Journal of Informetrics 15 (2021) 101127

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

SCiMet: Stable, sCalable and reliable Metric-based framework for quality assessment in collaborative content generation systems

Mohammad Allahbakhsh[a,b,c,*], Haleh Amintoosi[b], Behshid Behkamal[b], Amin Beheshti[c], Elisa Bertino[d]

[a] The University of Zabol, Iran
[b] Ferdowsi University of Mashhad, Iran
[c] Macquarie University, NSW 2109, Australia
[d] Purdue University, Lafayette, IN 47907, United States

ARTICLE INFO

Article history:
Received 9 June 2020
Received in revised form 25 October 2020
Accepted 17 December 2020

Keywords:
Quality assessment
Quality metric
Scientometrics
Collaborative content
Attack-resilient

ABSTRACT

In collaborative content generation (CCG), such as publishing scientific articles, a group of contributors collaboratively generates artifacts available through a venue. The main concern in such systems is the quality. A remarkable range of research considers quality metrics partially when dealing with the quality of artifacts, contributors, and venues. However, such approaches have several drawbacks. One of the most notable ones is that they are not comprehensive in terms of the metrics to evaluate all entities, including artifacts, contributors, and venues. Also, they are vulnerable to potential attacks.

In this paper, we propose a novel iterative definition in which the quality of artifacts, collaborators, and venues are defined interconnectedly. In our framework, the quality of an artifact is defined based on the quality of its contributors, venue, references, and citations. The quality of a contributor is defined based on the quality of his artifacts, collaborators, and the venues. Quality of a venue is defined based on both quality of artifacts and contributors. We propose a data model, formulations, and an algorithm for the proposed approach. We also compare the robustness of our approach against malicious manipulations with two well-known related approaches. The comparison results show the superiority of our method over other related approaches.

1. Introduction

1.1. Background

One of the main results of Web 2.0 technologies is enabling people to collaborate in content generation. Wiki websites, open-source software, SVN repositories and tools such as Google Docs[1] and Dropbox Paper[2] are examples of efforts and

* Corresponding author.
E-mail addresses: allahbakhsh@uoz.ac.ir (M. Allahbakhsh), amintoosi@um.ac.ir (H. Amintoosi), behkamal@um.ac.ir (B. Behkamal), amin.beheshti@mq.edu.au (A. Beheshti), bertino@purdue.edu (E. Bertino).

[1] https://docs.google.com/
[2] https://paper.dropbox.com/

https://doi.org/10.1016/j.joi.2020.101127
1751-1577/

platforms supporting the collaborative generation of contents. These systems are introduced earlier in the literature and are called collaborative systems and collective intelligence (Doan, Ramakrishnan, & Halevy, 2011). In this paper, we refer to these systems as Collaborative Content Generation (CCG) systems. In a Collaborative Content Generation system, a group of people, called *contributors*, contribute to the process of creating a content, also called an *artifact*. This artifact is available to the community via a *venue*.

Publishing scientific articles is another example of collaborative content generation, which has been since centuries. In such a system, a group of authors collaboratively write a scientific article and publish it in a venue such as a journal. To be more clear, this application is the most suitable application for the work we propose in this article. A scientific article is an artifact that a group of authors generate it. Throughout the paper, we use the terms artifact, contributor and venue to keep the model general, but they can be used/replaced with article, author and journal/conference, respectively. To make it more clear, we investigate it more, as the main motivating scenario of the proposed model.

Assessing and controlling the quality of the human-generated artifacts, which is mainly considered in scientometrics and some other more generic quality assessment frameworks, is a challenging task that has been investigated by a large number of research efforts (Allahbakhsh et al., 2013; Bollen, Van de Sompel, Hagberg, & Chute, 2009; Braithwaite et al., 2019; Daniel, Kucherbaev, Cappiello, Benatallah, & Allahbakhsh, 2018). From this point of view, the quality of an artifact depends on a broad range of factors, mainly the quality of the contributors (trust, expertise, credentials, etc.) and the quality attributes of the artifact itself (such as accuracy, completeness, consistency, etc.) (Allahbakhsh et al., 2013; Daniel et al., 2018; Salk, Sturn, See, Fritz, & Perger, 2016; Thuan, Antunes, & Johnstone, 2016). Several approaches have been proposed, such as expert review, majority consensus, contributor evaluation and ground truth, that are widely used for assessing the quality of artifacts. For instance, movie rating mechanisms used in web sites such as IMDb[3] and Rotten Tomatoes[4] are examples of majority consensus, and peer-reviewing of research articles is a kind of expert review quality assessment technique.

The other important factor in assessing the quality of an artifact is the quality of the venue in which the artifact is published. An artifact published in a high-quality venue is believed to have a higher quality than an artifact published in a low reputable venue. This is due to the quality control mechanisms the high quality venues put in place to make sure that they only publish high quality artifacts. For instance, research journals use expert review techniques, a.k.a peer-reviewing, to assess the quality of submitted manuscripts. The impact of the venue on the quality of artifacts has also been studied in past research (Braithwaite et al., 2019).

Furthermore, when it comes to collaboration, there are some other significant aspects that should be taken into account (Braithwaite et al., 2019; Doan et al., 2011), some of which are: (i) *Share*: Have all collaborators equally contributed to a given artifact? or are the contribution uneven?; (ii) *Quality*: Is collaborator quality the same for all collaborators? and (iii) *Role*: Have collaborators the same role or their roles are different?

Previous works have addressed quality for certain categories of *CCG* activities, namely writing books, authoring research articles or contributing to a Wikipedia[5] . A remarkable range of research considers these metrics partially when dealing with the quality of artifacts, contributors and venues (Bollen et al., 2009; Daniel et al., 2018; Haustein, Costas, & Larivière, 2015; Matei, Jabal, & Bertino, 2018; Teplitskiy, Duede, Menietti, & Lakhani, 2020). In some models, the reputation of the contributors as well as their expertise and experiences are taken into account to assess the credibility of contributors. Also, in some research articles, the share and roles of contributors have been analyzed (Braithwaite et al., 2019).

Finally, the generated contents available through a venue take different levels of users' attention reflected in the number of references to an article, number of likes/thumb ups of an online post, number of re-shares of online content, and so on. The level of attention is generally considered as an indicator of quality or importance of an artifact (Bollen et al., 2009; Braithwaite et al., 2019).

1.2. Research problem

Regarding the above-mentioned quality metrics, there is a plethora of research proposed for quality assessment in CCG systems (Bollen et al., 2009; Daniel et al., 2018; Haustein et al., 2015; Matei et al., 2018; Teplitskiy et al., 2020). However these approaches have major drawbacks. One of the drawbacks is that most existing quality assessment frameworks are incomplete. They either produce quality metrics for one or two of the three components of CCG systems which are contributors, artifacts and venues; or they ignore some quality factors and only use a limited number of them such as community attention (e.g., citations) to compute quality metrics.

Another drawback is related to the way the quality is defined in the existing approaches. Such approaches define and compute quality in a one-way manner, and ignore the correlation between the quality of the CCG components. For instance, in existing approaches, the quality of a paper is computed based on the number of its citations. Then, the quality of people and venues is computed based on the quality of papers. The reverse relation which is ignored is that a paper written by a high profile author is probably a high-quality one, compared with a paper written by a low profile author. This can be also

[3] https://www.imdb.com/
[4] https://www.rottentomatoes.com/
[5] https://www.wikipedia.org/

true for the venues. In other words, being published in a high-ranked venue can be considered as a quality indicator. This correlation between qualities of components is ignored in the existing literature.

The last drawback is that the one-way definition of quality metrics, which are based on a small amount of data, makes those approaches vulnerable to manipulation. Defining weak quality metrics makes it easier to manipulate them individually, or by collusion. A researcher can easily boost his/her profile by self-citations, or a group of researchers can easily promote their works by citing each others' works.

1.3. Contributions

In this paper, we propose a quality assessment framework for CCG systems to address those drawbacks. The proposed method is based on the idea that the quality of artifacts, contributors, and venues are correlated. To be specific, we define and compute quality metrics in an iterative manner. In our proposed method, the quality of an artifact is related to the quality of its venue and its contributors; the quality of a contributor is related to the quality of his/her contributions, quality of his/her collaborators, and quality of the venues that have published his/her artifacts; and finally, the quality of a venue is related to the quality of the artifacts published in it and the quality of the contributors to these artifacts. This iterative definition of quality not only makes quality metrics robust against manipulation, but these metrics are more meaningful and generic as they reflect all aspects of quality of artifacts, venues, and contributors.

We also propose a comprehensive framework in which three quality metrics are proposed and computed corresponding to contributors, artifacts, and venues.We propose a graph data model for better representation of these correlations. We also use this data model to simplify the understanding of the mathematical formulations proposed for the computation of quality metrics. In summary, our main contributions are as follows:

- We propose a graph data model for representing the concepts of artifacts, contributors, and venues, and the relationships between them in CCG systems.
- Based on our graph data model, we propose a novel iterative definition for quality of artifacts, contributors, and venues, identify different factors affecting quality in such an iterative manner, and propose mathematical well-formed formulations for computing quality metrics.
- We evaluate our proposed method both theoretically and empirically. In the theoretical evaluation, we show that the proposed metrics are meaningful, well-formed, and accurate. In the empirical part, we assess the robustness of the model against three organized attacks and compare its performance with two popular related approaches (Bergstrom, 2007; Hirsch, 2005). The empirical evaluations show the superiority of our proposed model.

The remainder is organized as follows. In Section 2, we discuss an example scenario for the proposed framework. In Section 3, we introduce the data model. In Section 4, we introduce our proposed framework. We evaluate the performance of our framework in Section 5, discuss related work in Section 6, and finally conclude in Section 7.

2. Motivating scenario

Publishing a scientific article is a good example of collaborative content generation. A group of scholars collaborates in solving a specific research problem and usually publish the outcomes of their research as a scientific article. The article is published in a venue, i.e., a journal or a conference proceeding. Fig. 1 shows a sample representing three papers, authored by four people, published in three venues.

The quality of the articles is an important concern, which is formally assessed through peer-review. The ultimate result of the peer-review process is either accept or reject, and there is no difference between a paper with a borderline acceptance and a paper with strong acceptance. So, there should be other metrics in place to find high quality articles.

Currently, the number of citations of a paper, i.e., the number of papers that refer to an article is the main quality metric. A paper with a high number of references is considered high quality; those papers with high citations then enhance the visibility of their authors and the reputation of their publication venues. It is thus clear that manipulating the number of paper citations can directly impact the visibility of its authors and reputation of its venue. This has been the main motivation for self-citations, citing each others' papers, and other forms of misbehaviour.

In general, any scenario in which groups of people collaborate in generating contents and other people endorse them, by like or dislike, thumb-up or thumb-down, citations or any other form of direct or indirect referencing is an example scenario for the model proposed in this paper.

In this paper, we use the article publication scenario to show the main essence of the proposed model and the dynamics behind its definitions and formulations. In Section 5, we use this scenario for the experimental evaluation.

3. Data model and abstractions

Assume that in a CCG system, a set of n_C collaborators, denoted by $C = \{c_i | 1 \leq i \leq n_C\}$, have collaboratively generated a set of n_A artifacts, denoted by $A = \{a_j | 1 \leq j \leq n_A\}$. These artifacts are available/published in a set of n_V venues, denoted by $V = \{v_i | 1 \leq k \leq n_V\}$. We represent such system via a directed graph $G = (V, E)$ called CCG Graph, in which V represents the

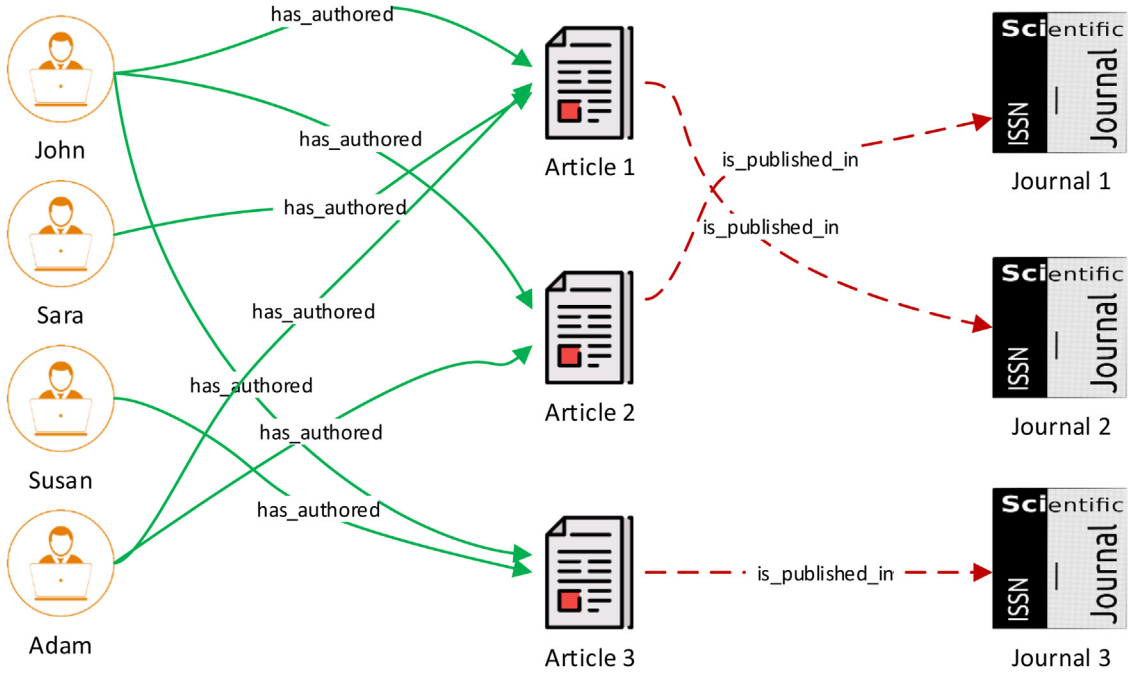


Fig. 1. Motivating Scenario: Application of the proposed model in the research publishing area.

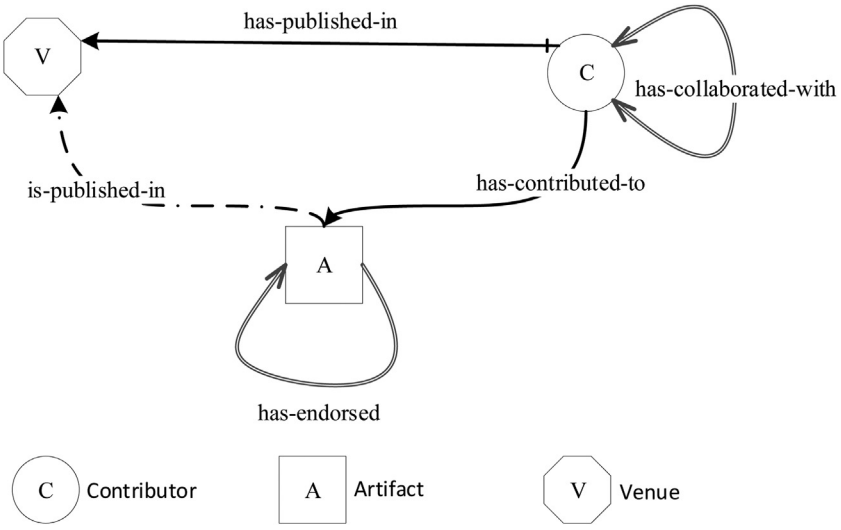


Fig. 2. CCG graph data model.

set of graph nodes and E represents the set of relationships between the entities. The data model is illustrated in Fig. 2. In what follows, we introduce entities and relationships in a CCG graph.

3.1. Entities

3.1.1. Artifact

An artifact is a content generated collaboratively. Each artifact a is identified by a unique Id, and has an associated quality score, denoted by R_a, which is the rating score of a. Moreover, an artifact might also have other application-specific attributes, such as name, title, URI, etc. In our motivating scenario, the artifacts are the research papers.

3.1.2. Contributor

A contributor is a human or a software agent who contributes to the process of generating content. Each contributor c is identified by a unique Id, and has an associated quality metric denoted by T_c, which is the trust score of c. A contributor might

also have other application-dependent attributes, such as name, address, etc. In our motivating scenario, the contributors are the authors of the papers.

3.1.3. Venue

A venue is a physical or virtual place at which the generated contents are available. Each venue v, identified by a unique *Id*, has an associated quality metric called quality impact, denoted by I_v. A venue might have some other application-dependent attributes, such as name, address, etc. In our motivating scenario, the venues are the journals/conferences where papers are published/presented.

3.2. Relationships

Five types of relationships exist in a CCG graph: *has_contributed_to*, *has_collaborated_with*, *has_published_in*, *is_published_in* and *has_endorsed*. These relationships are classified into basic and inferred categories. The basic relationships independently exist between the entities, while the inferred ones are derived based on the basic relationships. *has_contributed_to*, *is_published_in* and *has_endorsed* are basic relationships, and *has_collaborated_with* and *has_published_in* are inferred from the three basic relationships.

3.2.1. has_contributed_to relationship

This is a basic relationship established between a contributor c and an artifact a when c has contributed to the process of generating a, denoted by $c \rightarrow a$. For instance, when an author contributes to a paper, a *has_contributed_to* relationship is established between the author and the paper.

3.2.2. has_collaborated_with relationship

When a contributor c_i has collaborated with another contributor say c_j in generating artifact a, the *has_collaborated_with* relationship is established between them and is denoted by $c_i \overset{a}{\rightleftharpoons} c_j$. The has_collaborated_with relation is an inferred relationship which is established based on has_contributed_to relationships. More precisely,

$$if(c_i \rightarrow a \ \ and \ \ c_j \rightarrow a) then \ \ c_i \overset{a}{\rightleftharpoons} c_j$$

For example, John and Sara, in Fig. 1, both have contributed to the $Article_1$. So, there will be a *has_collaborated_with* relationship between them.

3.2.3. is_published_in relationship

When an artifact a is accessible via a venue v, it means that the artifact is published in v. To reflect this, the *is_published_in* relationship is established between a and v and denoted by $a - \rightarrow v$. Referred to the motivating scenario, $Article_1 - \rightarrow Journal_2$.

3.2.4. has_published_in

When a contributor c contributes to the artifact a that is published in the venue v, a *has_published_in* relationship is established between a and v and denoted by $a \overset{a}{\mapsto} v$. In fact, a has_published_in relationship is an inferred relationship which is derived from a pair of has_contributed_to and is_published_in relationships. In other words:

$$if(c \rightarrow a and a - \rightarrow v) then c \overset{a}{\mapsto} v$$

Referred to Fig. 1, since the $Article_1$ which is coauthored by $Author_1$ is published in $Journal_2$, we establish a *has_published_in* relationship between $Author_1$ and $Journal_2$, i.e., $Author_1 \overset{Article_1}{\longmapsto} Journal_2$.

3.2.5. has_endorsed relationship

When an artifact, say a_i, contains a reference to another artifact a_j, we establish a *has_endorsed* relationship between them and denote it by $a_i \Rightarrow a_j$.

4. Proposed approach

In this section, we first explain the intuition behind our proposed model, its main architecture, and quality metrics. Then, we propose the formulations and the way we calculate quality metrics. Finally, we propose an iterative algorithm that calculates the quality of contributors, artifacts and venues based on the proposed formulations.

4.1. Overview

We assume that in a CCG system, there is no ground truth to be used for quality assessment. In other words, the quality of entities in such a system depends on the dynamics of the system itself, and no other parameters are involved from outside

of the system. In such a system, we propose a model to calculate quality metrics for artifacts, contributors and venues. To do so, we first introduce the following definitions as the core intuition behind our proposed model.

Quality of a human generated artifact, for example a scientific article, has always been one of the main concerns in human-centric systems (Allahbakhsh et al., 2013; Bollen et al., 2009; Braithwaite et al., 2019; Daniel et al., 2018). It depends on many factors such as the quality of its content, contributing people, venue, etc. Relying on just one or two of these factors, while popular, cannot represent the true quality of the artifact. For instance, considering the quality of the journal in which a paper is published, can not truly represent the quality of that paper. Specifically, as recommended by DORA (Declaration on Research Assessment)[6], journal-based metrics, such as journal impact factors, can not be used as a surrogate measure of the quality of individual research articles (Cagan, 2013; Schmid, 2017).

In order to adopt the DORA declaration and provide a more comprehensive definition, we define a high quality artifact as follows:

Definition 1. A high quality artifact is an artifact that:

- has a high quality content, and
- is generated by high quality contributors, and
- is endorsed by high quality artifacts, and
- endorses high quality artifacts, and
- is published in a high quality venue.

It is important to notice that there might be special situations, in which a large number of artifacts refer to an artifact not because of its quality, but to report or fix its problems. Even such an extreme situation, in contrast with the existing approaches, will not have a significant impact on the quality metric of an artifact, since it is used in combination with four other parameters.

We define a high quality contributor as follows:

Definition 2. A high quality contributor is a contributor who:

- has contributed to high-quality artifacts, and
- has published in high-quality venues, and
- has collaborated with high-quality contributors.

This implies that, quality of a contributor, depends on the quality of the contributions to which he/she has contributed to, the quality of the venues that he/she has published artifacts in, and the quality of the people with whom he/she has collaborated.

Finally, we define a high quality venue as follows:

Definition 3. A high quality venue is a venue in which:

- high quality contributors publish their artifacts, and
- high quality artifacts are published.

This definition implies that quality of a venue can be computed based on the quality of the papers that have been published in the venue, as well as, the quality of the people who choose this venue to publish their generated contents.

Looking at Definitions 1, 2 and 3, it is clear that in our proposed model, quality of artifacts, contributors and venues are interrelated, and the definition of quality is iterative. Our iterative model is inspired by, but is more comprehensive than the way Google PageRank establishes the quality of web pages (Page, Brin, Motwani, & Winograd, 1999). An important page is a page that is linked by other important pages, which is clearly an iterative definition. Based on the proposed model, we have three main interrelated quality metrics that should be computed iteratively: the quality of artifact a, denoted b by R_a, quality of contributor cm denoted by T_c, and quality of venue v, denoted by I_v.

4.2. Quality of artifact

As stated in Definition 1, quality of an artifact depends on: (i) quality of the content, (ii) quality of the contributors, (iii) quality of endorsing artifacts, (iv) quality of endorsed artifacts, and (v) quality of the venue. In what follows, we explain how we compute these parameters when assessing quality of artefact.

[6] https://sfdora.org/

4.2.1. Gain of artifact from its content

One of the main parameters that contributes to the quality of an article is the quality of its content. Let's denote the gain of the artifact a from the quality of its content by R_a^n. We assume that R_a^n is a real number in the range [0,1], where 0 means lack of quality and 1 means full quality.

The type of the contents of an article can vary in a broad range from a simple text, to graphics, audio and video. Quality factors for each of these content types are different. While, for example, for a text, accuracy, completeness, and consistency might be the most important factors, the parameters that represent the quality of a graphical content might be completely different. Depending on the content type, there are tools that can automatically examine the quality of the content and compute R_a^n, such as text, audio, image and video processors.

There are also cases in which quality of the content does not necessarily reflect the quality of the artifact. For instance, a paper that is well-written, in terms of quality of writing, is not necessarily a high quality paper, in terms of research contributions. Assessing quality of the content, in these cases, requires specific knowledge and expertise, and should be done by a human expert.

No matter, computed automatically or manually, an artifact has a gain from the quality of its content, i.e.,

$$R_a^n = \begin{cases} expertopinion & ifassessedmanually \\ machineoutput & ifassessedautomatically \end{cases} \tag{1}$$

4.2.2. Gain of artifact from its contributors

Assume that a set of one or more contributors have contributed to the artifact a, denoted by C_a. In some cases, one of the contributors might be considered different from the others. For example, the corresponding or the first author of a paper is assumed to have more important role than others. Let's denote that special contributor by C_a^f.

In our model, the gain of an artifact from the quality of C_a^f is different from its gain from other contributors, applied by their corresponding weights. Let's use w_f and w_o to denote the weight of C_a^f and weight of others, correspondingly. We denote the gain of the artifact a from the quality of its contributors by R_a^c, and calculate it as follows:

$$R_a^c = \frac{w_f T_f + \sum_{o:o\to a} w_o T_o}{w_f + (|C_a| - 1)w_o} \tag{2}$$

R_a^c is a weighted average of the quality. If $C_a^f = C_a^f$, the importance of all contributors is considered the same.

4.2.3. Gain of artifact from its venue

We assume that an artifact is published only in one venue. So, the gain of the artifact comes from just one venue, denoted by R_a^v, and calculated as follows:

$$R_a^v = I_v\,, \quad v : a \dashrightarrow v \tag{3}$$

4.2.4. Gain of artifact from its endorsing artifacts

Endorsing artifacts are the artifacts that have endorsed the artifact a. The quality and the number of endorsing artifacts are parameters that show the quality of an artifact. An artifact with a great number of high-quality endorsing artifacts is deemed to be a high quality artifact. Assume that a has n_g endorsing artifacts. The gain of artefact a for its endorsing artifacts, denoted by R_a^g, is calculated as follows:

$$R_a^g = \frac{\sum_{g:g\Rightarrow a} R_g}{n_g} \tag{4}$$

4.2.5. Gain of artifact from its endorsed artifacts

Endorsed artifacts of a are the artifacts that are endorsed by a. This endorsement means that the content of a somehow is related to the contents of the endorsed artifacts. Hence, if the endorsed artifacts have high quality ranks, this artifact is also of high quality. Assume that a has endorsed n_d other artifacts. The gain of a from its endorsed documents, denoted by R_a^d, is calculated as follows:

$$R_a^d = \frac{\sum_{d:a\Rightarrow d} R_d}{n_d} \tag{5}$$

4.2.6. Aggregating gains of artifact

As modelled by Eqs. (1), (2), (3), (4) and (5), the quality of the artifact a is composed of five elements: R_a^n, R_a^c, R_a^v, R_a^g and R_a^d. In different application domains, each of these elements might have a different level of importance. In some applications, the content might be the most important one, while in some others venue, contributors, or other elements. To represent different levels of importance, we assume that the importance of these elements is reflected in their weights, which are w_a^n,

w_a^c, w_a^v, w_a^g and w_a^d, for R_a^c, R_a^v, R_a^g and R_a^d, correspondingly. Based on these weights, we calculate the quality level of the artifact a as follows:

$$R_a = w_a^n \times R_a^n + w_a^c \times R_a^c + w_a^v \times R_a^v + w_a^g \times R_a^g + w_a^d \times R_a^d \tag{6}$$

In Eq. (6), weighs are positive numbers in the range [0,1] and selected so that $w_a^n + w_a^c + w_a^v + w_a^g + w_a^d = 1$.

4.3. Quality of contributor

Looking back at the Definition 2, quality of the contributor c, denoted by T_c, depends on: (i) quality of artifacts, (ii) quality venues and (iii) quality of his collaborators. A contributor receives gain from each of these parameters. Let's T_c^a be the gain of the contributor form his contribution to artifacts, T_c^v be his/her gain from publishing in quality venues, and $T_c^{\hat{C}}$ be his/her gain from collaborating with other contributors. In what follows, we explain how to compute these gains.

4.3.1. Gain of contributor from artifacts

The gain of c from all the artifacts to which he/she has contributed is denoted by T_c^a, and is calculated as follows:

$$T_c^a = \frac{\sum_{a:c\to a} \omega(c,a) \times R_a}{\sum_{a:c\to a} \omega(c,a)} \tag{7}$$

In Eq. (7), $\omega(c,a)$ is a weight which specifies the impact of an artifact on T_c^a. Based on what we explained in Section 4.2.2, $\omega(c,a)$ is computed as follows:

$$\omega(c,a) = \begin{cases} w_f & \text{if } c \text{ is the creator/first author of } a \\ w_o & \text{otherwise} \end{cases} \tag{8}$$

4.3.2. Gain of contributor from venues

The gain of contributor c for publishing articles in n_{cv} venues is denoted by T_c^v and calculated as follows:

$$T_c^v = \frac{\sum_{v:c\longmapsto v} I_v}{n_{cv}} \tag{9}$$

Note that if a contributor has published several artifacts in one venue, each of them will be considered separately.

4.3.3. Gain of contributor from collaborations

Another effective parameter is the quality of collaborators. The gain of the collaborator c for his/her collaborations with others is denoted by $T_c^{\hat{C}}$, and is computed as follows:

$$T_c^{\hat{C}} = \frac{\sum_{c':c\overset{a}{\rightleftharpoons}c'} \omega(c',a) \times T_{c'}}{\sum_{c':c\overset{a}{\rightleftharpoons}c'} \omega(c',a)} \tag{10}$$

In Eq. (10), $\omega(c',a)$ is a weight which specifies the impact of $T_{c'}$ on $T_c^{\hat{C}}$ and is computed based on Eq. (8).

4.3.4. Aggregating gains of contributor

We use the three gains obtained based on Eqs. (7), (9) and (10) to compute the quality of a contributor, T_c, as follows:

$$T_c = w_c^a \times T_c^a + w_c^v \times T_c^v + w_c^{\hat{C}} \times T_c^{\hat{C}} \tag{11}$$

In Eq. (11), w_c^a, w_c^v and $w_c^{\hat{C}}$ are the weights for T_c^a, T_c^v and $T_c^{\hat{C}}$, correspondingly Wights are positive numbers in the range [0,1] and selected so that $w_c^a + w_c^v + w_c^{\hat{C}} = 1$.

4.4. Quality of venue

With reference to the Definition 3, the quality of a venue, denoted by I_v, depends on: (i) gain of the venue from the quality of the contributors who publish in it, denoted by I_v^c, and (ii) gain of the venue from the quality of the artifacts published in it, denoted by I_v^a. In what follows, we explain how we compute the quality of a venue.

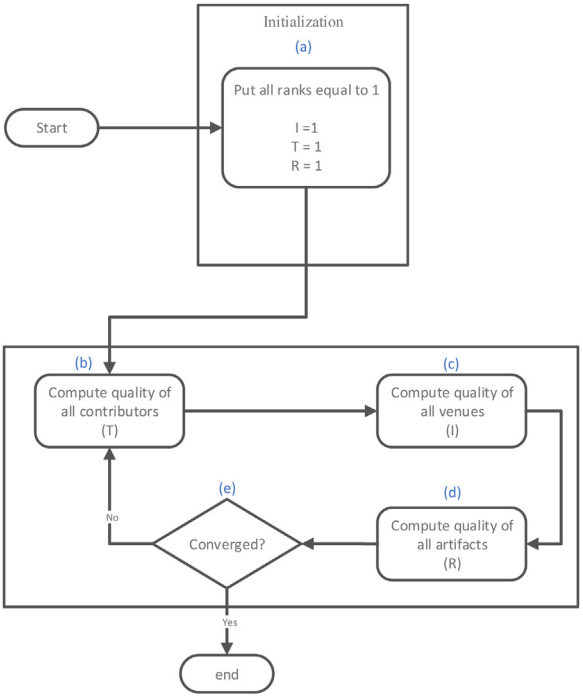


Fig. 3. Iterative Quality assessment model.

4.4.1. Gain of venue from contributors

Assume that n_v^c is the number of contributors who have published in venue v. Note that a contributor with more than one artifact in v will be counted more than once in n_v^c. We denote the gain of v from the quality of the contributors published in v by I_v^c and calculate it as follows:

$$I_v^c = \frac{\sum_{c\,:\,c \xrightarrow{a} v} \omega(c,a) \times T_c}{\sum_{c\,:\,c \xrightarrow{a} v} \omega(c,a)} \tag{12}$$

In Eq. (12), $\omega(c,a)$ is a weight which specifies the impact of T_c on I_v^c and is computed based on Eq. (8).

4.4.2. Gain of venue from artifacts

Assume that there are n_v^a artifacts published in the venue v. We denote the gain of v from these artifacts by I_v^a, and compute it as follows:

$$I_v^a = \frac{\sum_{a\,:\,a \longrightarrow v} R_a}{n_v^a} \tag{13}$$

4.4.3. Aggregating gains of venue

From Eqs. (12) and (13), we have two parameters that to be combined to compute the quality of a venue. These parameters, might have different levels of importance in different application domains. So, we define w_v^a and w_v^c as the weight of I_v^a and I_v^c, respectively, and use them to compute I_v as follows:

$$I_v = w_v^c \times I_v^c + w_v^a \times I_v^a \tag{14}$$

In Eq. (14), weighs are positive numbers in the range [0,1] and selected so that $w_v^c + w_v^a = 1$.

4.5. Quality assessment algorithm

We explained earlier that the nature of quality in our model is iterative. It means that the quality of artifacts, contributors and venues are interconnected. Iterative techniques are widely used in such situations (Allahbakhsh & Ignatovic, 2015; de Kerchove & Van Dooren, 2010; Laureti, Moret, Zhang, & Yu, 2006; Page et al., 1999). Inspired from those ranks and considering the promising performance of iterative techniques, we also propose an iterative approach for computing quality metrics.

The overall architecture of the proposed method is shown in Fig. 3. It starts with the initialization of the quality metrics (see Fig. 3(a)). The initialization is necessary due to the interdependency of the quality metrics, but the initial values do not have any impact on the final result. They might just impact the number of iterations before reaching convergence point. In

the next three steps, that are (Fig. 3(a), (c) and (d)), quality metrics for contributors, venues and artifacts are calculated. In Fig. 3(e) the convergence of the iterative algorithm is checked, which is reached when the values of R_a in two consecutive iterations are close enough to say they have not changed. This check is done using the root mean square error (RMSE). Let $R_a^{(p)}$ be the set of values of R in the p^{th} iteration and $R_a^{(p+1)}$ be the set of values of R in the $(p+1)^{th}$ iteration. The RMSE is the root mean square of $R_a^{(p+1)}$ and $R_a^{(p)}$. If the RMSE is larger than a very small threshold, the iteration will continue, otherwise the model has converged, and iteration stops.

When the algorithm is converged and finished its computations, the values in the three vectors R, T and I are finalized. In other words, upon the convergence of the algorithm, each artifact, contributor and venue will have a corresponding quality score, stored in R, T and I, respectively.

5. Experimentation and evaluation

In this section, we evaluate the accuracy and robustness of our proposed method CCG. We use a theoretical validation approach to assess the accuracy and validity of the proposed quality metrics. Then we report results from experiments that evaluate the robustness of the metrics. Finally, we discuss the results of the experiments.

5.1. Theoretical validation

Since we have no gold standards to evaluate the proposed metrics, it is crucial to show that the defined metrics are well-formed, generic (i.e., they are not defined for a particular application) and rigorous (i.e., defined based on precise mathematical bases). To do so, the metrics should be examined using specific measurement concepts (Briand, Morasca, & Basili, 1996). In other words, metrics should be theoretically analyzed within the framework of measurement theory. There are two main approaches for the theoretical validation of metrics: (i) frameworks that are based on measurement theory principles and define a set of mandatory properties for any type of metric to be considered as an acceptable measurement-theoretic metric., such as the DISTANCE framework (Poels & Dedene, 2000), and (ii) generic frameworks based on the desirable properties of the numerical relational system, such as the property-based measurement framework (Briand et al., 1996). In this paper, we use both to examine the properties of our metrics.

The DISTANCE framework proposes a set of mandatory properties, including identity, non-negativity, symmetry, and triangular inequality that need to be satisfied by any kind of metric. The property-based measurement framework proposes a set of desirable properties including size, length, complexity, coupling, and cohesion. Size is used when metrics are defined based on numeration, e.g. LOC (Lines of Code). Length measures are defined based on the length concept. Based on the definitions of the metrics presented in Sections 4.2 to 4.4, it is obvious that our metrics are neither of the size type, nor the length type. The last two metric types are cohesion and coupling which are meaningful only with reference to modular systems and they are not applicable to our domain. Thus, all metrics we propose here are of the complexity type. A complexity metric type is characterized by five desirable properties, namely, *non-negativity*, *null value*, *symmetry*, *additivity*, and *monotonicity*.

Before investigating these properties, we redefine our system based on the concepts presented in the aforementioned theoretical frameworks. In the property-based measurement framework, system S is represented as a pair $< E, R >$, where E is a set of elements of S, and R is a binary relation on $E(R \subseteq E \times E)$ that represents the relationships between the elements of S. Based on our data model (see Fig. 2), our system is a CCG Graph represented as a pair $< V, E >$, where V is the set of graph nodes and E is the set of a binary relation on $V(E \subseteq V \times V)$. There are three entity types, i.e., artifact, contributor and venue that have five types of relationships including has-contributed-to (between a contributor and an artifact), has-collaborated-with (between two contributors), has-published-in (between a contributor and a venue), is-published-in (between an artifact and a venue), and has-endorsed (between two artifacts). Here, we present the definition of all mandatory and desirable properties proposed in both frameworks in Poels and Dedene (2000) and Briand et al. (1996) and show how our metrics satisfy them.

5.1.1. Property 1. Non-negativity

The complexity of a system $S =< E, R >$ is non-negative, i.e., Complexity $(S) \geq 0$. Based on the metrics formulations presented in Sections 4.2–4.4, we expect all metrics values to be positive or zero (non-negative).

5.1.2. Property 2. Symmetry

The complexity of a system $S =< E, R >$ does not depend on the convention chosen to represent the relationships between its elements. The complexity measure should be insensitive to representation conventions with respect to the direction of arcs representing system relationships, i.e., a relation can be represented in either an "active" (R) or "passive" (R-1) form. In a CCG graph, there are two types of relationships between entities: unary and binary. The unary relations are "has_collaborated_with" and "has_endorsed" which are defined between entities of the same types of contributors and artifacts, respectively. These relations are symmetric. For binary relations, it is straightforward to show that all three relations can be defined in either active or passive form. Suppose that the relation of "has-contributed-to" is defined between a contributor and an artifact in passive form. This relation can be expressed in an active form of "has-contributor". Other two relations, "has-published-in" and "is-published-in", can be defined in an active form.

Table 1
Theoretical properties of the defined metrics (Note: M and D signify that the property is defined as a Mandatory in the DISTANCE framework or Desired in the property-based measurement framework, respectively).

	Poels and Dedene (2000)	Briand et al. (1996)	R_a^c	R_a^v	R_a^g	R_a^d	R_a	T_c^a	T_c^v	$T_c^{\hat{C}}$	T_c	I_v^a	I_v^c	I_v
Non- negativity	M	D	√	√	√	√	√	√	√	√	√	√	√	√
Symmetry	M	D	√	√	√	√	√	√	√	√	√	√	√	√
Identity	M	-	√	√	√	√	√	√	√	√	√	√	√	√
Triangular inequality	M	-	√	√	√	√	√	√	√	√	√	√	√	√
Null value	-	D	√	√	√	√	√	√	√	√	√	√	√	√
Monotonicity	-	D	√	√	√	√	√	√	√	√	√	√	√	√
Additivity	-	D	×	×	×	×	×	×	×	×	×	×	×	×

5.1.3. Property 3. Identity

Based on the metrics formulas presented in Sections 4.2–4.4, it is clear that all proposed metrics satisfy identity, i.e., if x is identical to y, then any property that is possessed by x is also possessed by y and vice versa.

5.1.4. Property 4. Triangular inequality

This property is intuitively justified by the fact that the proposed measures satisfy triangular inequality, since the graph's edges which show the relations between entities, act as a distance function.

5.1.5. Property 5. Null Value

The complexity of system $S =< E, R >$ is null if R is empty. In other words, metrics values are null when there are no relationships between system elements. In CCG graph, if there is no relationship between artifacts, contributors, and venue, all of the metrics' values will be null.

5.1.6. Property 6. Monotonicity

The complexity of a system $S =< E, R >$ is no less than the sum of the complexities of any two of its modules with no relationships in common. It means that adding relationships between elements of a system does not decrease its complexity; i.e., if any kind of indirect (or transitive) relationships between entities is considered in the computation of complexity, then the complexity of S will be larger than the sum of its modules' complexities (Briand et al., 1996). In our CCG graph, the complexity metrics consider only direct relationships between entities. For instance, when a contributor contributes to an artifact and artifact is published in a venue, we consider all possible binary relationships between these three entities, including "has-contributed-to", "is-published-in" and "has-published-in". As a result, no transitive relationship has remained to be considered in the computation of complexity metrics. Therefore, for all proposed metrics, the complexity of S is more than the sum of its modules' complexities.

5.1.7. Property 7. Additivity

The complexity of a system $S =< E, R >$ composed of two disjoint modules m_1 and m_2 is equal to the sum of the complexities of the two modules, i.e., it requires if two models are merged, the complexity of the resulting model is equal to the sum of the complexity of the two source models. This is not applicable to our system and does not necessarily need to be respected. For example, assuming a metric of M and two systems A and B, and let the system A be more complex than B, say $M(A) = 0.8$ and $M(B) = 0.4$. If M is an additive measure it is required to satisfy $M(A + B) = 1.2$ but a monotone M would only require $M(A + B) > 0.8$. Therefore, the monotonicity as respected by the proposed metrics is strict enough to satisfy the complexity, and a metric does not need to be additive (while being monotone) to be a valid complexity measure.

The summary of theoretical validation of our metrics is presented in Table 1. As shown, all of the proposed metrics respect the mandatory properties required by the DISTANCE framework, meaning that they are theoretically valid metrics. Furthermore, Table 1 shows that our metrics respect four out of the five desirable properties defined by the property-based measurement framework.

5.2. Empirical evaluation

In this section, we evaluate the robustness of our proposed quality metrics against attacks and manipulations. We use the application scenario we introduced in Section 2 for the purpose of performance comparison.

Regarding the selected application scenario, we choose one representative from each category of quality assessment approaches, namely: Hirsch as a popular citation-based and Eigenfactor as a well-known network-based approach.

The Hirsch model, proposed by Jorge E. Hirsch in 2005, is an index to quantify researchers' research outputs (Hirsch, 2005). This approach presents three different metrics for assessing the quality of people, papers and venues, which are h-index, number of citations and impact factor, respectively. The Hirsch model is the most popular approach adopted by almost all publishers and institutions, and that's why we select this approach for performance comparison.

Eigenfactor, proposed by Carl Bergstrom in 2007, is another well-known model for research output qualification (Bergstrom, 2007). As a Pagerank style approach, it relies on the structure of the research network and employs an iter-

Table 2
Statistical distributions in the dataset.

Statistical distribution	α	β
Number of authors per paper	0.221	3456.289
Number of papers per author	1218.148	19241.524
Number of citations per paper	0.597	6957.071
Number of references per paper	0.587	1315.053
Number of papers per venue	1.016	1.030

ative approach to compute more robust quality metrics. This approach assigns each venue an Eigenfactor index to venues and an influence score to articles. Although there is research in which some quality scores are assigned to authors based on Eigenfactor influence score (West, Jensen, Dandrea, Gordon, & Bergstrom, 2013), Eigenfactor officially does not propose any metric for evaluation of researchers.

5.3. Dataset

We use a synthetic dataset to analyze the robustness of our algorithm. The main reason for using the synthetic dataset is that we use the dataset to analyze the behaviour of models mainly in the presence of collusive behaviours. As we do not know the volume of collusion in the existing datasets, it is hard to use them for robustness measurement.

We use the *DBLP-Citation-network V11* AMiner dataset (Tang et al., 2008) which contains data extracted from DBLP, ACM, Microsoft Academic Graph, and other sources[7] as a guide to generate our synthetic dataset. The initial dataset contains information about 4,107,340 papers and 36,624,464 citations. We preprocessed the dataset and removed some nodes with incomplete information. Finally, we came up to a dataset containing information of with 3246606 papers, authored by 3655048 authors, published in 8771 venues. We use the Python NumPy library to estimate various statistical distributions parameters, which are beta distributions with the α and β parameters listed in Table 2 . We use these distribution parameters to generate the synthetic dataset.

In terms of the size, the generated dataset contains 3246 papers, authored by 3655 authors and published in 8 venues, which is one percent of the size of the real-world dataset. Moreover, each paper referenced by, and is cited by, at least one paper. To generate the dataset, we start by generating paper and then selecting authors, venues, references and citations based on the statistical distributions extracted from the real-world dataset.

Also, we use a subset of *DBLP-Citation-network V11* AMiner dataset, to show the applicability of our proposed model to the real-world scenarios. To do so, we extract the papers that are published in 2019. This dataset contains information about the 2255 papers published by 7727 authors in 128 venues. In the following sections, we will refer to this dataset as *DBLP19*. No enrichments or modifications have been applied to this dataset to make sure that the proposed model is applicable to real-world data directly, and not subject to data generation, aggregation or any other forms of modifications.

5.4. Applicability test

In this subsection, we show the applicability of our model to real world datasets. To this end, we first check the data items upon which our model relies. Looking at the formulations proposed in Section 4, CCG formulations are defined based on the following items:

- Papers: Only having a unique ID is enough for representing a paper. In addition to assigning a unique identifier to each paper, one can easily extract the title and some other metadata from the paper, itself.
- Authors: It is enough to have only a unique identifier assigned to each author, but one can extract the list of authors form the (pdf) file of the paper.
- Venues: Similarly, just having a unique identifier assigned to each paper is enough for CCG model. The title of the venue can also be extracted from the (pdf) file of the paper.
- References: References are also extractable from the paper, itself.
- Reviews: Subject to availability, the review results of the papers can also be taken into account when computing quality scores. However, the model will still work decently, in comparison with other related work, even if reviews are not available (see Section 5.7).

Excepting for the review results, which are not available in almost all bibliometric systems, the rest of required items are available and even are easily extractable from the pdf files of the papers. Moreover, these are the data items upon which almost all existing bibliometric systems, such as Hirsch and Eigenfactor, compute their quality metrics. This shows that theoretically, our proposed model is applicable to real-world scenarios.

[7] https://www.aminer.cn/citation

Table 3
Results of applying CCG to DBLP19 (Top 5 authors, papers and venues).

#	Author ID	Rank	#	Paper ID	Rank	#	Venue ID	Rank
1	2894635174	0.83	1	2895409706	0.18	1	189354248	0.53
2	2894867498	0.83	2	2893194824	0.18	2	7560371	0.38
3	2559411478	0.83	3	2898356931	0.18	3	147953040	0.27
4	1990054946	0.83	4	2893591486	0.18	4	61310614	0.12
5	2143705440	0.83	5	2894007473	0.18	5	186357190	0.10
..	..	..	..	..	..	..	..	..
10	2898308544	0.81	10	2895639546	0.16	10	165473669	0.08
..	..	..	..	..	..	..	..	..
50	2889478600	0.60	50	2892671625	0.08	50	21029587	0.06
..	..	..	..	..	..	..	..	..
100	2123686738	0.59	100	2886145918	0.0.4	100	204030396	0.06

Moreover, to show applicability of CCG practically, we apply our model to *DBLP19*. The results of this application are depicted in Table 3. We have shown in the table, the ranks of the top 5 authors, papers and venues, along with some more rows showing the 10th, 50th and the 100th authors, papers and venues as a subset of the whole results. Quality ranks are numbers in the range [0,1], where 0 means lack of quality and 1 means full quality.

5.5. Experimentation setup

The experimentation is conducted on a 64-bit machine with 8GB memory, running a Windows 10, with Intel dual Core i5 CPU. All algorithms are implemented in Python. We design three different attack scenarios for performance evaluation. The scenarios are designed based on the real-world scenarios that is reportedly being used in the research publication area. It is notable that these scenarios are also applicable to the generic form of the problem in the CCG area (Behkamal, Kahani, Bagheri, & Sezavar, 2016).

Moreover, currently there are no bibliometric datasets that include review results of the papers. On the other hand, as described earlier, it is not possible to assess quality of research articles automatically and it needs domain expert contributions. Due to this lack of information, we put the weight of quality of content, i.e., w_c^n to 0, to remove its impact on the computed scores. We change the weights so that $w_a^c + w_a^v + w_a^g + w_a^d = 1$.

5.5.1. Attack Scenario 1. Venue promotion

One of the popular attacks on quality metrics is promoting low-quality venues. One of the popular ways for promoting a venue is to ask others to cite the papers that are published in the target venue. Sometimes, journal editors ask authors to cite a specified number of papers published in that journal as a prerequisite to have their paper published in the venue. In this attack scenario, we thus simulate this scenario. We choose the venue with the smallest number of papers and try to promote it. To do so, we add a number of papers to the venue, each of which cites half of the venue papers. Subject to availability of information, the cited papers can be selected randomly, or based on content similarity. In our dataset, we do not have content information, so, we select the cited papers randomly. The number of added papers ranges from 20% of the number of venue papers to 100%, increased by 20. In this way, the number of citations to the papers of the venue changes. So, the quality score of the paper should increase. To control the impact of other factors, we choose for each paper just one author, and the paper cites only selected papers from the target venue.

5.5.2. Attack Scenario 2. Author promotion

Another known misbehaviour in CCG systems, mainly in research, happens when a group of authors cite each others' papers and, in this way, increase their quality metrics. In other words, every author in the collusion group cites the papers of others, rather than citing his/her own paper. This attack is harder to detect due to the collaborative nature of the attack. In this attack, we thus simulate this scenario. The average number of collaborating authors in the original real-world dataset is 4. So, we start from selecting 8 authors (twice the average group size) and increase this number to 40 which is 10 times the average group size. Members of each group cite half of the papers of other members.

5.5.3. Attack Scenario 3. Paper promotion

The third type of malicious behaviour is to promote a paper by adding citations to it. In other words, there are cases in which a paper is unfairly promoted by getting cited by a group of papers. These citations, coming from papers with different venues and authors, can simply increase the quality metrics, without getting noticed.

In this attack we simulate this scenario and try to promote quality metrics of 1% of the papers having the smallest number of citations. The average number of citations per paper in the real-world dataset is about 7. So, for each paper, we start by adding 40% of the number of citations and increase it up to 200%.

Table 4
System setup parameters.

Artifact related	Contributor related
$w_a^c = 0.5$	$w_c^a = 0.33$
$w_a^v = 0.3$	$w_c^v = 0.33$
$w_a^g = 0.1$	$w_c^{\hat{c}} = 0.34$
$w_a^d = 0.1$	
$\omega_f = 1.5$	
$\omega_o = 1$	
Venue Related	System related
$w_v^c = 0.5$	$\varepsilon = 0.01$
$w_v^a = 0.5$	

5.6. Evaluation metrics

To compare the robustness of our approach against the selected related approaches, we use the percent of the shift that occurs in the quality metric as the result of an attack and call it the *percentage error (PE)*. More precisely, the percentage error for a venue, for instance, is the percentage of the changes in the quality metric of the venue, due to an attack. Assume that the quality metrics of a venue before and after an attack are I_v^{before} and I_v^{after}, respectively. We calculate the percentage error for venue v as follows:

$$PE(v) = \frac{I_v^{after} - I_v^{before}}{I_v^{before}} \times 100 \tag{15}$$

A smaller score shift indicates a small change occurred in the quality score. This means that an attack has had a small impact on the quality score, and this implies the high level of robustness of the model.

We use *mean percentage error (MPE)* as the metric for comparing performance of model in a specific scenario. MPE is the mean of all PEs corresponding to each individual target venue/author/paper. For instance the MPE for a model in the first scenario is:

$$MPE = \frac{100}{n} \sum \frac{I_v^{after} - I_v^{before}}{I_v^{before}} \tag{16}$$

We follow the same approach for computing MPE for the other two attack scenarios. There are also some other weights and parameters that we use in the formulations of Section 4. Table 4 lists the values for those parameters. The values are selected just as a sample scenario, but the model is customizable with any other needed configuration.

5.7. Performance comparison results

In this section, we present the results of the experiments conducted for measuring the robustness of the three models: CCG, Hirsch and Eigenfactor.

5.7.1. Robustness comparison in Scenario 1

As the first scenario, we inject attacks to the synthetic dataset to promote a target venue. The size of the attacks ranges from 20% to 100%. Therefore, alongside the original synthetic dataset, we have 5 new datasets each of which is the original dataset that includes an attack with a specified size, based on what we explained earlier in Section 5.3.

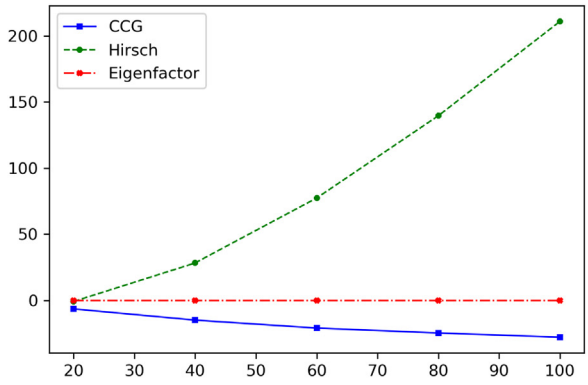
For robustness comparison, we apply three different models, namely CCG, Hirsch and Eigenfactor to each of these datasets. For each model, we use results obtained for applying the model to the original attack-free dataset as a measure to compute MPEs.

Fig. 4a reports the results of applying the three models to the datasets generated for the first scenario. The vertical axis shows the MPE and the horizontal axis represents the size of the attack. The Eigenfactor removes self-citations for venues, so the scores remain unchanged and the MPE for Eigenfactor is zero. The MPE for Hirsch increases drastically and for the attack with the size of 100%, it reaches a value of about 200%. The CCG shows an interesting behavior. The scores of the target venue decrease as the size of the attack increases. It means that CCG punishes target venues that try to get promoted due to organized attacks. The results suggest that CCG is more robust compared with the other two models against venue promotion attacks.

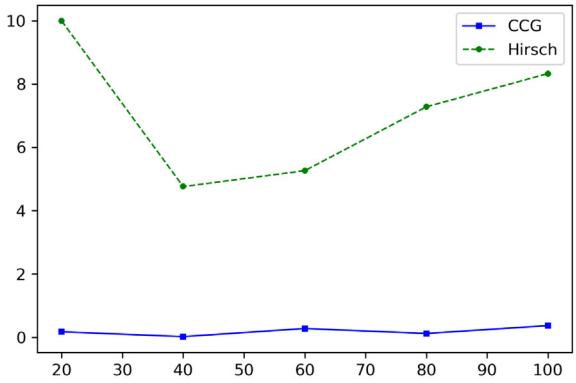
5.7.2. Robustness comparison in Scenario 2

In the second scenario, we attempt to promote a group of authors. We organize attacks with five different sizes. So, we generate five datasets, based on the original synthetic dataset, and use them for robustness comparison.

As we explained earlier, Eigenfactor does not provide any metrics for author quality assessment. So, in the second scenario, we only compare CCG and Hirsch. We apply these two models to the generated datasets and compare the MPE. The MPE



(a) Scenario 1.



(b) Scenario 2.

Fig. 4. Comparing robustness in scenarios 1 and 2.

in this scenario is the mean of percentage errors of all target authors. The results of the experiment are reported in Fig. 4b. As the chart shows, the MPE for the Hirsch model increases as the size of the attack increases, in general. In contrast the MPE for the CCG is very small and almost close to zero, showing that CCG is far more robust than Hirsch against the author promotion attack.

5.7.3. Robustness comparison in Scenario 3.

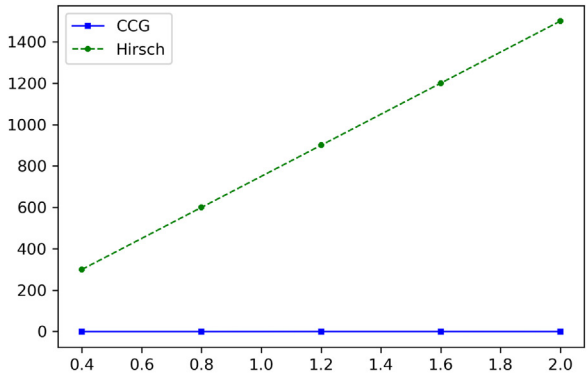
In the third scenario, we attempt to promote ten selected papers. We, again, inject attacks to the original synthetic dataset with five different sizes.

We apply all three models to the generated datasets. The MPE in this scenario is the mean of all percentage errors occurring in the quality scores of target papers. The results are reported in Fig. 5. Given the large values of MPEs for Hirsch, and relatively small MPE values for CCG and Eigenfactor, we present the comparison results of CCG with the other two models in two different charts. As shown in Fig. 5a, the MPE for the Hirsch model increases almost linearly, as the size of the attack increases. But the MPE for the CCG model remains very close to zero. Fig. 5b reveals more details about the behaviour of CCG. As shown in the figure, the MPE for Eigenfactor remains close to zero, but increases as the size of attack increases. But CCG again shows an interesting behaviour by punishing the targeted papers. After a slight increase in the MPE which is still negative, the MPE remains negative and decreases as the size of the attack increases.

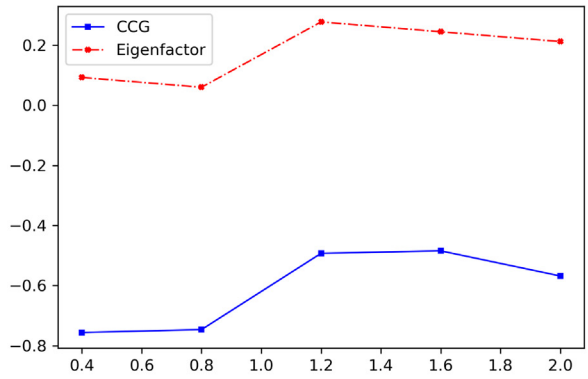
In summary, the experimental results show that the CCG is more robust against manipulation in all three scenarios in comparison with Hirsch and Eigenfactor.

6. Related work

Collaborative content generation is typically defined as generating artifacts by contributors and disseminating it to the community through a venue. Due to the rapid increase in the number of artifacts generated collaboratively in Web, assessing the quality of these artifacts has gained great attention in recent research.



(a) CCG vs. Hirsch



(b) CCG vs. Eigenfactor

Fig. 5. Comparing Robustness in Scenario 3.

Quality assessment is known to be a multi-dimensional problem (Bollen et al., 2009; Sidiropoulos, Gogoglou, Katsaros, & Manolopoulos, 2016). Several parameters affect the quality of an artifact. The first group of parameters characterizes the quality of the contributor, including his/her trustworthiness, experience, domain expertise, teamwork knowledge (Allahbakhsh, Samimi, Motahari-Nezhad, & Benatallah, 2014) and the relevant credentials and certificates achieved previously (Amintoosi & Kanhere, 2014). The second group of parameters is related to the artifact itself, such as its completeness, accuracy, reliability, and relevance (Amintoosi & Kanhere, 2014). The third group describes the venue the artifact has been presented in. The artifact presented in a highly reputable venue is believed to have high quality, due to being the subject of more strict selection criteria. In the following, we review the approaches proposed for quality assessment, and categorize them based on these parameters.

6.1. Contributor evaluation

Past research efforts have considered the contributor's role to assess the quality of the artifact. Shen and Barabafisi (2014) proposed a scheme to evaluate the credit of each author in a co-authored work that assigns more credit to the "senior author". To define the "seniority", they consider the number of papers published by the author ands the degree to which these papers share citations from papers citing the one under consideration. So, the author whose papers have more similarity to the one under consideration gains more "seniority". The approach by Persson (2017) is based on a similar idea but it considers the author ability and estimates the weight of author contributions to a body of co-authored work. The problem of contributor credit assessment has been also investigated by Perianes-Rodriguez, Waltman, and Van Eck (2016). They proposed a scheme to construct a bibliometric network by considering full as well as fractional counting approaches. Also, Dehdarirad and Nasini (2017) addressed the same problem and proposed a statistical model to analyse the research impact and quality in co-authorship structures.

Ciaccio et al. (2019) carried out a comparison between three author indices, i.e., h-index, c-index and c′-index, among the top 10 authors who have added celiac disease as the keyword in Google Scholar. By using the h-index, authors are ranked

based on their h publications with h or more citations. In the c-index calculation, the author's position affects the author's contribution weight. In c'-index, the author position and the total number of authors is considered in weighting author contribution. More comprehensive reviews on citation metrics can be found in Waltman (2016), Mingers and Leydesdorff (2015), Zeng et al. (2017). Panagopoulos, Tsatsaronis, and Varlamis (2017) proposed a computational methodology to evaluate the contributors (here, scholars writing papers) using an unsupervised machine learning approach. They considered a set of quantity (volume of publication), impact (number of citations it receives), and sociability (ability to collaborate with well-known scholars) for each contributor and monitor them over the time to create a profile for each contributor. These features are then used to cluster the contributors and identify/predict scholars with ever-growing number of citations per year.

6.2. Artifact evaluation

Some approaches have considered the role of the artifact itself and its characteristics in the assessment of artifact quality. Haustein et al. (2015) took into account the discipline, document type, title length, number of authors, number of references, and number of institutions and countries found in authors' addresses. Ferrara, Alipoufard, Burghardt, Gopal, and Lerman (2017) considered the contributor quality and the change in the contributor behaviour and analyzed its impact on the artifact quality. The concept of a knowledge network and its impact on the artifact quality has been introduced by Guan, Yan, and Zhang (2017), where they discuss the impact of knowledge elements (e.g., article keywords) and collaboration elements (e.g., author centrality) on the publication citation. Teplitskiy et al. (2020) carried out a study to evaluate the relation between citation, impact and quality and showed that the citation count is disadvantageous to a large numbers of papers due to giving a perception of the low quality of these papers. In other words, equating the number of citations with quality results in overestimating the quality of highly cited papers.

The well-known idea of ranking the Web pages has been introduced in the PageRank algorithm (Page et al., 1999) where the quality and the number of pages that have links to a page are considered to iteratively evaluate the rank of the page. PageRank is claimed to be robust to manipulations due to considering global variables that are stronger against manipulation. PageRank has inspired several ranking methods known as PageRank-inspired schemes. Examples are Eigenfactor and SCImago Journal Rank (SJR) (Bollen et al., 2009).

6.3. Venue evaluation

Past research efforts have also considered the role of the venue quality in the quality of the artifact presented in it. In the domain of article publication, a well-known indicator of the venue quality is the journal's impact factor. The paper published in a journal with a high impact factor is known to be a paper with high quality. Some research (e.g., Huang, 2016; Rousseau, 2016) claimed that there exists a positive correlation between impact factor and article number in scholarly journals. In other words, the quality of the venue is in direct relationship with the quantity of its published papers. High impact journals publish more articles. However, they did not discuss the impact of contributors in the quality of venues.

6.4. Comparison with our work

Even though several approaches have been proposed for quality assessment in CCG systems, these approaches do not consider all the three relevant factors (i.e., contributors, artifacts and venues) simultaneously. Moreover, they do not take into account all effective quality parameters into considerations. Another important drawback of those approaches is that they do not consider the two-way relation between the contributor-artifact-venue parameters. Also those approaches are prone to manipulation and collusion due to their non-comprehensive quality metric selection.

In other words, briefly speaking, the proposed methods can be generally classified into two broad categories: iterative solutions and non-iterative ones. Non-iterative solutions, as proven in the literature (Allahbakhsh & Ignatovic, 2015; Page et al., 1999; Rezvani, Allahbakhsh, Vigentini, Ignjatovic, & Jha, 2015), are prone to manipulation. The method we propose here, is an iterative method and is more robust against manipulation.

Regarding the iterative techniques such as Eigenfactor, or in general PageRank inspired algorithms, the proposed methods in the literature, mainly rely on the concept of citation count. They generally rely on a uni-partite graph of articles and their references and compute metrics based on this relationship. Different from the existing methods, we propose a method that takes into account all different aspects of the CCG system, which are people and their collaborations, venues, papers and their citations. Therefore, to the best of our knowledge, our proposed method is the most comprehensive method proposed that ignores none of the important factors that can have an impact on the quality of an artifact. This is what on which the DORA-declaration emphasises too.

In summary, the main differences of our proposed method with the related works is two-folded: Comprehensiveness and robustness. Comprehensiveness is shown in the formulations and definitions in the Section 4, and the robustness of the method is shown in Section 5.

7. Conclusion

In this paper, we proposed a novel framework for quality assessment in collaborative content generation (CCG) systems. To make quality scores comprehensive, the proposed method takes into account all referenced-based and network-based

parameters when computing scores. We consider parameters such as quality of contributors, the order of the contributors, quality of references and citations, and quality of the venue as contributing parameters to the quality of an artifact. Our model also considers the quality of collaborators, quality of the venues and quality of the artifacts when evaluating a contributor, and quality of artifacts and contributors when evaluating a venue.

We also proposed an iterative definition for quality and employed a novel iterative algorithm to compute quality scores for contributors, artifacts and venues. We used a theoretical validation approach to assess the accuracy and validity of the proposed quality metrics. We have also compared the robustness of our proposed method against manipulation with two well-known related approaches. The promising results of the experimentation suggest that the proposed model provides more robust and comprehensive quality scores in comparison with the related approaches.

As future research work, we plan to apply our model to a real-world large enough dataset and make the results available through a service or an online tool. We are also working on involving more parameters in our computations to make more comprehensive intuitive quality scores. Precisely, we plan to involve the quality of the content of artifacts, when possible (e.g., quality of the contents of a research paper) to study their impact on the performance of our model.

Currently, the conferences are going to be held on-line or hybrid more often, specifically after the COVID-19 pandemic. This can result in the availability of a corpus of blog posts, audio files, video recordings, ratings, likes and dislikes, and comments given by audience and communities, related to each paper. These are rich sources of information that can be investigated, as future works, to examine the quality of an article content.

Author contributions

Mohammad Allahbakhsh: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper; Other contribution.

Haleh Amintoosi: Conceived and designed the analysis; Contributed data or analysis tools; Performed the analysis; Wrote the paper; Other contribution: Conceptualization.

Haleh Amintoosi: Conceived and designed the analysis; Contributed data or analysis tools; Performed the analysis; Wrote the paper; Other contribution: Conceptualization.

Behshid Behkamal: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Wrote the paper; Other contribution: Conceptualization.

Amin Beheshti: Contributed data or analysis tools; Wrote the paper: Writing - Review & Editing; Other contribution: Conceptualization.

Elisa Bertino: Conceived and designed the analysis; Wrote the paper; Other contribution: Conceptualization, Supervision.

Acknowledgment

The authors would like to thank Mr. Yaser Kazemi for his contributions to developing the idea in preliminary steps.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.joi.2020.101127.

References

Allahbakhsh, M., & Ignatovic, A. (2015). An iterative method for calculating robust rating scores. *IEEE Transactions on Parallel and Distributed Systems*, *26*(2), 340–350.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, *17*(2), 76–81. https://doi.org/10.1109/MIC.2013.20

Allahbakhsh, M., Samimi, S., Motahari-Nezhad, H.-R., & Benatallah, B. (2014). Harnessing implicit teamwork knowledge to improve quality in crowdsourcing processes. *in: 2014 IEEE 7th International Conference on Service-Oriented Computing and Applications IEEE*, 17–24.

Amintoosi, H., & Kanhere, S. (2014). A reputation framework for social participatory sensing systems. *Mobile Networks and Applications*, *19*(1), 88–100. https://doi.org/10.1007/s11036-013-0455-x

Behkamal, B., Kahani, M., Bagheri, E., & Sezavar. (2016). A metric suite for systematic quality assessment of linked open data. *International Journal of Information and Communication Technology*, *8*(3), 27–45.

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, *68*(5), 314–316.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLOS ONE*, *4*(6).

Braithwaite, J., Herkes, J., Churruca, K., Long, J. C., Pomare, C., Boyling, C., . . . Shih, P., et al. (2019). Comprehensive researcher achievement model (cram): a framework for measuring researcher achievement, impact and inuence derived from a systematic literature review of metrics and models. *BMJ Open*, *9*(3), e025320.

Briand, L. C., Morasca, S., & Basili, V. R. (1996). Property-based software engineering measurement. *IEEE Transactions on Software Engineering*, *22*(1), 68–86.

Cagan, R. (2013). *The San Francisco Declaration on Research Assessment*.

Ciaccio, E. J., Bhagat, G., Lebwohl, B., Lewis, S. K., Ciacci, C., & Green, P. H. (2019). Comparison of several author indices for gauging academic productivity. *Informatics in Medicine Unlocked*, 100166.

Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: a survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, *51*(1) https://doi.org/10.1145/3148148

de Kerchove, C., & Van Dooren, P. (2010). Iterative filtering in reputation systems. *SIAM Journal of Matrix Analysis and Applications*, *31*(4), 1812–1834. https://doi.org/10.1137/090748196

Dehdarirad, T., & Nasini, S. (2017). Research impact in co-authorship networks: A two-mode analysis. *Journal of Informetrics*, *11*(2), 371–388.
Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, *54*(4), 86–96.
Ferrara, E., Alipoufard, N., Burghardt, K., Gopal, C., & Lerman, K. (2017). Dynamics of content quality in collaborative knowledge production. *in: Eleventh International AAAI Conference on Web and Social Media*.
Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics*, *11*(2), 407–422.
Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLOS ONE*, *10*(3), e0120495.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.
Huang, D.-w. (2016). Positive correlation between quality and quantity in academic journals. *Journal of Informetrics*, *10*(2), 329–335.
Laureti, P., Moret, L., Zhang, Y., & Yu, Y. (2006). Information filtering via iterative refinement. *EPL (Europhysics Letters)*, *75*, 1006.
Matei, S. A., Jabal, A. A., & Bertino, E. (2018). Social-collaborative determinants of content quality in online knowledge production systems: comparing wikipedia and stack overow. *Social Network Analysis and Mining*, *8*(1), 36.
Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web. Tech. rep*. Stanford InfoLab.
Panagopoulos, G., Tsatsaronis, G., & Varlamis, I. (2017). Detecting rising stars in dynamic collaborative networks. *Journal of Informetrics*, *11*(1), 198–222.
Perianes-Rodriguez, A., Waltman, L., & Van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, *10*(4), 1178–1195.
Persson, R. A. (2017). Bibliometric author evaluation through linear regression on the coauthor network. *Journal of Informetrics*, *11*(1), 299–306.
Poels, G., & Dedene, G. (2000). Distance-based software measurement: necessary and suficient properties for software measures. *Information and Software Technology*, *42*(1), 35–46.
Rezvani, M., Allahbakhsh, M., Vigentini, L., Ignjatovic, A., & Jha, S. (2015). An iterative algorithm for reputation aggregation in multi-dimensional and multinomial rating systems. In H. Federrath, & D. Gollmann (Eds.), *ICT Systems Security and Privacy Protection* (pp. 189–203). Cham: Springer International Publishing.
Rousseau, R. (2016). Positive correlation between journal production and journal impact factors. *Journal of Informetrics*, *10*(2), 567–568.
Salk, C. F., Sturn, T., See, L., Fritz, S., & Perger, C. (2016). Assessing quality of volunteer crowdsourcing contributions: lessons from the cropland capture game. *International Journal of Digital Earth*, *9*(4), 410–426.
Schmid, S. L. (2017). Five years post-dora: promoting best practices for research assessment. *Molecular Biology of the Cell*, *28*(22), 2941–2944.
Shen, H.-W., & Barabafisi, A.-L. (2014). Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, *111*(34), 12325–12330.
Sidiropoulos, A., Gogoglou, A., Katsaros, D., & Manolopoulos, Y. (2016). Gazing at the skyline for star scientists. *Journal of Informetrics*, *10*(3), 789–813.
Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. *in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 990–998.
Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2020). *Citations systematically misrepresent the quality and impact of research articles: Survey and experimental evidence from thousands of citers.* , arXiv preprint arXiv:2002.10033.
Thuan, N. H., Antunes, P., & Johnstone, D. (2016). Factors inuencing the decision to crowdsource: A systematic literature review. *Information Systems Frontiers*, *18*(1), 47–68.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.
West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., & Bergstrom, C. T. (2013). Author-level eigenfactor metrics: Evaluating the inuence of authors institutions and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, *64*(4), 787–801.
Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., & Stanley, H. E. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714*, 1–73.