(\frac{m(i,g)}{m(\max,g)}\right) \tag{8}$$

Where $m(s,g)$ is the number of reviews the source s has about product g and $m(max,g)$ is the number of reviews of the source which have maximum number of reviews.

4.3. Rating Trustworthiness

Rating trustworthiness determines the extent to which the reviewer provides evaluation that reflects its own feelings or opinions. In other words we can say that rating trustworthiness determines that whether the review is fake or genuine. Several words for ratings trustworthiness are used in literature such as spam detection, fake reviews detection or opinions fraud detection (Jindal and Liu, 2008; Lim et al., 2010). We implemented the method proposed in (Jindal and Liu, 2008) in order to detect spam reviews. There are three types of spam reviews on e-commence and review sites: untruthful reviews (type 1), reviews on brand only (type 2) and non reviews (type 3). Untruthful reviews are those reviews which are deliberately posted either to promote a product or to damage the reputation of a product. Reviews on brand are those where brand, manufacturer or seller are discussed instead of expressing opinions about product. Non reviews are those which contain

irrelevant text such as advertisement, questions and answers or random text. Most of the untruthful reviews are duplicate. In order to detect duplicate or near duplicate reviews, content comparison of textual ratings is performed. Reviews that have similarity at least 80 % are considered duplicate. On the other hand, a model is built based on the features (which differentiate spam reviews from genuine reviews) in order to detect the spam reviews of type 2, type 3 and type 1 which are not duplicate. Three main categories of features are defined in (Jindal and Liu, 2008) (i.e. review centric features, reviewer centric features and product centric features). We are using those features which can be define on the bases of information extracted from e-commerce and review sites (i.e. textual rating, reviewer id, reviewer name, review helpfulness, numeric rating given by reviewer, aggregated value, brand name, product features etc.) without any manual labeling. The model is built with logistic regression which produces a probability estimation which reflects the likelihood that the review is spam. The statistical package R (Ripley, 2001) is used for logistic regression. The result of logistic regression (which represents the probability that a review is spam) is weighted down in order to determine rating trustworthiness RT. On the other hand, the rating trustworthiness for each duplicate review is assign value zero because the intent is to manipulate reputation valve.

4.4. Rating Credibility

Rating credibility measures the degree to which the rating is credible to be considered for reputation. Rating credibility is a broader concept than source credibility as it considers parameters such as rating trustworthiness, source credibility and decay principle (i.e. time on which rating is posted, sufficient number of ratings, reviewer age, reviewer sex, reviewer location). This measure ranks the credibility of each rating within the range of 0 and 1. We proposed Equation 9 which aggregates these parameters in order to compute rating credibility RC for rating x. Simple product is used here again and this formulation also offers the same advantages as listed for Equation 4. In order to obtain a normalized value for each rating the product of these parameters is divided by the summation of the product (of same parameters) for all ratings.

$$RC(x) = \frac{TW(x) \times RT(x) \times SC(x)}{\sum_{i=1}^{n} (TW_i \times RT_i \times SC_i)} \tag{9}$$

Where $TW(x)$ is the time weight assigned to ratings x using decay principle, $RT(x)$ is the rating trustworthiness and $SC(x)$ is the source credibility of the web source from which rating x is extracted.

5. Reputation Values Computation

This section explains how the rating parameters and/or rating credibility can be used in order to compute reputation values. Several reputation values can be computed with rating parameters and rating credibility. However, due to space limit in this paper we propose aggregation method only for aggregated star value.

5.1. Aggregated Star Value

Aggregated star Value ranks the product into one to five stars. This value is obtained by aggregating numeric ratings. We have tailored weighted median so that it can be used for aggregating numeric ratings. Weighted median offers several benefits over existing aggregation methods.

In order to use weighted median we must have data in a specific form, i.e. set of numbers $\{x_1, x_2, x_3 \cdots\cdots x_n\}$ with each number having weight $\{w_1, w_2, w_3 \cdots\cdots w_n\}$. In product reputation context, we have set of numbers which is five stars (i.e. $\{1, 2, 3, 4, 5\}$) , however we don't have weights associated with each star. Therefore, the concept of star weight SW is introduced, which actually represents the weights associated with each star. We propose the Equation 10 to compute weights for all five stars, i.e. five star weight SW_5, four star weight SW_4 , three star weight SW_3 , two star weight SW_2 and one star weight SW_1. The concept of star weight associated with each star while computing weighted median is similar to the weights and frequencies of different numbers we have while computing weighted mean and mode respectively. However, the star weight associate with a star is not only based on the occurrences (i.e. frequency) of that star in numeric ratings but also on rating credibility associated with each occurrence. In short, the star weight actually represents the important of each star to be selected as an aggregate star value and it is based on occurrences and rating credibility. If a star has more occurrences associated with higher credibility then the star weight (for that star) will be high and subsequently that star is more likely to be selected as an aggregated star value.

$$SW_v = \sum_{i=1}^{k(v)} \left(\frac{RC(i) \times Nr(i,v)}{v}\right) \div \sum_{v=1}^{5} \left(\sum_{i=1}^{k(v)} \left(\frac{RC(i) \times Nr(i,v)}{v}\right)\right) \tag{10}$$

Where $SW(v)$ is the star weight, associated with star v, where $v \in \{1,2,3,4,5\}$, $Nr_(i,v)$ represents the occurrence i of star v, RC_i is the rating credibility associated with occurrence i of star v computed using Equation 9 and $k(v)$ is the total number of ratings with star v. The SW are normalized star weights, such that $\sum_{v=1}^{5} SW_v = 1$. When a user want to compute reputation value only based on specific rating parameter (such as rating trustworthiness, decay principle or source credibility) then instead of using RC_i a specific rating parameter can be used in Equation 10.

The stars $V = \{1,2,3,4,5\}$ associated with the corresponding star weights $SW = \{SW_1, SW_2, SW_3, SW_4, SW_5\}$ are aggregated using weighted median. The following steps are followed in order to compute aggregated star value ASr.

1. Sort the stars in descending order, Let $v_5 \geq v_4 \geq v_3 \geq v_2 \geq v_1$
2. Let the corresponding normalized star weights are $SW_5, SW_4, SW_3, SW_2, SW_1$
3. Find $SW_{j\psi}$ such that the Equation 11 and 12 hold, where l is the total number of stars (i.e five).

$$SW_{j\Psi} + \sum_{j=1}^{j\Psi} SW_j \geq \sum_{j=j\Psi+1}^{l} SW_j \tag{11}$$

$$\sum_{j=1}^{j\Psi-1} SW_j \leq \sum_{j=j\Psi}^{l} SW_j + SW_{j\Psi} \tag{12}$$

4. Then, the star $v_{j\psi}$ corresponding to $SW_{j\Psi}$ is the weighted median and the aggregated star value ASr.

6. Results and Discussion

In order to validate the results and to evaluate the performance of our reputation model we have used different criteria such as sensitivity, robustness, estimation accuracy and strategy proofness. We also compared the results of our model with existing models and aggregation methods using these criteria.

6.1. Sensitivity

Sensitivity measures the extent to which a reputation model or aggregation method reflects the recent ratings. The reputation model must not be too much sensitive so that malicious users cannot change the reputation value easily. On the other hand, the reputation model must also not be too much robust so that it can reflect the recent opinions quickly. There is a trade-off between robustness and sensitivity, so a balance between robustness and sensitivity must be maintained (Liu and Munro, 2012).

6.1.1. Sensitivity of Aggregated Star Value Method

In order to measure the sensitivity of aggregation methods for numeric ratings, we have extracted numeric ratings of twenty products from Amazon, Ebay and Cnet sites on September, 2015. These products belong to four different product categories, i.e. Cell Phones, Cameras, Computers and Appliances, represented by PC-1, PC-2, PC-3 and PC-4. Products with different number of ratings n are selected from each category as shown in Table 1. Ten of these products already maintained good reputation value (i.e. 5 or 4 aggregated star value) and same number of products already maintained bad reputation value (i.e. 2 or 1). We suppose that due to change in quality, behavior, technology or market, the opinions of customers changed and the customers started to rate the products with good reputation with the lowest rating (i.e. 1) and products with bad reputation with highest rating (i.e. 5). The sensitivity of different aggregation methods is measured by counting the number of lowest ratings (for product with good reputation) and highest ratings (for product with bad reputation) needed to fully reflect the recent opinions. For example, if a product has aggregated four or five stars rating, however with change in the quality the customers started to rate it with worst ratings (i.e. 1), then sensitivity is measured by counting the number of one star ratings needed to change the aggregated star value from five/four to one. One issue with mean and weighted mean is that it is not possible to bring the aggregated star value exactly to one when a product already maintained a good reputation. Therefore, for mean and weighted mean we count the number of 1 star ratings needed to bring the aggregated star to less than 1.5. The weights computed by the linear decay are also used with weighted mean.

The results of proposed method (PM) is compared with other aggregation methods in Table 1. The PM here refers to the whole process of computing rating credibility and then the reputation value (i.e. aggregated star value)

Table 1: Sensitivity of Aggregation Methods

Product Category	Product	n		Aggregation Methods				
				Mean	W. Mean	Mode	Median	Proposed Method
PC-1	P-1	10	No. of 1/5 Star Ratings	72	18	9	11	4
	P-2	20		42	24	11	18	8
	P-3	30		111	40	3	25	11
	P-4	40		246	67	31	38	17
	P-5	50		190	60	7	33	16
PC-2	P-6	60		349	95	39	53	22
	P-7	70		370	160	55	64	27
	P-8	80		490	170	58	72	30
	P-9	90		510	199	61	84	36
	P-10	100		550	155	60	90	35
PC-3	P-11	110		720	179	89	112	46
	P-12	120		570	159	56	102	40
	P-13	130		765	173	86	111	48
	P-14	140		817	175	86	137	56
	P-15	150		967	170	124	148	61
PC-4	P-16	160		849	154	75	157	67
	P-17	170		1010	153	115	128	72
	P-18	180		1043	149	117	161	73
	P-19	190		984	141	94	172	73
	P-20	200		1125	136	121	177	75
Average				589	128.8	64.8	94.6	40.8
n= Total Number of Ratings, PC=Product Category and P-1=Product-First								

using the proposed aggregation method. The result shows that our proposed method is very quick to reflect the recent ratings because on average only 40.8 ratings are needed to bring the aggregated star value either to one or five. Conversely, means and weighted means are not able to reflect the recent ratings quickly because very large number of ratings, i.e. 589 and 128.8 respectively are needed to bring the aggregated start ratings even closer to one/five star. Similarly, mode and median needed 64.8 and 94.6 ratings respectively, which is also relative large. In most products the number of 1/5 star ratings needed for mode and median is even very close to the total number of ratings, which is too much to reflect the recent opinions. For example in P-4, where the total numbers of ratings are 40, we need 31 and 38 one star ratings for mode and median respectively. The same results are observed in the case of P-7, P-8 and P-15. The results also shows that the number of one stars needed is directly proportion to the total number of ratings, however it also depends on the aggregated star value as well as the

stars distribution.

6.1.2. Sensitivity of Decay Principle

Another experiment is performed using the same experimental settings and method (with different data set) used in Section 6.1.1 in order to determine the sensitivity of different decay principles. The numeric ratings of ten products which already maintained good reputation are extracted from reviewer sites for this experiment. The proposed linear decay and hybrid decay principles are compared with exponential decay (Jsang and Ismail, 2002; Jøsang et al., 2003; ElSalamouny et al., 2009; Selcuk et al., 2004; Kinateder et al., 2005) using different exponential factors (i.e. 0.3, 0.5, 0.9).

Table 2: Sensitivity of Decay Principles

Aggregation Methods	PC-1				PC-2				PC-3				A.A
	NS			ANS	NS			ANS	NS			ANS	
Product	P-1	P-2	P-3		P-4	P-5	P-6		P-7	P-8	P-9		
n	20	65	112		50	136	40		150				
LD	8	22	47	25.6	14	16	51	27	17	36	61	38	30.2
HD	8	13	15	12	14	13	15	14	15	15	15	15	13.6
ED, F=0.9	5	6	7	6	7	6	7	6.6	7	7	7	7	6.5
ED, F=0.5	1	1	2	1.3	2	1	2	1.6	2	2	2	2	1.6
ED, F=0.3	1	1	1	1	1	1	1	1	1	1	1	1	1
PC=Product Category, NS=No. of 1 Stars, ANS=Average No. of 1 Stars , P-1=Product-1, A.A=Accumulative Average, n=Number of Ratings, LD= Linear Decay, ED=Exponential Decay and F=Factor													

Table 2 shows the sensitivity results of different decay principles. The results show that exponential decay with factor 0.3 only needed one false rating to bring the reputation value from highest to lowest. Similarly, the exponential decay with factor 0.5 and 0.9, on average require two and seven ratings respectively to reflect recent opinion fully. This shows that exponential decay is too much sensitive to false ratings and provides sufficient incentives to malicious users to change the reputation value even from highest to lowest and vice versa. From this, we can conclude that exponential decay is not suitable when we have 1-5 stars scale and some ratings can be false as in product reputation. On the other hand, the proposed linear decay needed thirty ratings to fully reflect the recent opinions. In other words, six ratings are needed to change the aggregated value by one rank (e.g. from 4 to 3), which is more suitable in product reputation context. Furthermore, the results also show that the hybrid decay can be used to increase the sensitivity

of the model because only fourteen ratings, which is almost half of the linear decay are needed to fully reflect the recent ratings.

6.2. Robustness

A reputation model or aggregation method is said to be robust if the false ratings may not easily affect the overall reputation value (Garcin et al., 2009a). Higher the number of false ratings needed to change the aggregation result, more robust the aggregation method will be. In order to measure the robustness of the aggregation methods we counted the number of biased ratings required to change the aggregation result. For this purpose, both highest possible ratings (i.e. 5 star) and lowest possible ratings (i.e. 1 star) are inserted until the result of aggregation method change. Ratings of twelve products are extracted from Amazon, Ebay and Cnet review sites on July, 2014. The products with varying number of ratings (as shown in Table 3), different aggregated star value and stars distribution are selected in order to validate the results on different scenarios.

Table 3: Robustness of Aggregation methods

Aggregation Method		Mean		Weighted Mean		Mode		Median		Proposed Method	
Product	n	HR	LR	HR	LR	HR	LR	HR	LR	HR	LR
P-1	10	1	1	1	1	8	8	8	7	4	3
P-2	15	1	1	1	1	10	10	18	4	8	3
P-3	20	1	1	1	1	6	5	3	7	2	3
P-4	25	1	1	1	1	17	18	23	13	10	4
P-5	30	1	2	1	1	14	6	29	5	14	2
P-6	44	1	3	1	1	3	6	4	20	8	8
P-7	50	2	2	1	1	1	3	10	12	7	7
P-8	60	1	3	1	2	1	3	14	12	6	9
P-9	70	3	2	1	2	16	18	47	12	24	4
P-10	80	1	1	1	2	18	25	14	49	5	23
P-11	100	3	4	1	3	30	27	10	52	5	33
P-12	120	6	1	1	3	14	20	24	40	14	16
Avg.		1.8	1.8	1	1.5	11.5	12.4	17	19.4	8.9	9.5
C.Avg		1.8		1.2		11.9		18.2		9.25	
HR=No. of Highest Ratings, LR=No. of Lowest Ratings, Avg.=Average, C.Avg=Cumulative Average, n=No. of Ratings											

Table 3 shows the number of false ratings (i.e. number of lowest ratings LR and number of highest ratings HR) required to change the aggregation results. While discussing the results of robustness in Table 3 we also need to keep in mind the result of sensitivity in Table 1 because there is a trade-off between sensitivity and robustness. The results in Table 3 show that mean and weighted mean are not robust to false ratings because only one biased rating is enough to change the result. Similarly, results in Table 1 shows that both mean and weighted mean are not able to reflect the recent opinions quickly. On the other hand, the median and mode are robust because it needed 18.2 and 11.9 ratings on average respectively to change the results (Table 3), however they are not able to reflect the recent ratings quickly (Table 1). The result of our proposed method is optimal because it is fairly robust (as it needed 9.2 ratings on average to change the reputation value (Table 3)) as well as able to reflect the recent opinions quickly (Table 1). In other words, we can say that proposed method is not too much robust, therefore able to reflect the recent opinions quickly. Simultaneously, it is not too much sensitive, hence cannot be compromised by false ratings easily. Our method also allows the users to increase the robustness by keeping the decay principle constant (i.e. $WT_i = 1$). In this case, the robustness will be increased and will be the same as the median. However, keeping the decay principle constant is suitable only when the behaviour of the product is static.

6.3. Estimation Accuracy

The estimation accuracy measures the degree to which a reputation model or aggregation method accurately estimates a true reputation value. The Absolute Error AE is used to compute the accuracy, which consists of computing the absolute value of the difference between the actual and the predicted reputation values (Massa and Avesani, 2007). The average accuracy is computed by Mean Absolute Error MAE, which determines the average of all Absolute errors. Lower the value AE and/or MAE for a reputation model, higher the accuracy. AE and MAE are computed for different reputation models and aggregation methods using the Equation 13 and 14 respectively.

$$AE = |R_a - R_p| \tag{13}$$

$$MAE = \frac{\sum_{i=1}^{o} AE_i}{o} \tag{14}$$

Where R_a is the actual and R_p is the predicted reputation values, o is the total number of products for which AE is computed. The predicted reputation value is computed by using our proposed method. On the other hand, the actual reputation value is computed by using the field experts whose are provided with ratings and associated actual rating parameters and rating credibility.

Table 4: Scenario for Computing Accuracy

Product	Source_1			Source_2			Source_3			n
	$m(s_1, g)$	$r(s_1, g)$	$\bar{r}(s_1, g)$	$m(s_2, g)$	$r(s_2, g)$	$\bar{r}(s_2, g)$	$m(s_3, g)$	$r(s_3, g)$	$\bar{r}(s_3, g)$	
Product_1	19	17	2	17	14	3	14	8	6	50
Product_2	25	22	3	20	15	5	15	8	7	60
Product_3	28	25	3	23	16	7	19	9	10	70
Product_4	38	33	5	34	24	10	28	13	15	100

Table 5: Source Credibility Computation

Product	Source	ST	ES	G	RI	SC
Product_1	$s-1$	0.5955	0.3706	0.3907	0.38	0.7196
	$s-2$	0.2205	0.3314	0.3596	0.34	0.1962
	$s-3$	0.1838	0.2979	0.2495	0.28	0.084
Product_2	$s-1$	0.5955	0.3201	0.4067	0.4166	0.708
	$s-2$	0.2205	0.3972	0.3466	0.3333	0.2218
	$s-3$	0.1838	0.2825	0.2465	0.25	0.0701
Product_3	$s-1$	0.5955	0.2727	0.4329	0.4	0.6841
	$s-2$	0.2205	0.3582	0.3373	0.3285	0.213
	$s-3$	0.1838	0.3689	0.2296	0.2714	0.1028
Product_4	$s-1$	0.5955	0.3058	0.4259	0.38	0.7
	$s-2$	0.2205	0.3157	0.3462	0.34	0.1946
	$s-3$	0.1838	0.3783	0.2277	0.28	0.1052

In order to measure the accuracy of proposed method we build a scenario based on four products. The scenario is shown in Table 4. We assume the following things while building the scenario: (i) the reviews about four products (P-1, P-2, P-3, P-4) are extracted from three different sources ($s-1$, $s-2$, $s-3$), (ii) some sources do not have good credibility, (iii) similarly, some reviews are also spam and (iv) most of the numeric rating provided in spam reviews are far from genuine ratings. In our method all three rating parameters (i.e. Source Credibility, Linear Decay and Rating Trustworthiness) are

Table 6: Accuracy of Aggregation Methods

Aggregation Method	Accuracy (AE)				Average Accuracy (MAE)
	P-1	P-2	P-3	P-4	
Mean	0.77	0.63	0.51	0.41	0.58
W. Mean	0.6	1.3	0.77	0.49	0.79
Mode	0.29	1.08	1.17	1.37	0.97
Median	0.29	0.08	0.32	0.62	0.32
PM(PM+LD)	0.29	0.08	0.17	0.37	0.22
PM+ED, F=0.9	0.29	0.91	0.83	0.37	0.60
PM+ED, F=0.5	0.29	2.91	1.83	1.37	1.6
PM+ED, F=0.3	0.29	2.91	1.83	1.37	1.6
PM=Proposed Method, LD=Linear Decay, ED=Exponential Decay, F=Factor					

considered while computing the predicted reputation. The source credibility measure method is used to compute the credibility of different sources, as shown in Table 5. In order to compute and compare the accuracy of different decay principles, all decay principles are used with our aggregation method and rating parameters.

Table 6 shows the accuracy of proposed method in comparison with other models and aggregation methods. The results show that proposed method achieved the best average accuracy, i.e. MAE is 0.22. Our proposed method also offered the best accuracy AE for all four individual products. On other hand, the average accuracy of mode is 0.97, which is worst of all aggregation methods. Similarly, the median is the nearest counter part of the proposed method, where the MAE is equal to 0.32. The average accuracy of both mean and median is 0.58 and 0.79 respectively which is better than the mode but worst than the median and proposed method. The results also show that exponential decay with different factors (i.e. F=0.9, F=0.5 and F=0.3) is not able to estimate a good reputation value, even if it is used with the proposed method.

6.4. Strategy-Proofness

An aggregation method is said to be strategy proof if it does not provide incentive to a reviewer in order to obtain their preferred reputation value by lying or hiding the actual evaluation (Garcin et al., 2009a). The aggregated value shown on a reputation system influences the way the users rate the

product. Each user has a preferred score that they want to see as an aggregated value, this score represents the true perception of a particular user about the product quality. However, when the preferred score is different than the aggregated value, the users are motivated to post unfair ratings in order to make the aggregated value either equal to the preferred score or even closer. In such case, the aggregation method influence this biased behavior of users. When an aggregation method does not provide any incentive to the users to lie or hide their preferred evaluation, then the users will avoid the biased behavior. This strategy motivates the users to post honest ratings, which ultimately improve the accuracy of the overall reputation model. Robustness and strategy proofness are related but different concepts. Robustness is actually resilience to false and biased ratings. However, strategy proofness go some steps further because it is not only resilience to false and biased ratings but also to those reviewers who want to obtain their preferred reputation value by hiding the actual evaluation. In addition, the strategy proofness limits a single reviewer to change the reputation value along the real line instead of changing it substantially. The incentive that an aggregation method provides to change the reputation value is computed by Equation 15.

$$I = |R_n - R_o| \tag{15}$$

Where I is the incentive provided by the aggregation method to change the reputation value from R_o to R_n after adding false rating. In the case when aggregation method does not provide any incentive then $R_n = R_o$. If the number of cases/occurrences in which $R_n = R_o$ is higher for an aggregation method and I is not substantial when $R_n \neq R_o$, then that aggregation method is said to be strategy proof.

The experiment is performed for nine products extracted from review sites on July, 2014. The products with varying number of ratings, star distribution and aggregated star value are selected. False ratings with lowest star are inserted in the same way as a user U-1, a group of three users U-3 and five users U-5, who intended to give false ratings in order to obtain their preferred aggregated value. We have twenty-seven cases (9*3=27) as shown in Table 7, three for each product. The incentive is measured by determining the change a user or a group of users bring to the previous reputation value.

Table 7 shows the results of strategy proofness for different aggregation methods. Mode, median and the proposed method provided zero incentives

Table 7: Strategy-Proofness of Aggregation Methods

Aggregation Method			Mean	W. Mean	Mode	Median	P.M
Incentive (I)	P-1	U-1	.06	.09	0	0	0
		U-3	.19	.20	0	0	1
		U-5	.30	.42	4	1	1
	P-2	U-1	.09	.14	0	0	0
		U-3	.27	.39	0	0	0
		U-5	.43	.61	0	0	0
	P-3	U-1	.11	.22	0	0	0
		U-3	.31	.58	0	0	1
		U-5	.49	.86	.5	.5	1
	P-4	U-1	.05	.12	0	0	0
		U-3	.17	.36	0	0	0
		U-5	.27	.58	0	.5	0
	P-5	U-1	.08	.17	0	0	0
		U-3	.24	.48	0	0	0
		U-5	.39	.75	0	0	0
	P-6	U-1	.03	.07	0	0	0
		U-3	.11	.22	0	0	0
		U-5	.18	.36	0	0	0
	P-7	U-1	.03	.06	0	0	0
		U-3	.09	.19	0	0	0
		U-5	.16	.31	0	0	0
	P-8	U-1	.02	.05	0	0	0
		U-3	.07	.14	0	0	0
		U-5	.12	.24	0	0	0
	P-9	U-1	.02	.04	0	0	0
		U-3	.07	.14	0	0	0
		U-5	.12	.23	0	0	0

to change the aggregated value in most cases. However, both mean and weighted mean provided some incentives to change the aggregated value in all twenty-seven cases. On the other hand, the proposed method provided incentives in four cases to change the result. Similarly, the median provided incentives only in three cases. The mode provided incentives only in two cases but in one case it given maximum incentive (i.e. 4 in case P-1 and U-5). Furthermore, if we make the group of four users in the same case the mode will provide incentive 2 to change the reputation value. This shows that in some situations the mode can provide maximum incentives in two consecutive false ratings. Conversationally, both median and our proposed method only allow to change the reputation value along the real line which is more desirable in product reputation context.

7. Conclusion

People have more opportunity to express their opinions and to evaluate products and services on the web nowadays. These opinions should be used to compute reputation values which are useful for both customers and organizations to make decisions. In this paper, a multi-source reputation model dedicated to products is proposed. The proposed model increases the computational feasibility of reputation values and also solved the issues of availability and vulnerability of single source-based approaches. A source credibility measure method is also proposed which deals with non credible sources in order to increases accuracy of reputation value. In addition, more robust aggregation methods are proposed to be resilient to false ratings and are able to estimate a fairer reputation value even if some ratings are false. Suitable decay principles for product reputation are proposed, which allow the model to reflect recent opinions about product quickly. This provides the recent reputation value to the decision makers and protects the reputation model from being cheated by opportunistic manufacturer after gaining a good reputation value. Furthermore, a balance between sensitivity and robustness is maintained so that it cannot be compromised by malicious users but still consider newest opinions quickly. Results show that our model is robust, strategy-proof and estimate a good reputation value.

As future works, other reputation values such as feature reputation, feature based product reputation and aggregated general opinion will be integrated into our model in order to increase customer personalization and to improve decision making of organization. A multi-agent based product repu-

tation system will be developed, which will be further integrated with PLM (Product Life Cycle Management System) in order to allow the organization to make decision throughout product life cycle.

Acknowledgement

This project has been funded with support from the European Commission (cLink Project: 372242-1-2012-1-UK-ERA MUNDUS-EMA21). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

Abdel-Hafez, A., Xu, Y., Tjondronegoro, D., 2012. Product reputation model: An opinion mining based approach. In: SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data. p. 16.

Cho, J., Kwon, K., Park, Y., 2009. Q-rater: A collaborative reputation system based on source credibility theory. Expert Systems with Applications 36 (2), 3751–3760.

Connors, L., Mudambi, S. M., Schuff, D., 2011. Is it the review or the reviewer? a multi-method approach to determine the antecedents of online review helpfulness. In: System Sciences (HICSS), 2011 44th Hawaii International Conference on. IEEE, pp. 1–10.

ElSalamouny, E., Krukow, K. T., Sassone, V., 2009. An analysis of the exponential decay principle in probabilistic trust models. Theoretical computer science 410 (41), 4067–4084.

Garcin, F., Faltings, B., Jurca, R., 2009a. Aggregating reputation feedback. In: Proceedings of the First International Conference on Reputation: Theory and Technology-ICORE. Vol. 9. p. 63.

Garcin, F., Faltings, B., Jurca, R., Joswig, N., 2009b. Rating aggregation in collaborative filtering systems. In: Proceedings of the third ACM conference on Recommender systems. ACM, pp. 349–352.

Hawkins, D., Best, R., Coney, K., Koch, E., 2004. Consumer behavior: building marketing strategy. McGraw-Hill/Irwin series in marketing Show all parts in this series.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 168–177.

Jindal, N., Liu, B., 2008. Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, pp. 219–230.

Jøsang, A., Hird, S., Faccer, E., 2003. Simulating the effect of reputation systems on e-markets. In: Trust Management. Springer, pp. 179–194.

Josang, A., Ismail, R., Boyd, C., 2007. A survey of trust and reputation systems for online service provision. Decision support systems 43 (2), 618–644.

Jsang, A., Ismail, R., 2002. The beta reputation system. In: Proceedings of the 15th bled electronic commerce conference. pp. 41–55.

Kinateder, M., Baschny, E., Rothermel, K., 2005. Towards a generic trust model–comparison of various trust update algorithms. In: Trust Management. Springer, pp. 177–192.

Koch, G. G., 1982. Intraclass correlation coefficient. Encyclopedia of statistical sciences.

Korfiatis, N., García-Bariocanal, E., Sánchez-Alonso, S., 2012. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. Electronic Commerce Research and Applications 11 (3), 205–217.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H. W., 2010. Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, pp. 939–948.

Liu, L., Munro, M., 2012. Systematic analysis of centralized online reputation systems. Decision support systems 52 (2), 438–449.

MacLennan, R. N., 1993. Interrater reliability with spss for windows 5.0. The American Statistician 47 (4), 292–296.

Massa, P., Avesani, P., 2007. Trust-aware recommender systems. In: Proceedings of the 2007 ACM conference on Recommender systems. ACM, pp. 17–24.

Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T., 2002. Mining product reputations on the web. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 341–349.

Popescu, A.-M., Etzioni, O., 2007. Extracting product features and opinions from reviews. In: Natural language processing and text mining. Springer, pp. 9–28.

Ripley, B. D., 2001. The r project in statistical computing. MSOR Connections 1 (1), 23–25.

Sabater, J., Sierra, C., 2002. Reputation and social network analysis in multiagent systems. In: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1. ACM, pp. 475–482.

Scott, Ward, T. S., Robertson, J., Zielinski, 1984. Consumer Behavior (Scott, Foresman Series in Marketing). Addison Wesley Publishing Company.

Selcuk, A. A., Uzun, E., Pariente, M. R., 2004. A reputation-based trust management system for p2p networks. In: Cluster Computing and the Grid, 2004. CCGrid 2004. IEEE International Symposium on. IEEE, pp. 251–258.

Sun, M., 2012. How does the variance of product ratings matter? Management Science 58 (4), 696–707.

Turney, P. D., 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, pp. 417–424.

Wiener, J. L., Mowen, J. C., 1986. Source credibility: On the independent effects of trust and expertise. Advances in consumer research 13 (1), 306–310.

Xiong, L., Liu, L., 2004. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. Knowledge and Data Engineering, IEEE Transactions on 16 (7), 843–857.

Yin, L., Yang, R., Gabbouj, M., Neuvo, Y., 1996. Weighted median filters: a tutorial. Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on 43 (3), 157–192.

Zacharia, G., Maes, P., 2000. Trust management through reputation mechanisms. Applied Artificial Intelligence 14 (9), 881–907.