RC25419 (IRE1310-013) October 25, 2013
Mathematics

IBM Research Report

Suppressing Deceitful Influence in Cites-or-Writes Social Networks

Adam Hackett[1], Shoukat Ali[2], Stephen Kirkland[3], Massimiliano Meneghin[2]

[1]Department of Mathematics and Statistics
University of Limerick
Limerick, Ireland

[2]IBM Research
Smarter Cities Technology Centre
Mulhuddart
Dublin 15, Ireland

[3]Department of Mathematics
University of Manitoba
Winnipeg, Manitoba
Canada

Research Division
Almaden - Austin - Beijing - Cambridge - Dublin - Haifa - India - T. J. Watson - Tokyo - Zurich

IBM Technical Report: Suppressing deceitful influence in cites-or-writes social networks

Adam Hackett*, Shoukat Ali[†], Stephen Kirkland[‡], and Massimiliano Meneghin[†]
*Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
adam.hackett@ul.ie
[†]Exascale Systems Group, IBM Dublin Research Laboratory, Dublin, Ireland
{shoukat.ali, massimiliano.meneghin}@ie.ibm.com
[‡]Department of Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada
stephen.kirkland@umanitoba.ca

***Abstract*—This paper presents a method for analysing influence in a class of social networks that can be viewed as consisting of messengers and messages. A messenger may either write a message or cite (e.g., re-tweet, repost, grade, or like) a message originally written by another messenger. We term such social networks as *cites-or-writes networks*. The free citability of messages may lead to collusion among a small set of messengers to gain unwarranted influence. We propose a mathematically rigorous method for inferring the *influence* of a messenger and the *impact* of her message. As a key feature, our method does not require any configurable parameters. It only uses a matrix of messenger-writes-message data and a matrix of messenger-cites-message data. Another key feature of our technique is the suppression of collusion; i.e., messengers who deliberately promote each other by preferential citations are automatically suppressed in their influence. We discuss several examples of social networks that follow the cites-or-writes model. We show that our technique is fast enough to do large-scale social network analysis, and present a sample analysis for a large network.**

I. Introduction

The proliferation of online social networking platforms has provided a deluge of data concerning the communication activities of their various users. A large portion of this activity involves the exchange of publicly or selectively visible messages between users. Furthermore, this exchange typically involves user activities that can be identified as belonging to either of two broad categories: writing and citing. For example, Facebook users usually communicate with other users of the site by posting status messages to their personal 'timeline' which can then be 'liked', 'shared', or otherwise commented upon by those with whom the message has been shared. Similarly, on Twitter a user may post a message – a so-called 'tweet' – which may then be 'retweeted' by any other user of the site if it has been made publicly visible. More broadly, these two categories of communication activity can be identified as the cornerstone of a large and growing number of networks ranging from microblogging platforms such as Tumblr, to user-content driven news and entertainment websites such as Reddit. This communication paradigm has even infiltrated traditional news outlets such as broadsheet newspapers, with the online versions of some of the most established of these outlets now containing comment sections in which readers can post messages which can then be recommended by other readers.[1]

However, while the online form of this mode of communication may be a relatively recent phenomenon, generally speaking it is far from novel. The most conspicuous example of an offline social network in which this activity has played a crucial role since its earliest days is that formed by the authors of academic papers and the papers they write or cite.

Mathematically speaking, in data sets where this mode of communication is evident we can identify an instance of a class of bipartite networks in which vertices belong to either of two disjoint sets labelled *messengers* and *messages*, and in which a weighted directed edge, either of type *writes* or *cites*, from a messenger vertex to a message vertex indicates that the messenger in question has either written the message or cited the message. We refer to this class of networks as cites-or-writes networks. For example, in the case of scientific publications we observe a network in which a directed edge from an author to a paper indicates that the author has either written the paper or cited it in one of her own publications. Similarly, if the relevant data is available – namely, messenger-writes-message and messenger-cites-message data – we can identify and configure instances of cites-or-writes networks for each of the aforementioned online social networks.

A fundamental problem for cites-or-writes networks

[1] See, for example, (http://www.nytimes.com).

is the identification of messengers with high influence and messages of high impact. We reserve the term *influence* for messengers and *impact* for messages. In this paper we develop a parameter-free mathematical method that simultaneously quantifies influence and impact. We adopt the view that these two notions are interdependent: a messenger is influential if, on average, that messenger's messages have high impact, and a message has high impact if it is endorsed by a large number of influential messengers. This approach means that our method is resistant to collusive attempts by messengers to gain a position of influence and also allows us to quickly identify phony messages, such as those seen in *astroturf* campaigns [1].

The remainder of this paper is structured as follows. In Sec. II we elaborate further on the problem addressed by our ranking method and discuss some previous mathematical work on related problems. In Sec. III we show how the structure of cites-or-writes networks can be modelled using matrices, provide a detailed description of our ranking method, and comment on its suitability for large scale analysis. In Sec IV, we use a real dataset to illustrate the collusion detection feature of our method.

II. Preliminaries

A. Problem Statement

A large proportion of the communication activities we engage in and the items of information we consume nowadays are centered around web-based user-generated content. Any frequent user of the World Wide Web will no doubt have experienced the phenomenon of information overload, and will agree that determining which items of information are of value to us and which sources of information are worth paying attention to has never seemed more crucial. Similarly, scientists are constantly faced with the problem of determining which papers and journals they should read to obtain high quality information that is relevant to their chosen field of study, and which conferences they should participate in to allow their research to reach the largest possible audience of influential peers.

In the abstract language of cites-or-writes networks, practical problems of this type can be expressed more generally as those of determining which messengers are influential, and are therefore worth paying attention to, and which messages have high impact, and are therefore worth citing.[2] It is clear that not every message a user of a social network may happen to observe can be of equal value to that user. Similarly, not all messengers deserve an equal reputation for producing valuable messages. The notions of messenger influence and message impact have been introduced in order to allow us to quantify the reputation of each messenger and the value of each message and thereby produce a qualitative ranking of both messengers and messages. Our central problem, therefore, is to determine a rigorous mathematical method that will allow us to measure the relative influence of messengers and the relative impact of messages in cites-or-writes networks.

B. Existing Methods

The problem we have posed can be viewed as a generalization of a number of well-established problems from an extensive list of scientific areas of study. We discuss some of that work here.

In the field of scientometrics the problem of determining the reputation of authors and the quality of the papers they produce has been intensively studied for more than fifty years. The earliest attempts to quantify the idea of author reputation simply used the total number of citations received by the papers written by a particular author as a measure of that author's influence [2], [3]. However, it has long been recognized that this approach does not take into account the quality of the citations received. Evidently, a citation received from an acknowledged expert in a particular field should be of greater value to the reputation of the recipient than a citation from a relatively obscure author. This problem is often compounded by the fact that novice authors tend to heavily cite more establish (already highly-cited) authors in order to gain the attention of their peers. The now widely-used measure of author influence known as the 'h-index' – where h is the number of the author's papers that have been cited at least h times each – also suffers from this problem [4].

More recently, various attempts have been made to address this problem by linking author influence to citation quality. To achieve this it is useful to introduce the notion of paper impact. The method known as EigenScore [5] uses ideas borrowed from the PageRank algorithm [6] to provide a ranking of authors and papers. This method provides a quality-based ranking of individual authors;

[2] The generality of cites-or-writes networks means that they can be found in many different contexts. All that we require is data adhering to the form messenger-writes-message and messenger-cites-message. For instance, we can easily view conferences as messengers and the papers published at individual conferences as messages and thereby apply our method to obtain a ranking of influential conferences.

however, it does not provide a ranking of individual papers. A very high impact paper may not be able to derive its high score from the influence of its authors. Eigenscore only gives an average impact score for the papers of each author.

Another recently proposed method, called APrank [7], measures author influence and paper impact simultaneously and does provide individual impact rankings for papers. This method is closely related to our approach, and, of all the ranking methods discussed in this section, is the only one which may be considered a direct competitor to ours. However, we shall demonstrate later in our experimental results that unlike our approach APrank lacks the ability to successfully suppress disingenuous authors (messengers) that may try to gain a position of influence by selectively sharing citations.[3]

Given the generality of our approach, a myriad of other ranking methods including some of those from the fields of recommender systems [8], [9], [10], [11], [12], [13], machine learning [14], reputation management [15], [16], and web search [17], [18], may be seen as addressing similar or related problems to ours. However, all of these techniques depend on initialization inputs from their users in order to provide ranking results. This makes all of these methods susceptible to collusion. As we have stated previously, our method is parameter-free. This feature coupled with the mathematical approach on which our method is based means that our method is resistant to gaming. The generality of our approach also means that it can be usefully applied to the problem of messenger and message ranking in each of these fields.

III. Our Methodology

A. Modeling Cites-or-Writes Networks

Suppose that we have a record of messages that have been written and cited by the users of a social network. Specifically, assume that there are n users, and k messages. We want to simultaneously measure the influence of each user, and the impact of each message. We take the view these two notions are interdependent: a user is influential if, on average, that user's messages have high impact, and a message has high impact if it is cited by a large number of influential users.

With that viewpoint in mind, we set up the following two matrices. For each $i = 1, \dots, n$, let d_i be the number of messages written by user i. We assume here that $d_i \geq 1$ for each i; i.e., we focus only on users who write at least one message, and can therefore be defined as messengers, as clearly those writing no messages can have no influence. We construct the $n \times k$ matrix $A = \left[\begin{array}{c} a_{ij} \end{array} \right]_{i=1,\dots,n,j=1,\dots,k}$, where $a_{ij} = \frac{1}{d_i}$ if messenger i writes message j, and $a_{ij} = 0$ otherwise. Next, we construct the $k \times n$ matrix $B = \left[\begin{array}{c} b_{lm} \end{array} \right]_{l=1,\dots,k,m=1,\dots,n}$, where $b_{lm} = g$, $g \in \mathbb{N}_0$, is the number of times that messenger m cites message l.

Here we take a specific interpretation of the term 'cite': a user is understood to have cited a message only in the case that this user takes some action with the message. In the context of scientific publications this meaning is self-evident. However, it is important to clarify that when speaking of citation activity on online social networks, such as, for example, Twitter what we are referring to are activities such as retweeting, or mentioning. We do not consider the merely passive receipt of a message as citation activity on the part of the receiver. Furthermore, in order to maintain a clear distinction between writing and citing activity, we stipulate that a messenger cannot cite her own message.

B. Our Algorithm

We want to set up two sequences of vectors $u(p) \in \mathbb{R}^n$ and $t(p) \in \mathbb{R}^k$ such that for each $p \in \mathbb{N}$, the vector $u(p)$ approximates the influence of the various messengers, and the vector $t(p)$ approximates the impact of the various messages. Without loss of generality we normalise each $u(p)$ and $t(p)$ so that the entries sum to 1. Start with $u(0) = \frac{1}{n}\mathbf{1}_n$ and $t(0) = \frac{1}{k}\mathbf{1}_k$, where $\mathbf{1}_n$ and $\mathbf{1}_k$ are the all ones vectors of orders n and k, respectively. Now set up the following recursions. For each $p \in \mathbb{N}$, let

$$\tilde{u}(p) = At(p-1); \;\; u(p) = \frac{1}{\mathbf{1}_n^t \tilde{u}(p)} \tilde{u}(p), \qquad (1)$$

and

$$\tilde{t}(p) = Bu(p-1); \;\; t(p) = \frac{1}{\mathbf{1}_k^t \tilde{t}(p)} \tilde{t}(p). \qquad (2)$$

Observe that if $u(p-1)$ and $t(p-1)$ approximate the messenger influences and message impacts, respectively, then at the p–th step of the iteration, the i–th entry of $\tilde{u}(p)$ approximates the average impact of the messages sent by the i–th messenger, and the l–th entry of $\tilde{t}(p)$ approximates the sum of the influences of the messengers that receive message l. In other words, these vectors reflect the viewpoint that we have adopted regarding the interdependence of influence and impact.

[3] We readily acknowledge that gaming the ranking system in order obtain a position of influence is an unlikely activity for scientists to engage in. However, it is undoubtedly the case that this activity does occur in online social networks, see for example [1]. Therefore, we envisage that this aspect of our method will be of relevance primarily to these networks. In this paper, we use scientific citation data in our experiments merely for illustrative purposes, as this is the only data available to us at this time.

From the iterations above, we have, for each $p \geq 2$, that

$$\tilde{u}(p) = At(p-1) = \frac{1}{\mathbf{1}_n^t Bu(p-2)} ABu(p-2), \quad (3)$$

so that

$$u(p) = \frac{1}{\mathbf{1}_n^t ABu(p-2)} ABu(p-2). \quad (4)$$

Similarly, it follows that

$$t(p) = \frac{1}{\mathbf{1}_n^t BAt(p-2)} BAt(p-2). \quad (5)$$

Suppose that the matrix AB is primitive (i.e. some power has all positive entries). Then letting $p \to \infty$, we find that $u(p)$ converges to a dominant right Perron vector $\overline{u}$ of AB (normalised so that its entries sum to 1), while $t(p)$ converges to a dominant right Perron vector $\overline{t}$ of BA (also normalised so that the entries sum to 1). Observe that $\overline{u}$ is a scalar multiple of $A\overline{t}$ and $\overline{t}$ is a scalar multiple of $B\overline{u}$, which is precisely the kind of relationship that we were looking for in adopting our viewpoint on the interdependence between influence and impact.

Further note that the (i, j) entry of the matrix AB is given by $\sum_{m=1}^{k} a_{i,m} b_{m,j}$. It is easy to see that $(AB)_{i,j}$ is the proportion of the messages written by messenger i that are cited by messenger j. In particular, $(AB)_{i,j}$ is positive if and only if there is some m such that $a_{i,m} > 0$ and $b_{m,j} > 0$ – i.e. if and only if at least one message written by messenger i is cited by user j. Similarly, the (p, q) entry of BA is positive if and only if there is at least one user that cites message p and writes message q. Recall that the Perron value ρ of AB is increasing in each entry of AB, and since each entry of AB is bounded above by 1, it follows that $\rho \leq n$, with equality holding if and only if every entry in AB is 1. This means that we might interpret ρ as a measure of the overall intensity of communication between messengers in the network – if ρ is large, then messengers are citing a large proportion of each others' messages, while if ρ is small, then messengers are citing just a small proportion of messages written by other messengers.

C. Running Time

We use an iterative method (the power method [19]) to compute the leading eigenvectors of AB and BA. The method only requires a matrix-vector product instead of a full matrix multiplication. One could use other standard iterative methods like the Lanczos method [20].

IV. Experiments and Results

Without loss of generality, we focus our experiments on cites-or-writes networks obtained from scientific citation data, where the messengers are authors and messages are the papers that are either written or cited. For convenience we refer to the influence and impact scores from our cites-or-writes analysis as *CoWrank*.

We tested our algorithm's collusion detection feature on the same dataset that was used in [7]. It consists of a total of 2012 econophysics papers, written by 1990 authors, taken from 78 scientific journals and arXiv.org, published between April 1995 and September 2010. We refer to this dataset as the *econophysics dataset*.

We asked ourselves the following questions? What if a small subset of authors started preferentially citing each other's messages? As a specific experiment, we ranked the authors in the econophysics dataset using both the APrank method and our method, CoWrank. Next, we selected a set C of three authors randomly from the intersection of the bottom 20% bands of the ranking vectors produced by APrank and CoWrank (referred to as the *selection band*).[4] We then added *collusive edges* such that each author in C cited every other author $\lfloor k\bar{c} \rfloor$ times, where $\bar{c}$ is the average number of citations in the network before collusion, and then ranked the author influence again using both APrank and our method. We repeated this procedure for 30 trials where two trials differed from each other only in the selection of the collusive authors, i.e., the set C was different for a different trial ($|C|$ was always 3). At the end of the experiment we calculated, for each author in C, its average influence scores from APrank and CoWrank over 30 trials. We then determined the average rank of an author in C relative to other authors in the network based on the values of these average scores, and plotted this as *influence* on the y-axis in Figure 1. One can see that APrank and CoWrank behave very differently. With APrank, the authors in C gain more and more influence as the collusive activity increases (i.e., as k increases). However, with CoWrank, the collusive citationing is not rewarded at all; the author influence actually falls somewhat.

The collusive authors in the above experiment were all "weak" to begin with. A natural question is how the influence graph would change if these authors were already mid-level influential or highly influential. Figures 2 and 3 answer this question. Figure 2 is plotted for a selection band from 40% to 60%, i.e., each of the three

[4] In other words, an author is selected if both APrank and CoWrank tell us that at least 80% of the authors in the network should be ranked higher.

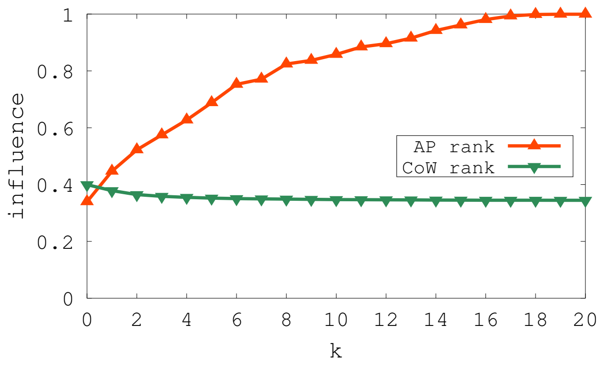

Fig. 1: A what-if scenario created with the econophysics data to show how author influence changes when three originally low-influence authors start citing each other preferentially. Such authors gain influence with APrank, and lose influence with CoWrank.

authors in C has an influence value greater than at least 40% of the other authors in the network and less than at least 60%. Similarly, Figure 3 is plotted for a selection band of 80% to 100%. As can be seen from these figures, CoWrank continues to punish collusive citationing. A more revealing experiment is shown in Figure 4 that explores the effect on influence of a mixed group of weak and strong authors that engage in preferential inter-citations. Three of these six authors are selected randomly from the bottom 20% of the influence spectrum and the remaining three authors are selected randomly from the top 20% of the influence spectrum. The figure shows the interesting result that, up to a certain value of k, CoWrank increases the influence of the weak authors. However the influence of the weak authors decreases when the extent of inter-citations increases past a certain value of k. As before, APrank does not detect collusion and significantly boosts the influence of the colluding authors.

We also determined the smallest group of authors for which excessive inter-citations are not viewed as unfair. Obviously this number would depend on the structure and size of the network. For the econophysics data, this number was 19, i.e., about 10% of the total number of authors. Figure 5 shows the change in influence with increasing value of k, i.e., increasing extent of inter-citations. It can be see that authors build their influence in this scenario.

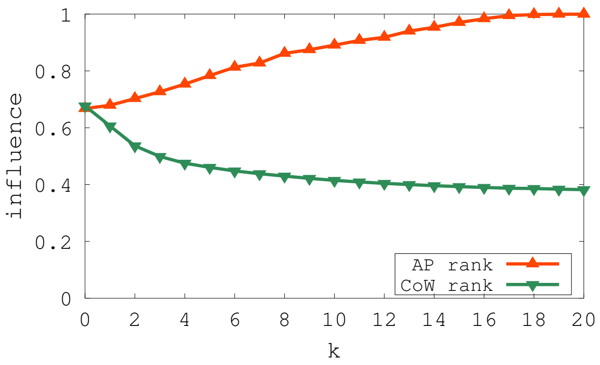

Fig. 2: The what-if scenario of Figure 1, except that the collusive authors have mid-level influence originally.

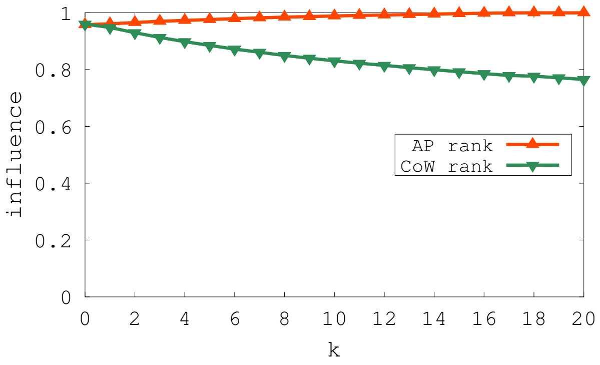

Fig. 3: The what-if scenario of Figure 1, except that the collusive authors are highly influential originally.

V. Conclusions

In this paper we identify a class of bipartite networks that can be used to model a large number of real social networks. Such cites-or-write networks consist of messengers that can not only write messages but also cite (grade, like, or repost) messages. We propose a mathematically rigorous method for inferring the *influence* of a messenger and the *impact* of her message. Our method does not require any configurable parameters and is very efficient at suppressing collusion (a natural possibility granted by the free citability of messages).

We apply our method to a data set consisting of a large subset of authors and their publications in the field

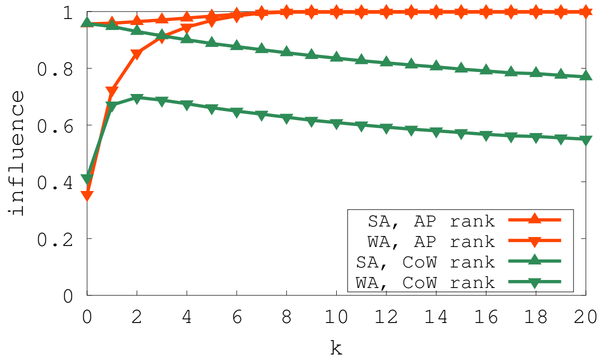

Fig. 4: The what-if scenario of Figure 1, except that (a) there are six authors and (b) three of the authors are from the bottom 20% of the influence spectrum and the remaining three authors are from the top 20% of the influence spectrum. SA: strong author, WA: weak author.

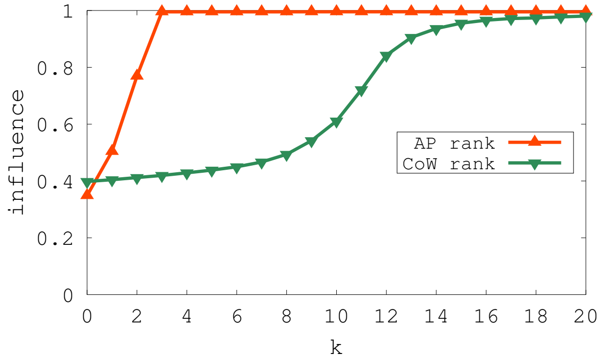

Fig. 5: For the econophysics dataset, 19 was the smallest group size for which excessive internal citations were deemed legitimate.

of econophysics to obtain a ranking of these authors' influence. We show through experimental results how our method allows us to identify and punish collusive behavior. Specifically, we show how a group of authors who share citations preferentially with each other in order to artificially boost each others' author influence and paper impact scores will, in fact, have their respective influence and impact scores reduced relative to the scores of genuine authors and papers in the network under our scheme. Since our method allows a messenger to be a journal as well, this collusion suppression can be used to control the quality of journals indexed by a publisher. Our experiments verify that no exiting method of ranking which is comparable to ours provides this capability.

References

[1] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*. AAAI Press, 2011.

[2] E. Garfield, "Citation indexes for science: A new dimension in documentation through association of ideas," *Science*, vol. 122, no. 3159, 1955.

[3] ——, "Citation frequency as a measure of research activity and performance," *Essays of an Information Scientist*, vol. 1, 1973.

[4] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[5] C. T. Bergstrom, "Eigenfactor: Measuring the value and prestige of scholarly journals," *College & Research Libraries News*, vol. 68, no. 5, 2007.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[7] Y.-B. Zhou, L. Lu, and M. Li, "Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity," *New Journal of Physics*, vol. 14, 2012.

[8] A. Cheng and E. Friedman, "Sybilproof reputation mechanisms," in *Proc. of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems (P2PCON '05)*. ACM, 2005, pp. 128–132.

[9] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the web*, vol. 1, no. 1, 2007.

[10] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: experiments on recommending content from information streams," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, 2010, pp. 1185–1194.

[11] A. Ghosh, S. Kale, and P. McAfee, "Who moderates the moderators?: crowdsourcing abuse detection in user-generated content," in *Proceedings of the 12th ACM conference on Electronic commerce (EC '11)*. ACM, 2011, pp. 167–176.

[12] S. Kywe, E. P. Lim, and F. Zhu, "A survey of recommender systems in twitter," in *Social Informatics*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7710, pp. 420–433.

[13] D. R. Karger, S. Oh, and D. Shah, "Efficient crowdsourcing for multi-class labeling," in *Proc. of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems (SIGMETRICS '13)*. ACM, 2013, pp. 81–92.

[14] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: mapping the spread of astroturf in microblog streams," in *Proc. of the 20th international conference companion on World wide web (WWW '11)*. ACM, 2011, pp. 249–252.

[15] T. Hogg and L. Adamic, "Enhancing reputation mechanisms via online social networks," in *Proc. of the 5th ACM conference on Electronic commerce (EC '04)*. ACM, 2004, pp. 236–237.

[16] S. Kamvar, M. Schlosser, and H. Garcia-Molina, "Eigenrep: reputation management in p2p networks," in *Proc. of the 12th international world wide web conference (WWW '03)*. ACM, 2003.

[17] H. Zhang, A. Goel, R. Govindan, K. Mason, and B. V. Roy, "Making eigenvector-based reputation systems robust to collusion," *Algorithms and models for the web-graph*, vol. 3243, 2004.

[18] K. Mason, “Detecting colluders in pagerank: finding slow mixing states in a markov chain,” Ph.D. dissertation, 2005.

[19] R. von Mises and H. Pollaczek-Geiringer, “Praktische verfahren der gleichungsauflösung,” *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 9, 1929.

[20] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations: Volume 1, Theory*, ser. SIAM classics in applied mathematics. SIAM, 2002, no. 41.