Title	Distortion as a validation criterion in the identification of suspicious reviews
Authors(s)	Wu, Guangyu; Greene, Derek; Smyth, Barry; Cunningham, Pádraig
Publication date	2010-05-02
Series	UCD CSI Technical Reports; UCD-CSI-2010-04
Publisher	University College Dublin
Link to online version	http://www.csi.ucd.ie/content/distortion-validation-criterion-identification-suspicious-reviews
Item record/more information	http://hdl.handle.net/10197/1949

Distortion as a Validation Criterion in the Identification of Suspicious Reviews

University College Dublin Technical Report
UCD-CSI-2010-04
May 2010

Guangyu Wu, Derek Greene, Barry Smyth, Pádraig Cunningham
School of Computer Science and Informatics
University College Dublin, Ireland
{guangyu.wu,derek.greene,barry.smyth,padraig.cunningham}@ucd.ie

ABSTRACT

Assessing the trustworthiness of reviews is a key issue for the maintainers of opinion sites such as TripAdvisor. In this paper we propose a distortion criterion for assessing the impact of methods for uncovering suspicious hotel reviews in TripAdvisor. The principle is that dishonest reviews will distort the overall popularity ranking for a collection of hotels. Thus a mechanism that deletes dishonest reviews will distort the popularity ranking significantly, when compared with the removal of a similar set of reviews at random. This distortion can be quantified by comparing popularity rankings before and after deletion, using rank correlation. We present an evaluation of this strategy in the assessment of shill detection mechanisms on a dataset of hotel reviews collected from TripAdvisor.

Categories and Subject Descriptors

E.0 [**Data**]: General – Data quality; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Evaluation, Algorithms

Keywords

User-generated content, Credibility, Shilling

1. INTRODUCTION

Perhaps the greatest commercial success derived from user-generated content is the use of reviews and recommendations on sites such as Amazon and TripAdvisor [13, 17]. It is recognized that the fact that Amazon has a more extensive collection of user-generated reviews and recommendations than its competitors confers a significant sales advantage [17]. However, this reliance on user-generated content comes at a price. Tripadvisor claims to be the largest site for "unbiased travel reviews" on the internet [13] and if this unbiased claim is brought into question then it can be very damaging for them [9].

This vulnerability of recommender systems to 'shilling' attacks is widely recognized, and there is already an extensive literature on identifying such attacks and on making systems robust to malicious influence [14, 19]. While much of this work has addressed automatic collaborative filtering (ACF) systems (*e.g.* [14]), in the work described here we focus on identifying bogus reviews and ratings that are not necessarily being used in an ACF framework.

In this paper we explore the conjecture that shill reviews are likely to distort popularity rankings given that the objective is to improve the online reputation of a hotel. For instance, the Four Seasons Hotel in Las Vegas is ranked second of 286 hotels in Las Vegas based on 446 reviews. It would be difficult to influence this ranking because of the volume of reviews and ratings available, making it an unlikely target for shilling.

A major challenge for research in this area is the lack of annotated datasets for assessing the effectiveness of shill detection strategies. For this reason, we have gathered a dataset of approximately 30,000 TripAdvisor reviews covering Irish hotels, which we used in our evaluation. This evaluation has two components: an assessment of the distortion impact of a number of review deletion policies, and an analysis of the ability to detect artificial shill reviews that we add to the dataset. The findings of both aspects of the evaluation are quite encouraging.

The paper proceeds as follows. In the next section we provide a brief overview of shilling and ballot stuffing and in Section 3 we present the basic shill detection strategies that we use in our evaluation. In Section 4 we introduce the idea of using distortion as a principle for validating shill detection strategies, and in Section 5 we present an evaluation of this on the Irish TripAdvisor data. The paper concludes with a summary and some suggestions for future work.

2. RELATED WORK

People often seek recommendations from others in relation to which movies to see, which books to read, or which hotel to choose. In our purchasing decisions, we are happy to be influenced by others – not just by friends but by strangers as well. We will steer clear of restaurants that have few customers, while on e-commerce sites we are more comfortable with merchants that have many endorsements from satisfied customers.

This openness to recommendations is vulnerable to shilling whereby an agent, in collusion with the seller of an asset or service, heaps praise on mediocre offerings. This practice, which has existed in the real world for centuries, has found its way into online opinion and recommendation sites [14, 19]. The proliferation of such practices can lead us to question whether the gap in quality and unreliability between user-generated content and expert editorial opinion could render the former valueless [1]. Notably while people are relatively adept at assessing the reliability of agents in the real world and valuing recommendations from such sources accordingly, it is much more difficult to make such judgments in online environments.

Indeed shilling has become so much a part of recommendation and opinion websites, that the term is more likely to be encountered in an online context than in the real world[1]. This is also reflected in the emergence of a new vocabulary for describing shilling and related activities:

- **Astroturfing** refers to activity that gives the impression of the existence of popular or 'grassroots' support for an organization, product or service.
- A **sockpuppet** is a second online identity used by a member of an internet community to talk themselves up, and lend credence to their opinions.
- **Ballot stuffing** can occur on sites such as eBay that have explicit reputation reporting systems for customers to rate merchants. This activity involves the submission of fake ratings designed to boost or damage reputation scores [6].
- **Bad mouthing** entails assigning unfairly low ratings to a merchant or service provider[6]. It can also cover unfair commentary in reviews without explicit ratings[7].

In this paper, while we are interested in identifying shill reviews, we are essentially addressing ballot stuffing as we are particularly concerned with how the numeric ratings in shill reviews affect popularity rankings. This work is not typical of mainstream research on shilling in computer science which has so far addressed two main themes, shilling in ACF [14, 12, 5] and ballot stuffing on online auctions sites [3].

[1]Articles that mention shilling in the context of restaurant reviews, hotel reviews and cosmetic surgery: `http://ny.eater.com/tags/adventures-in-shilling` `http://www.usatoday.com/tech/hotsites/2009-07-15-trip-advisor_N.htm` `http://www.nytimes.com/2009/07/15/technology/internet/15lift.html`.

If we consider the identification of spam reviews as a subset of the larger problem of identifying reviews that are authoritative, credible or helpful, then there is some interesting research to draw on. Both O'Mahony & Smyth [16] and Hsu et al. [8] cast the problem of ranking reviews in a supervised learning framework, and show impressive results. O'Mahony & Smyth use customer feedback on the helpfulness of reviews on Amazon to provide the *supervision*, while Hsu et al. use feedback provided from Digg. Unfortunately in the TripAdvisor scenario there is no user feedback to support a supervised learning approach.

There are many related or analogous problems that have received attention – in particular e-mail spam [4], link spam (search engine spam) [2], detecting attacks on recommender systems [15], and assessing authoritativeness on sites such as Wikipedia [11].

The idea of using *distortion* to identify anomalous behavior is not new. For instance this general principle has been used to reveal link spam [2] and to identify untrustworthy participants in peer-to-peer search networks [18].

Before concluding this brief review it is worth mentioning *bid shilling* in online auctions. Here shilling is sometimes used to refer to bids placed by a merchant in order to run up the price in an online auction [10]. This is somewhat different to what we are concerned with here, as the objective is to inflate the *price* rather than the reputation of the asset or service.

3. SHILL DETECTION

The focus of this paper is on distortion as a validation criterion in shill detection rather than on features that are predictive of shills so the features we employ are quite basic. The two features we consider are based on the idea of positive singletons as shown in Figure 1. Positive singletons are positive reviews from reviewers who have posted no other reviews.

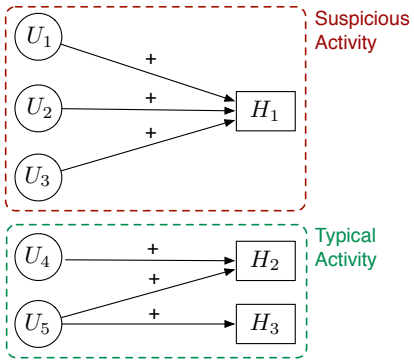

Figure 1: Bipartite graph representing a simple scenario involving five users and three hotels.

Proportion of Positive Singletons (PPS): The PPS score for hotel H is the proportion of reviews on that hotel that are positive singletons:

$$PPS(H) = \frac{N_{ps}}{N} \quad (1)$$

where N_{ps} is the number of positive singleton reviews, and N is the total review count for the hotel.

Concentration of Positive Singletons (CPS): Often multiple shill reviews will be injected in quick succession.[2] The greater the degree of temporal clustering between positive singletons, the more suspicious these reviews appear.

Given the list of of positive singleton reviews $\{r_1, \ldots, r_P\}$ for a hotel H arranged in ascending order by submission date, we define a score for H as a function of the average date distance D (*i.e.* number of days) between each review r_i and its temporally nearest neighbor:

$$CPS(H) = \frac{1}{P}\sum_{i=1}^{P} e^{-\lambda \times min(D(r_i, r_{i-1}), D(r_i, r_{i+1}))} \quad (2)$$

where λ is a bandwidth parameter that controls the influence of the proximity of reviews. We found that a value of $\lambda = 1$ was most effective on the TripAdvisor data.

4. VALIDATION USING DISTORTION

Our proposal for using distortion to validate the filtering of suspicious reviews is based on the prominence given to user-based popularity rankings on many e-commerce sites. For instance, TripAdvisor assigns a ranking to each hotel in a given region (*e.g.* 2^{nd} of 446 hotels in Las Vegas). Our contention is that a common objective of shilling will be to influence this ranking. Deleting a set of reviews chosen at random should not overly disrupt the ranked list of hotels, while deleting shill reviews should significantly alter or distort the ranking of hotels to reveal the "true" ranking.

It is important to state that TripAdvisor do not disclose the details of their ranking algorithm. However, it is clear that the main component is the average reviewer rating, as their ranked lists are strongly correlated with lists ordered simply based on average rating. Since we can recalculate the average reviewer rating after review deletion, we use this to produce the popularity ranking used in our experiments. We first calculate a raw distortion score resulting from the deletion of suspect reviews. We subsequently calculate an adjusted distortion score which takes account of the impact of deleting a similar number of reviews chosen at random.

Raw Distortion: The raw distortion score simply quantifies the change in popularity ranking resulting from deleting a number of suspicious reviews. It is calculated as the rank correlation between the original popularity ranking and the popularity ranking after the suspicious reviews have been deleted. More formally, if P is the original popularity ranking where P_i is the rank of the i^{th} hotel and S is the ranking after deleting shills, then the raw distortion after deleting suspected shills is:

$$RD = SRC(P, S) = \frac{\sum_i (P_i - \bar{P})(S_i - \bar{S})}{\sqrt{\sum_i (P_i - \bar{P})^2 \sum_i (S_i - \bar{S})^2}} \quad (3)$$

where $SRC(P, S)$ is the Spearman rank correlation of the two rankings and $\bar{P}$ is the average rank in P. Lower values indicate a higher level of distortion.

Adjusted Distortion: To allow comparisons across hotels where different numbers of reviews may be deleted, it is useful to adjust the raw distortion score to account for this. This is done by assessing the impact of deleting a similar number of positive reviews from a hotel with a similar number of overall reviews. The adjusted distortion score is the difference between this expected distortion score and the raw distortion score. Significant adjusted distortion scores will be positive and insignificant scores will be close to zero. This signifies that there is no difference between deleting the suspected reviews and simply deleting reviews at random. The adjusted distortion score AD for S, which incorporates an expected distortion of ED based on a ranking R after random deletions, is given by:

$$AD = ED - RD = \overline{SRC(P, R)} - SRC(P, S) \quad (4)$$

In practice, the expected distortion score ED is calculated by repeatedly choosing hotels at random of a similar size, removing positive reviews, and calculating the raw distortion – the expected is score is given by the average raw distortion over many runs.

5. EVALUATION

In our evaluation we explore whether the distortion in popularity rankings is an effective mechanism for validating the output of shill detection processes. Firstly we examine the impact on distortion of review deletion based on the PPS and CPS scores described previously in Section 3. Then in Section 5.2 we present an evaluation based on the insertion of artificial shills into the TripAdvisor dataset.

The Irish TripAdvisor dataset[3] used here comprises 29,799 reviews from 2,1851 unique reviewers, covering hotels from all regions of Ireland over a two-year time window from September 2007 to September 2009. Note that we only consider a subset of 843 hotels which received four or more reviews during this time. Approximately two thirds of the reviews are positive – *i.e.* awarding at least four out of five stars. Of these roughly 30% are positive singletons as defined in section 3.

A time-plot of the reviews for a typical hotel from the TripAdvisor dataset is shown in Figure 2. For this hotel there is a reasonable balance between singleton reviews and reviews from users who have submitted multiple reviews. Other, perhaps more suspicious cases, are shown in the time-plots in Figures 4 and 6.

5.1 Evaluation on TripAdvisor Data

The scatter plot in Figure 3 shows the top 20 most suspicious hotels as ranked by the PPS score, with PPS scores plotted against adjusted distortion in the popularity ranking. Half of the top hotels have negligible or negative distortion when the 'suspicious' reviews are deleted, suggesting that these reviews are unlikely to be shills.

An example time-plot for one of these hotels is shown in Figure 4. This hotel has a highly suspicious PPS score because an overwhelming 83% of the positive reviews are singletons

[2]The review spam recently discovered on Apple's App Store had this characteristic `http://edition.cnn.com/2009/TECH/12/09/wired.apple.apps/index.html`

[3]Available at: `http://mlg.ucd.ie/datasets/trip`.

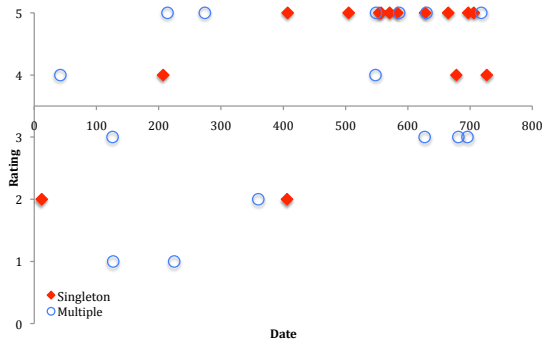

Figure 2: A time plot of the reviews for a typical hotel covering a two year period (730 days).

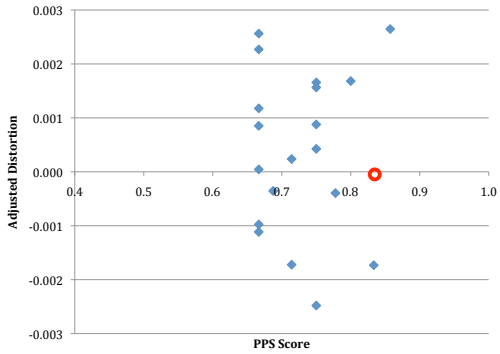

Figure 3: The top 20 hotels as ranked by the PPS score. The chart plots the PPS score against the adjusted distortion. The corresponding time-plot for the hotel marked by the circle is shown in Figure 4.

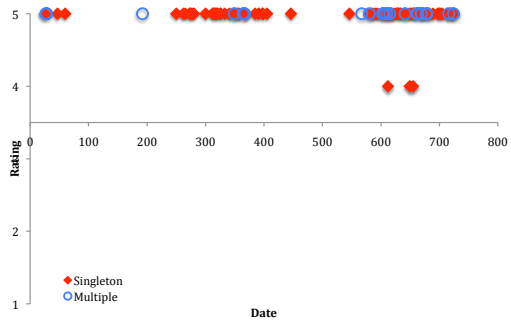

Figure 4: The time-plot of the hotel marked with the circle in Figure 3.

(101 of 121 reviews). Remember that the baseline for the whole dataset is that ≈ 30% of positive reviews are singletons. However, it can be seen in Figure 3 that the distortion score for this hotel is close to zero suggesting that this is a *false positive*. This is because all the non-singleton reviews are also positive so deletion of the purported suspicious reviews does not distort the popularity ranking. Furthermore, an inspection of the text of the suspicious reviews suggests that they might be genuine. We speculate that there may be something more innocent than full-scale shilling going on here – perhaps the hotelier is soliciting reviews from satisfied customers?

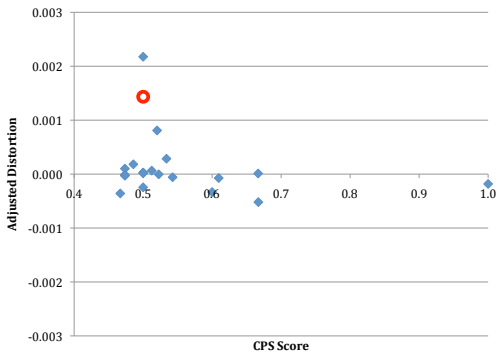

Figure 5: The scatter plot for the top 20 hotels based on the CPS score. The corresponding time-plot for the hotel marked by the circle is shown in Figure 6.

In Figure 5 we show the scatter plot for the top 20 hotels as ordered by the CPS score. It is interesting to note that there is far less negative distortion in this plot. This is because the CPS score has no bias towards hotels with few reviews. Thus distortion will be positive or close to zero. This contrasts with the situation for the PPS score, which is inclined to select hotels with few reviews and thus can result in significant negative distortion when a large fraction of a small review set is deleted.

The time-plot for the hotel marked with the circle is shown in Figure 6. The reviews producing the high CPS score are the two shown in the top right of the plot. When these are deleted, the average rating goes from 4⋆ to 3.3⋆, resulting in a significant distortion. We feel this is valid as the two positive singletons look suspiciously like a management response to the strongly negative review that immediately precedes them. This view is supported by an inspection of the text of the reviews.

While we believe this analysis highlights some suspicious behavior, it is difficult to draw any concrete conclusions in the absence of a 'ground truth' to compare against. It must also be remembered that TripAdvisor claim that they "catch the vast majority of suspicious reviews"[4]. For this reason, in the next section we extend the analysis by including artificial shills in the data.

5.2 Evaluation on Artificial Data

The model we use to produce artificial shills is based on recent reporting of shilling behavior on the web. In November 2009 Apple removed a number of iPhone apps from their App Store because the supplier of these apps was found to be shilling the reviews and ratings for these apps. The artificial shills we use here are based on the publicly available details regarding these shills[5]. In the shilling example presented in detail it appears that 42 of 44 reviews are 5⋆ singleton shills. The artificial data used in this evaluation are less extreme examples based on this template. Table 1 provides details on the construction of these artificial hotels.

[4]http://www.elliott.org/blog/TripAdvisors-kaufer-we-catch-the-vast-majority-of-suspicious-reviews/

[5]http://www.iphoneography.com/journal/2009/11/28/apple-investigates-possible-us-appstore-ratings-scam.html

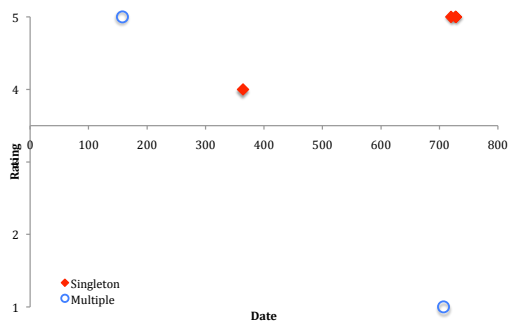

Figure 6: The time-plot of the hotel marked with the circle in Figure 5.

Hotel Id.	Real Reviews	Shills
H_1	$1 \times 5\star$, $2 \times 1\star$	$40 \times 5\star$
H_2	$1 \times 5\star$, $2 \times 1\star$	$30 \times 5\star$
H_3	$1 \times 5\star$, $2 \times 1\star$	$20 \times 5\star$
H_4	$1 \times 5\star$, $2 \times 1\star$	$10 \times 5\star$
H_5	$1 \times 5\star$, $2 \times 1\star$	$5 \times 5\star$
H_6	$1 \times 5\star$, $2 \times 1\star$	$2 \times 5\star$
S_1	$1 \times 5\star$, $2 \times 1\star$	$10 \times 5\star$
S_2	$1 \times 5\star$, $2 \times 2\star$	$10 \times 5\star$
S_3	$1 \times 5\star$, $2 \times 3\star$	$10 \times 5\star$
S_4	$1 \times 5\star$, $2 \times 4\star$	$10 \times 5\star$
S_5	$1 \times 5\star$, $2 \times 5\star$	$10 \times 5\star$

Table 1: Details on the artificial data included in the TripAdvisor dataset. For instance, hotel H_5 has one $5\star$ and two $1\star$ reviews that are legitimate, and also five shill $5\star$ reviews.

Unfortunately, because these App Store shills were all posted over a very short time period, they are easy to detect using the proposed CPS score. Therefore the CPS score is not really challenged in this evaluation, and so we focus on the performance of the PPS score for the remainder of this section. It is interesting to note that, even though this shilling is not very sophisticated, it was not noticed by Apple until it was brought to their attention by a user of the store.

Figure 7 shows how these shills are ranked by the PPS score. Artificial hotels H_1-H_4 appear in the top 20. Hotel H_5 (with five $5\star$ shills) is ranked 29^{th} of the 849 hotels (843 real + 6 artificial) hotels, while H_6 (with only two $5\star$ shills) does not show up on the PPS score. H_1-H_5 all produce appreciable distortion – the distortions for H_1-H_4 are shown in the scatter plot.

The performance of CPS on H_1-H_6 is shown in Figure 8. Note that, because the shill reviews in these examples all have the same date, they have a very strong signature when we apply the CPS score. It can be seen that the distortion decreases from H_1 to H_6, with H_6 having no appreciable distortion.

The objective with artificial hotels S_1-S_5 is to explore the behavior of distortion. For this reason the extent of the shilling in these examples is kept constant and the legitimate reviews are changed. S_1 should show significant distortion, while for S_5 the distributions of legitimate and shill reviews

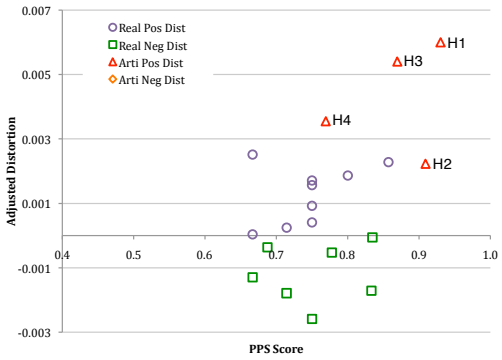

Figure 7: PPS scores on the artificial data H_1-H_6.

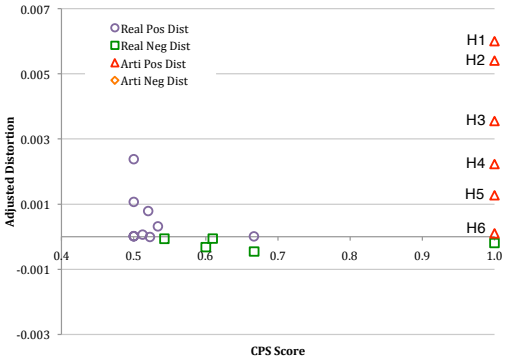

Figure 8: CPS scores on the artificial data H_1-H_6.

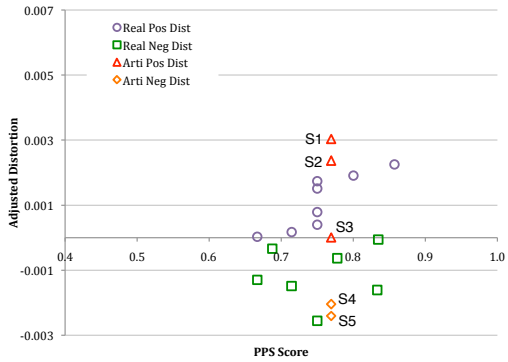

Figure 9: PPS scores on the artificial data S_1-S_5.

should be same (since all reviews are $5\star$). The results for the PPS score on S_1-S_5 are shown in Figure 9. It is apparent that S_1 and S_2 have a clear distortion signature, while S_3 has no adjusted distortion, and S_4 and S_5 both have negative distortion.

This evaluation on artificial data does not tell us much about the CPS score as it is trivial to detect shills that all share the same date. However, it does show that PPS is an effective shill detection mechanism when the generators of the fake reviews do not take the trouble to produce realistic personas that have generated multiple reviews. It also shows that our notion of distortion is effective for highlighting shilling.

6. CONCLUSIONS

The objective of the work described in this paper is to explore distortion in popularity ranking as a measure for highlighting shilling. We have presented a preliminary evaluation on real and artificial data that supports this. We have used two scores based on the proportion of positive singleton reviews and the concentration of positive singletons to highlight suspicious behavior, and have then shown that distortion helps to separate out true positives (Figures 5 & 6) from false positives (Figures 3 & 4).

Clearly, if distortion is effective for validating other shill scoring mechanisms, then it would make sense to integrate it into a multi-variate shill detection mechanism. The difficulty with integrating the validation mechanism into the detection process is the problem of validating results. We plan to explore this issue in future work.

Acknowledgments: This research was supported by Science Foundation Ireland (SFI) Grant No. 08/SRC/I1407.

7. REFERENCES

[1] R. Baeza-Yates. User Generated Content: How Good Is It? In *3rd Workshop on Information Credibility on the Web (WICOW 2009)*, pages 1–2, 2009.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Using rank propagation and probabilistic counting for link-based spam detection. In *Proc. Workshop on Web Mining and Web Usage Analysis (WebKDD)*, 2006.

[3] R. Bhattacharjee and A. Goel. Avoiding ballot stuffing in ebay-like reputation systems. In *Proc. 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems*, page 137, 2005.

[4] P. O. Boykin and V. P. Roychowdhury. Leveraging social networks to fight spam. *IEEE Computer*, 38(4):61–68, 2005.

[5] K. Bryan, M. O'Mahony, and P. Cunningham. Unsupervised retrieval of attack profiles in collaborative recommender systems. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 155–162, 2008.

[6] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proc. 2nd ACM Conference on Electronic Commerce*, pages 150–157, 2000.

[7] C. Dellarocas. Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms. *Management Science*, 52(10):1577–1593, October 2006.

[8] C. Hsu, E. Khabiri, and J. Caverlee. Ranking Comments on the Social Web. In *Proc. 2009 International Conference on Computational Science and Engineering-Volume 04*, pages 90–97, 2009.

[9] R. Jurca and B. Faltings. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34(1):209–253, 2009.

[10] R. Kauffman and C. Wood. Running up the bid: detecting, predicting, and preventing reserve price shilling in online auctions. In *Proc. 5th International Conference on Electronic Commerce*, page 265, 2003.

[11] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative sources using social networks: an insight from Wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[12] S. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *Proc. 13th international conference on World Wide Web*, pages 393–402, 2004.

[13] S. Litvin, R. Goldsmith, and B. Pan. Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, 29(3):458–468, 2008.

[14] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre. Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology (TOIT)*, 4(4):344–377, 2004.

[15] M. P. O'Mahony, N. J. Hurley, and G. C. M. Silvestre. Recommender systems: Attack types and strategies. In M. M. Veloso and S. Kambhampati, editors, *AAAI*, pages 334–339. AAAI Press / The MIT Press, 2005.

[16] M. P. O'Mahony and B. Smyth. Learning to recommend helpful hotel reviews. In L. D. Bergman, A. Tuzhilin, R. D. Burke, A. Felfernig, and L. S-Thieme, editors, *RecSys*, pages 305–308, 2009.

[17] T. O'Reilly. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Technical report, Communications & Strategies, No. 1, p. 17, First Quarter 2007. Available at SSRN: http://ssrn.com/abstract=1008839, 2007.

[18] J. Parreira, D. Donato, C. Castillo, and G. Weikum. Computing trusted authority scores in peer-to-peer web search networks. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 80. ACM, 2007.

[19] J. Staddon and R. Chow. Detecting reviewer bias through web-based association mining. In *2nd Workshop on Information Credibility on the Web (WICOW 2008) at ACM CIKM'08*, 2008.