How to Support the Lay Users Evaluations of Medical Information on the Web?

Katarzyna Abramczuk[1(✉)], Michał Kąkol[2], and Adam Wierzbicki[2]

[1] Institute of Sociology, University of Warsaw, Warsaw, Poland
k.abramczuk@uw.edu.pl
[2] Polish-Japanese Academy of Information Technology, Warsaw, Poland
{michal.kakol,adamw}@pjwstk.edu.pl

Abstract. In this paper we present a study on the credibility of lay users evaluations of health related web content. We investigate the differences between their approach and the approach of medical experts, analyse whether we can increase their accuracy using a simple support system, and explore the effectiveness of the wisdom of crowds approach. We find that a support system based on expert ratings is effective while relying on the wisdom of crowds can be risky. There is a clear positive bias in lay evaluations that is very difficult to correct. Moreover lay users perceive health related web-content differently than medical experts.

Keywords: Web credibility · Wisdom of crowds · World wide web · Information retrieval · Health information

1 Introduction

People often search the Web for medical information. As a matter of fact eight in ten people browse the Internet for health related content, which makes it one of the most common Internet activities. Unfortunately, what the users ultimately find is often misleading, incomplete, and non-credible. The Web is filled with a myriad of humbug therapies, mysterious super-drugs and pseudo doctors. As it can be easily guessed it may, and often does, lead to grave consequences for both private and, as can be seen by the example of the anti-vaccine movement, social matters.

We present a study on the process by which people evaluate the credibility of medical web content and a system supporting lay evaluations. We try to understand what guides peoples decisions, to what extent we can influence them and how a supporting system should be constructed in order to maximise its positive effect.

The paper is structured as follows. Firstly, we introduce the concept of web content credibility and describe the study design. Then we contrast expert and lay evaluations to see how they differ and whether lay users are able to discern credible and non-credible medical content on their own. Finally, we analyse whether and to what extent lay users are prone to following advice provided by

S. Yamamoto (Ed.): HIMI 2016, Part I, LNCS 9734, pp. 3–13, 2016.
DOI: 10.1007/978-3-319-40349-6_1

a supporting system. We do this by comparing how accurate and inaccurate system suggestions affect users, while contrasting this with the power of the wisdom of crowds.

2 Related Work

Credibility is a multifaceted concept. The majority of researchers agree that it is composed of at least two components: expertise and trustworthiness. Both these components are naturally associated with the source of the information being assessed [1]. Expertise is defined in terms of the competency or experience of the author and trustworthiness in terms of the goodwill and agenda of the source [2]. There is also a number of other source characteristics that are related to the concept of credibility such as completeness of information or information accuracy. One of the dimensions which was assessed in our study is controversy of the information on the page. In fact this dimension turns out to be important for explaining the differences between lay and expert evaluations. Jankowski-Lorek et al. in their work make attempts to automatically detect controversy in the content based on crowdsourced ratings distributions [3] or Wikipedia category structure [4].

Apart from the correlates of credibility related to the information piece and its source a number of viewer traits influence the individual process of credibility assessment. For example Rafalak et al. [5,6] investigated how different psychological and demographical traits of Internet users affect the credibility perception. Flanagin and Metzger [7] add the internet/web experience of the viewer as being one of the crucial factors. Similarly in [8] it is pointed out that Internet usage efficacy is an important determinant of the assessment of credibility of web sources. Another obvious viewer related factor is the viewers familiarity with the assessed information [9]. Lucassen et al. [10] also found that various user characteristics such as domain expertise and information search skills affect credibility evaluations. The effect of the viewers traits has also been reported specifically in the health domain e.g. Crawford et al. [11] have explored health information search behaviors in this context.

The usefulness of Google as an indicator of medical Websites credibility has been evaluated by Frick et al. [12] who evaluated the PageRank score as one indicator of quality. Their results show that it is not inherently useful for discrimination or helping users to avoid inaccurate or poor information. Griffiths et al. [13] evaluated PageRank scores with evidence based quality scores for depression websites (expert ratings). Again PageRank scores correlated weakly with the evidence based scores. This shows that, considered alone, PageRank indicator is clearly insufficient to account for the quality of medical Websites.

In this paper we address the problem of increasing the accuracy of lay credibility evaluations. We evaluate the usefulness of a system supporting credibility evaluations and test the reliability of the wisdom of crowds approach. The idea of augmenting web content in order to enhance credibility judgments is not new. For example Schwarz and Morris [14] present visualizations based on expert user

Table 1. The list of medical topics used in the study together with the number of webpages and the number of lay user evaluations.

Category	No. of webpages	No. of evaluations
Celiac disease treatment	23	1112
Depression treatment	17	776
Diabetes treatment	22	1023
Heart disease treatment	19	901
Hormonal contraception	21	979
Norovirus treatment	21	979
Twitch eye treatment	24	1143
West Nile virus	19	913
Whooping cough treatment	24	1152

behavior additionally to search results. Yamamoto and Tanaka [15] also augment web search results with visualizations and re-rank the results according to the users predicted credibility model. A simpler augmentation in a similar setting is also described in an Amin et al. [16] study covering culture related web content.

3 Study Design

The study we present is based on an experiment. A single task in this experiment consisted of evaluations of three websites drawn randomly from a corpus of 190 webpages on 9 health-related topics. We chose websites using most popular, and those increasing the most in popularity, medical searches performed on Google from the US in 12 months preceding the study and chose 9 topics. We constructed corresponding search phrases aimed at finding both credible and not credible information. We selected random websites from those that were shown on the first results pages (as searched for in the US) and filtered out those that did not contain any meaningful content that could be subject to credibility evaluation. The list of medical topics used in the study together with the number of webpages and the number of lay user evaluations used in this paper for each topic is presented below (Table 1):

To run the study we used the Reconcile[1] (Robust Online Credibility Evaluation of Web Content) system which is a product of a joint research project of two universities: Polish PJIIT and Swiss EPFL. The system is a prototype of an online support platform for credibility evaluations.

Two types of users took part in our study: lay users recruited via Amazon Mechanical Turk and medical experts from the Medical University of Warsaw. The medical experts took part in the study during two medical congresses and from their houses. Based on their evaluations each website was assigned one

[1] http://www.reconcile.pl.

unique expert rating that was a result of experts consensus. The lay users were MTurk workers from English-speaking countries (mostly US). All users were paid accordingly for their participation in the study.

The lay user data presented in this paper is based on 3 experimental treatments that were part of our study. Participants were assigned to the treatments randomly. In each case they were asked to evaluate three websites with respect to 6 dimensions: credibility, expertise, intentions of the authors, completeness, controversy of the information provided, and appearance of the website. For all the dimensions except for controversy we used a 5-point ordinal scale where choices were labeled (e.g. 1 completely not credible 2 mostly not credible etc.). Controversy was measured using dichotomous choice (controversial vs. not controversial).

The experimental treatments we are reporting here were as follows:

NH (No Hints) Treatment: In this condition we asked participants to evaluate the webpages without any external support.

EH (Expert Hint) Treatment: In this condition we used Reconcile interface to present the suggested evaluations of the webpages. The suggested evaluations were based on previously gathered expert ratings of the given page and were presented in a form of a traffic light (e.g. green corresponded to credible and red to not credible content) next to the field where the user was submitting his own rating. This mimics the workings of most of the existing support systems.

RE (Reversed Expert) Treatment: This condition was identical to the EH condition except for the fact that the suggested ratings were exactly opposite to the ratings provided by the experts. We chose 158 webpages that received expert credibility ratings other than 3 (the middle of the scale) and marked the credible websites as not credible and vice versa.

To avoid learning that the system suggestions are misleading in RE conditions, this condition and the EH condition were joined. The participants were served either three webpages in the EH condition or two pages in the EH condition and one (the last in the package) in the RE condition. In all the conditions we tracked the subjects activity i.e. page clicks, time spent on each task, going back and forth within the quest etc.

4 Lay and Expert Evaluations

The distribution of credibility ratings for the experts and for the lay users in all the experimental conditions is presented in Fig. 1. 45 (24 %) out of the 190 webpages in the corpus were rated by the experts as not credible (evaluations 1 and 2). 113 (60 %) were rated as credible (evaluations 4 and 5). This distribution for the experts shows a familiar skew that is present in most of the online assessment

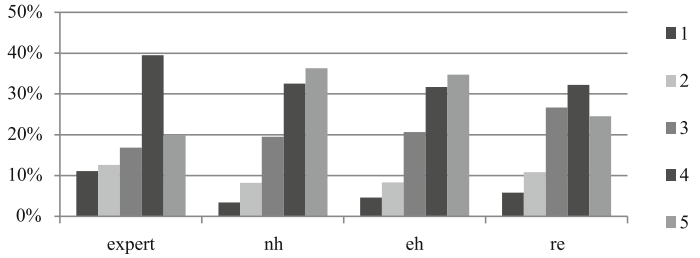

Fig. 1. The distribution of credibility ratings for the experts and for the lay users in all the experimental conditions.

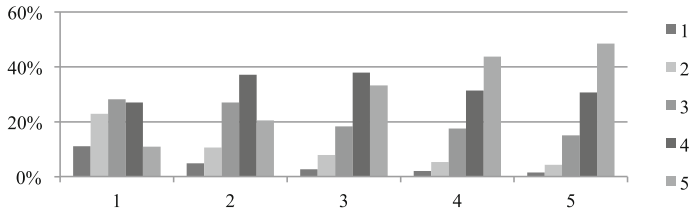

Fig. 2. The distribution of unsupported lay ratings (NH) in the subgroups of web-pages grouped by expert evaluation.

systems. In this case it probably results from the fact that websites were chosen using Google search which successfully filtered out much of the noncredible content.

There is a visible and statistically significant (p = 0,01) difference between expert distribution and lay distributions in any of the conditions presented. In particular the experts are much less prone to using the maximal (5) rating and use the negative (1 and 2) ratings more often. The maximal rating is the most popular choice for the lay users both without any support (NH) and with an expert suggestion (EH). It is slightly different for the RE treatment which will be discussed in the next section. The remaining part of this section is based entirely on the data from the NH treatment.

The lay evaluations and the expert evaluations are correlated. Figure 2 presents the distribution of unsupported lay ratings in the subgroups of webpages grouped by expert evaluation. The share of negative ratings rises systematically with the falling expert rating, starting with 7 % for the most credible and ending with 34 % for the least credible content. Unfortunately, even in the latter case there are more positive (4 and 5) than negative ratings. They amount to 38 % of all the evaluations of the completely noncredible web pages. As a result, the ratings distribution in this case is highly dispersed. The Leik measure [17] of ordinal dispersion equals 0.47 here while in the other cases it oscillates around 0.40.

There are clear limits to the ability of the individual lay users to discern the valuable and trash health information on the web. When no support is provided only 27.72 % of all their evaluations are exactly the same as expert

evaluations of the given webpages while 49.72 % are higher. The main problem is therefore a tendency to overestimate the credibility of the medical information encountered on the Internet. If we take into account the fact that both values 4 and 5 indicate credible content while choosing ratings between 1 and 3 indicates various amounts of doubt, the problem becomes much less dramatic. For the following analyses we divided all the ratings into positive (4 or 5) and negative (other values). In this case 62.53 % of the lay evaluations are the same as expert evaluations i.e. both groups think the content as (generally) credible or both group consider it (generally) not credible. The share of overly positive ratings equals 23.43 % and the share of overly negative ratings equals 14.03 %.

The situation improves even further when we rely on the wisdom of crowds. Under this approach individual lay ratings are aggregated to produce a group-based evaluation that in theory should be more accurate as it eliminates the individual biases. As [18] suggests that when imperfect judgments are aggregated in the right way, the collective intelligence is often excellent. To test this supposition we computed median lay ratings for each page. Two facts became clear. First, the share of accurate lay ratings did indeed increase from 62.53 % to 70.52 %. Second, the share of overly optimistic ratings remained practically unchanged and equaled 23.68 %. In other words, while the wisdom of the crowds approach can help to evaluate accurately the websites that the individual users tend to unjustifiably see as not credible, it does little to identify the content that is seen as credible and is in fact treacherous. This is due to a clear positive bias in lay evaluations.

We investigated the possible sources of the differences between expert and lay evaluations. In particular we wanted to know which other characteristics that were rated by the participants of the study are most useful when predicting their credibility evaluations. Towards this purpose we ran logistic regressions with the dichotomised credibility rating as the dependent variable and the dichotomised other ratings provided by the same group as the independent variables. The results are presented in Table 2. In general all the evaluated dimensions are very

Table 2. Logistic regressions results. Dichotomised credibility rating dichotomised other dimensions.

	Lay users				Experts			
	Coef.	Robust std. error	Z	Sig.	Coef.	Std. error	Z	Sig.
Appearance	0.562	0.15	3.80	<0.001	0.973	0.47	2.06	0.04
Completeness	1.101	0.15	7.44	<0.001	1.210	0.58	2.09	0.04
Expertise	2.604	0.15	17.38	<0.001	1.265	0.59	2.14	0.03
Intentions	1.372	0.16	8.38	<0.001	1.657	0.48	3.44	<0.001
Controversy	−1.213	0.15	−8.11	<0.001	−1.696	0.48	−3.57	<0.001
Pseudo R squared	0.5232				0.5060			

strongly correlated in both groups. For the lay users clearly the most important factor is the perceived expertise of the author. The content in which the author is thought of as an expert has 13.5 (OR) times higher chance of being rated as credible than the content written by someone not perceived as an expert. The least important determinant of the perceived credibility for this group is the evaluation of the website appearance. In the case of experts the role of all the other dimensions is weighted in a more balanced way. The most important factor being the perceived intentions of the author. The content in which the author is thought to have good intentions has 5.2 (OR) times higher chance of being rated as credible than the content written by someone whose intentions seem dubious. Once again the websites perceived appearance is of least importance.

The relatively small role of a websites appearance may come as a surprise and, as it shall be shown later, its actual influence is much larger. Please note that appearance is a characteristic that can in principle be evaluated with equal competence by both the medical experts and the lay users. We therefore computed the correlation coefficient between appearance evaluations made by the experts and the lay users. They are significantly (p=0,01) correlated, but the correlation coefficient is only 0,135. It is small in comparison to correlations between different dimensions rated by the same group, virtually all of which exceed 0,3. Moreover the same pattern applies to all other evaluated characteristics i.e. all of them are tightly connected within the group (the experts or the lay users) and loosely related to the same characteristics as seen by the other group. Both the experts and the lay users form very coherent images of the content they evaluate but their images do not correspond with each other. In other words the experts and non-experts almost literally see medical web content differently.

5 To Follow or to Resist

In this section we will investigate whether the lay evaluations can be improved by a support system offering simple suggestions concerning the contents reliability. In general the answer is positive. The share of lay evaluations that are exactly the same as the expert evaluations rises by almost 10 % points (from the previously mentioned 27.72 % to almost 37 %). Unfortunately the share of overly positive ratings falls only slightly (from the 49.72 % to 46.22 %). When concentrating on the dichotomous evaluations (4–5 credible, 1–3 not credible) the share of accurate ratings equals 74.51 % and the overly positive ratings amount to only 17.63 % of all the ratings. When using the wisdom of crowds approach an impressive 87.9 % of the webpages are generally classified adequately. The remaining 12 % is however still overrated.

We compared median evaluations of all the webpages in the NH and the EH treatment. We found that all the webpages that were underrated when no suggestions were made got positive evaluations when the support system was available. However, from the webpages that were spontaneously overrated by the lay participants only 53 % got negative evaluations when the system revealed that they were not trustworthy. Once again it appears that the positive evaluation bias is exceptionally persistent.

Table 3. The average shares of webpages that got positive evaluations on the remaining four dimensions within the webpages with corrected and uncorrected group ratings.

	EH treatment			NH treatment		
	Corrected	Not corrected	Sig.	Corrected	Not corrected	Sig.
Expertise	0.125	0.762	<0.001	0.875	0.952	0.181
Intentions	0.750	0.952	0.031	1.000	1.000	-
Completeness	0.125	0.667	<0.001	0.750	0.905	0.088
Appearance	0.250	0.762	<0.001	0.750	0.952	0.031

Trying to understand which mechanism stands behind this unfortunate case, we inspected the differences between those webpages overrated in the NH treatment where the ratings were and were not corrected when the support system was available. First we note that almost all the webpages that were wrongly thought credible when no suggestions were offered were rated as controversial by the lay people. We therefore computed the average shares of webpages that got positive evaluations on the remaining four dimensions within the webpages with corrected and uncorrected group ratings. The results are presented in Table 3.

Three facts are worth noticing. First, in the EH condition the measured characteristics are once again highly correlated with the credibility rating. In the case of the websites whose ratings were corrected to not-credible, the mean share of positive ratings on all the other dimensions is significantly lower. Second, the general image of the analysed websites in the EH and NH conditions is visibly different. When the support system suggests that the website is a bad one the mean share of positive evaluations becomes much lower. Last but not least, if we were to predict, the image of which overrated websites will be corrected when we put the support system to work, relying solely on the ratings provided without this system, the only useful cue would be the perceived website appearance. Only in this case there is a statistically significant difference between the websites with corrected and uncorrected group ratings. We conclude that the bad webpages that look good are less likely to receive negative evaluation even if a support system reveals their treacherousness. The website appearance is therefore more important than it seems to be.

So far we have been analysing only reliable system feedback. Now we will explore whether the wise crowd can fight off system suggestions that are plainly wrong. Here we must analyse data from the RE condition. The accuracy of the lay evaluations in this case is the lowest. Only 23.4 % of all of them is exactly the same as the corresponding (correct) expert evaluations, After dichotomising the ratings this share reaches 50.88 % and when using the wisdom of the crowd approach it rises by a further 3 % points.

Similarly to the case of completely noncredible content in the NH treatment the best sign of the ongoing fight for correct evaluation is the dispersion of ratings in this condition. The Leik measure of ordinal dispersion in this case, irrespective of the actual website credibility, oscillates around 45 % of the

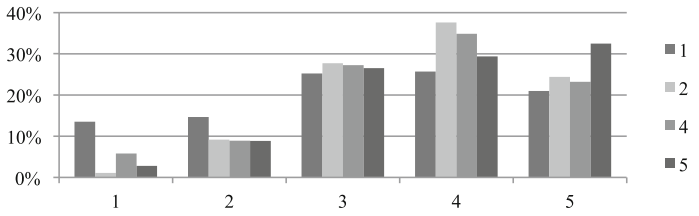

Fig. 3. The distributions of the lay evaluations (horizontal axis) for all the possible values of false suggestions made by the system (colours).

Table 4. The average time in seconds spent on a single evaluation in the RE condition.

	Mean estimation time	Std. Err.	95 % Conf. Interval	
Negative suggestions followed	36.947	2.396	32.215	41.679
Negative suggestion rejected	43.683	1.352	41.012	46.353
Positive suggestion rejected	46.152	3.471	39.296	53.008
Positive suggestion followed	43.341	2.162	39.070	47.612

maximum. Figure 3 depicts the distributions of the lay evaluations for all the possible values of false suggestions made by the system (as explained at the beginning the middle ratings were excluded). It can be seen that not all lay users are keen on following the deceitful advice. However, once again there is a visible asymmetry in how positive and negative information is treated. The probability that deceptive positive advice will be rejected equals 13 % if the rejection is defined as choosing either 1 or 2, and equals 39.8 % if we also include 3 as a negative evaluation. At the same time the probability that deceptive negative advice will be rejected equals over 53.3 %. It looks as though the subjects relied on some decision heuristics that are particularly sensitive to all the positive cues and they disregarded the warning signs.

The disregard for the warning signs, however, is costly as indicated by the time spent on evaluation. Table 4 presents the average time in seconds spent on a single evaluation in the RE condition. The rejection of the system suggestion is always more time consuming than following it. Interestingly this difference is statistically significant only in the case of negative system suggestions. The decision to follow them takes on average about 37 s while when deciding to reject it the subjects need on average about 44 s.

6 Summary

In this paper we presented a study on how the lay users evaluate the credibility of health related content on the web. We investigated the differences between their approach and the approach of medical experts and analysed whether we can increase their accuracy using a simple support system.

The general conclusion is positive. A simple support system based on expert choices does visibly increase the share of accurate evaluations. Even the wisdom of crowds is effective to some extent. Yet in this case it is important to pay attention not only to the central tendency measures for the group but also to the dispersion of ratings that is indicative of an inner struggle for better evaluations in the most problematic cases. These problematic cases can aslo be identified by asking for controversy evaluations.

We learned that lay users exhibit a clear positive evaluation bias that cannot be easily corrected under the wisdom of crowds approach. Moreover, it is fairly resistant to support system suggestions. On one hand, it prevents the lay users from correcting their overly positive evaluations when informed about their inaccuracy. On the other, it leads them to believe the false positive suggestions more keenly than the false negative suggestions.

Furthermore, we saw that lay users and experts both form coherent images of the evaluated web content that have much less in common than expected. It seems that these two groups literally see the evaluated medical websites differently. The lay users seem to perceive their evaluations as based mostly on the expertise of the source. However, when we analysed the predictors for their resilience against (rightfully) negative system suggestions, the perceived appearance of the websites becomes important.

Acknowledgments. This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 690962.

References

1. Fogg, B.J., Tseng, H.: The elements of computer credibility. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 80–87. ACM (1999)
2. Fogg, B.J.: Persuasive technology: using computers to change what we think and do. Morgan Kaufmann Publishers, San Francisco, CA (2003)
3. Jankowski-Lorek, M., Nielek, R., Wierzbicki, A., Zieliński, K.: Predicting controversy of wikipedia articles using the article feedback tool. In: Proceedings of the 2014 International Conference on Social Computing, p. 22. ACM (2014)
4. Jankowski-Lorek, M., Zieliński, K.: Document controversy classification based on the wikipedia category structure. Comput. Sci. **16**(2), 185–198 (2015)
5. Rafalak, M., Abramczuk, K., Wierzbicki, A.: Incredible: is (almost) all web content trustworthy? Analysis of psychological factors related to website credibility evaluation. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 1117–1122. International World Wide Web Conferences Steering Committee (2014)
6. Rafalak, M., Bilski, P., Wierzbicki, A.: Analysis of demographical factors' influence on websites' credibility evaluation. In: Kurosu, M. (ed.) HCI 2014, Part III. LNCS, vol. 8512, pp. 57–68. Springer, Heidelberg (2014)
7. Flanagin, A.J., Metzger, M.J.: The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. New Media Soc. **9**(2), 319–342 (2007)

8. Kakol, M., Jankowski-Lorek, M., Abramczuk, K., Wierzbicki, A., Catasta, M.: On the subjectivity and bias of web content credibility evaluations. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 1131–1136. International World Wide Web Conferences Steering Committee (2013)
9. Kakol, M., Nielek, R.: What affects web credibility perception? An analysis of textual justifications. Comput. Sci. **16**(3), 295–310 (2015)
10. Lucassen, T., Muilwijk, R., Noordzij, M.L., Schraagen, J.M.: Topic familiarity and information skills in online credibility evaluation. J. Am. Soc. Inform. Sci. Technol. **64**(2), 254–264 (2013)
11. Crawford, J.L., Guo, C., Schroeder, J., Arriaga, R.I., Mankoff, J.: Is it a question of trust? How search preferences influence forum use. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp. 118–125. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2014)
12. Frické, M., Fallis, D., Jones, M., Luszko, G.M.: Consumer health information on the internet about carpal tunnel syndrome: indicators of accuracy. Am. J. Med. **118**(2), 168–174 (2005)
13. Griffiths, K.M., Tang, T.T., Hawking, D., Christensen, H.: Automated assessment of the quality of depression websites. J. Med. Internet Res. **7**(5), e59 (2005)
14. Schwarz, J., Morris, M.: Augmenting web pages and search results to support credibility assessment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1245–1254. ACM (2011)
15. Yamamoto, Y., Tanaka, K.: Enhancing credibility judgment of web search results. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1235–1244. ACM (2011)
16. Amin, A., Zhang, J., Cramer, H., Hardman, L., Evers, V.: The effects of source credibility ratings in a cultural heritage information aggregator. In: Proceedings of the 3rd Workshop on Information Credibility on the Web, pp. 35–42. ACM (2009)
17. Leik, R.K.: A measure of ordinal consensus. Pac. Sociol. Rev. **9**(2), 85–90 (1966)
18. Surowiecki, J.: The Wisdom of Crowds. Anchor, Garden City (2005)