International Journal of Modern Physics C
Vol. 29, No. 5 (2018) 1840006 (10 pages)

DOI: 10.1142/S0129183118400065

World Scientific
www.worldscientific.com

Measuring diversity of music tastes in online musical society

Hao Li

DMI Research Center, Hangzhou Normal University
Hangzhou 311121, P. R. China
leo.t5@outlook.com

Xiao-Pu Han*

Alibaba Research Center for Complexity Sciences
Hangzhou Normal University
Hangzhou 311121, P. R. China
xp@hznu.edu.cn

Linyuan Lü*

Institute of Fundamental and Frontier Sciences
University of Electronic Science and Technology of China
Chengdu 610054, P. R. China

Alibaba Research Center for Complexity Sciences
Hangzhou Normal University
Hangzhou 311121, P. R. China
linyuan.lv@gmail.com

Zhigeng Pan*

DMI Research Center, Hangzhou Normal University
Hangzhou 311121, P. R. China

Institute of Virtual Reality, Guangzhou NINED Corp
Guangzhou, P. R. China
zgpan@hznu.edu.cn

Received 30 September 2017
Accepted 14 March 2018
Published 24 May 2018

The diversity of people's musical tastes is one of the significant parts which helps people to better understand the behavior trends and cultural preferences of people. In this paper, based on Hill-type true diversity, we propose an improved diversity metric that fairly captures the diversity of musical tastes. This diversity efficiently considers all the three aspects of diversity definitions: variety, balance, and disparity, and keeps higher discriminatory power. Using this diversity metric, one can analyze users' music tastes on Xiami.com, one of the largest social music media in China; we explore the association between the diversity and various variables which represent users' personal traits, as well as the difference between different genre levels and map the cultural pattern of difference genres. Our findings dig out many efficient factors that

*Corresponding authors.

deeply impact users' music tastes, and provide the global pattern of musical cultural structure on the Chinese online music society.

Keywords: Diversity; music tastes; genre distances; analysis on user's preference.

PACS Nos.: 89.65Ef, 89.75Hc.

1. Introduction

Music, which does much more than conveying significantly through nonverbal means, plays a vital role in human society.[1,2] As our listening behavior comes to digital age and considering the fast development of online music websites, the big-sized datasets from online musical society have been an important source in updating our understanding of human musical behaviors, and have attracted much more attention from researchers.

Many culture-relevant information, such as users' psychological tendency, aesthetic, cultural background, social surroundings, and so on, hide in people's musical preferences. The analysis on musical tastes therefore is an efficient way in the digging of new insights on social influence and cultural diffusion. In this issue, based on the datasets from several well-known online music societies, such as Last.fm[3,4] and Spotify,[5,6] many interesting and worthy results were found. In Refs. 7 and 8, one's diversity of musical tastes could serve as a proxy of the degree of one's openness. Reference 9 confirms that this information may influence one's music listening behavior. It was pointed out that the recommendations from the system will affect users' choices of music.[10] In the retrieval of music information, some researchers have explored to achieve the optimal balance between the similarity and diversity on recommendation.[7,11] Reference 12 considers certain types of music as similar or dissimilar, and use co-consumption behavior of users to measure it. Besides, several mature concepts of diversities, such as Rao–Stirliing diversity, were applied to analyze the music consumption behaviors in recent studies.[8,13,14]

In this paper, we focus on the analysis of diversity of user's music tastes, collect the dataset of a large number of users' music lists on a Chinese music online society Xiami (Sec. 2), propose an improved true diversity metric that fairly captures the diversity of users' music tastes (Sec. 3). By analyzing the music tastes of users in the dataset using a novel diversity metric, we find out a series of efficient associations between users' music tastes and various intrinsic or external factors, as well as the musical cultural pattern of users of Xiami (Sec. 4).

2. The Dataset

The dataset in our study is crawled from the API of Xiami (http://xiami.com) by using web crawler. Xiami is one of the most popular online music society in China, which owns millions of songs and hundreds of millions of registered users. To obtain a valid sample of Xiami users, we used two typical sampling methods, namely the

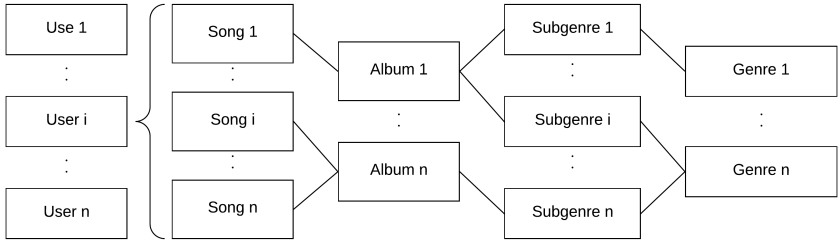


Fig. 1. The structure of the dataset of Xiami.com.

stratified sampling and random sampling. As of September 14, 2017, the date we carried out the sampling, there are about 325 million registered users on Xiami. Firstly, we clustered all users into 100 groups, ordered by their unique user ID. For each group, we made the simple random sampling, aimed to acquire 32 500 users, namely 1% of the whole set.

Next, we identified the users with valid information. The dataset includes six distinct fields: the library of favorite songs, the list of favorite songs of each user, list of users' profile, the list of albums which contains these songs, the lists of artists who produced these songs, and the lists of genres. Here, the genre information includes two levels: genre level and subgenre level (see its detailed information and discussions later).

From the above 1% sampling dataset of Xiami, 585 801 unique users with valid information were selected, and 820 663 unique songs, 8 338 301 records of favorites, 267 911 unique albums, 82 589 unique artists, 23 unique genres and 471 unique subgenres were also collected separately. The users who didn't set their full information and the albums that lack information of genre level or subgenre level were removed from our analysis. Finally, we obtained a set with 59 454 users, which could be used for further analysis. All these users hold 265 456 unique music at the genre level and 256 195 unique music at the subgenre level. Figure 1 shows the data structure. Each song in Xiami belongs to one certain album. Most songs belong to one certain artist, and some of them are created by several artists. Considering the style evolution of every artist, the music styles that users listen to were extracted from albums at both genre and subgenre levels.

3. Diversity and Genre Distances

Diversity is a concept which is widely used in many studies on social analysis.[15] Various diversity measures have been proposed for different applications.[16,17] Volume and Shannon entropy are the two most commonly-used methods.[18–20] However, due to genres of music usually are hierarchical and have internal relations, these diversity metrics still have some deficiencies in using of analysis on user's music tastes.

Generally, in the definition of most type of diversities, variety (i.e. the number of different categories) and balance (i.e. evenness of distribution) are two mainly-considered

factors. The variety describes how many types of things we have. In our analysis on musical tastes, it is the number of musical categories one favors. When it's considered alone, the greater the variety is, the larger the diversity it has.

The balance is relevant to what the relative number of items of each type is. In our analysis, it is a pattern of the count of user's favorites across musical categories. The more flattened the distribution is, the more balanced the distribution is, and the larger the diversity it has. Let's consider two users, one favors eight rock songs and two classical songs, the other favors five rock songs and five classical songs. Obviously, without the consideration of balance of a user's favorites and only using variety as measurement, they have the same results. However, from the perspective of balance, their music habits are different. The former likes rock music better than classical music, while the latter has no particular preference for the two types. Note that although users have the same variety and balance, their music taste patterns may also be different because the similarity between the songs matters. For example, the similarity between Opera and Symphony is apparently lower that that between New Age and R&B.

We use disparity to quantify the similarity between certain types of groups or classes. Here, disparity answers the question: How different from each other are the types of things that we observe? It refers to the manner and the degree in which things may be distinguished. Higher disparity leads to larger diversity. In our analysis, disparity relates to the type of genres of music.[12]

In Ref. 21, authors considered variety, balance, and disparity and proposed Rao–Stirling diversity which reads

$$D = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} (d_{ij})(p_i p_j), \tag{1}$$

where p_i and p_j are proportions of elements i and j in the system (balance) and d_{ij} is the degree of difference (disparity) attributed to elements i and j. Some recent studies have introduced this measure into music consumption.[8,13,14] However, according to the study reported in Ref. 22, the Rao–Stirling diversity shows low discriminatory power, which is further confirmed in the observation reported in Ref. 23. Importantly, according to the discussion in Ref. 24, Rao's measure actually is not a true diversity. Inspired by the discussions in Refs. 24–27, we propose a novel diversity using their Hill-type indicator:

$${}^{q}D^{S} = \left(\sum_{i=1}^{N} p_i \left(\sum_{j=1}^{N} s_{ij} p_j \right)^{q-1} \right)^{\frac{1}{1-q}}, \tag{2}$$

where p_i (> 0) is the fraction of user's preference for genre i, and $\sum_{i=1}^{n} p_i = 1$. Here, q is a parameter between 0 and ∞, indicating how much significance is attached to

genre's abundance. The similarity between genres is represented in a similarity matrix $\boldsymbol{S} = s_{ij}$. We assume $0 \leq s_{ij} = s_{ji} \leq 1$ and $s_{ii} = 1$. Reference 24 has proved that ${}^{q}D^{S}$ satisfies all properties a true diversity should possess for all q in their research. Without loss of generality and considering the computational complexity, we use the case when $q = 2$:

$$ {}^{2}D^{S} = \left(\sum_{i=1}^{N} p_i \left(\sum_{j=1}^{N} s_{ij} p_j \right) \right)^{-1} = \frac{1}{\sum_{i,j=1}^{N} s_{ij} p_i p_j}. \tag{3} $$

Due to $s_{ij} = 1 - d_{ij}$, and putting it into Eq. (3), we can find the relationship between ${}^{2}D^{S}$ and Rao–Stirling diversity:

$$ {}^{2}D^{S} = \frac{1}{\sum_{i,j=1}^{N} (1 - d_{ij}) p_i p_j} = \frac{1}{\sum_{i,j=1}^{N} p_i p_j - D} = \frac{1}{1 - D}. \tag{4} $$

Note that if a user only favors one music style, his diversity measured by Rao–Stirling measure should be zero, but ${}^{q}D^{S} = 1$. In our analysis, we use this method to measure the diversity of user's musical tastes. To optimize the results, three types of dissimilarity of genres are computed and compared by using different distance measures and different datasets.

The first dissimilarity is obtained by using cosine distance based on users' preference, which is adopted by previous research.[14] Cosine distance = 1 − cosine similarity. Given two vectors, $\boldsymbol{A}$ and $\boldsymbol{B}$, the cosine similarity is defined as

$$ \cos\theta = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2} \sqrt{\sum_{i=1}^{N} B_i^2}}, \tag{5} $$

where A_i and B_i are components of vector $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.

The second one is obtained by using cosine distance and computing pairwise occurrence of musical genres. To achieve this goal, an $M \times N$ album-genre matrix is constructed. Each row represents one unique album, and g_{ij} in column j represents whether album i owns genre j ($g_{ij} = 1$) or not ($g_{ij} = 0$). By using this matrix, we calculate the cosine distance between each pair of matrix columns.

The last one is derived from combining Jaccard distance and our album-genre matrix. The Jaccard distance, which measures dissimilarity between sample sets, is obtained by subtracting the Jaccard similarity from 1. Jaccard similarity, also known as intersection over union, is defined as the size of the intersection divided by the size of the union of the sample dataset.

$$ J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{6} $$

4. Findings

Using Eq. (3), we calculate the diversity of each user at the genre level and the subgenre level. Figure 2 shows the distribution of diversity at the genre level. Obviously, the diversity of the majority is lower than 2.

Figure 3 shows a positive relationship between the diversity at the genre level and diversity at the subgenre level, which is similar to the findings on interdisciplinary diversity.[29] The ${}^2D^S$ diversity is obviously dependent on the classification of music style, their Pearson correlation coefficient r is 0.791 with an extreme significance $P = 0$ (lower than 10^{-8}), indicating that the users with diverse genre tastes usually have more diverse tastes on subgenres. This fact illustrates that our measure is a discriminative indicator because such a fact hasn't been disclosed by the Rao–Stirling measure in previous research.

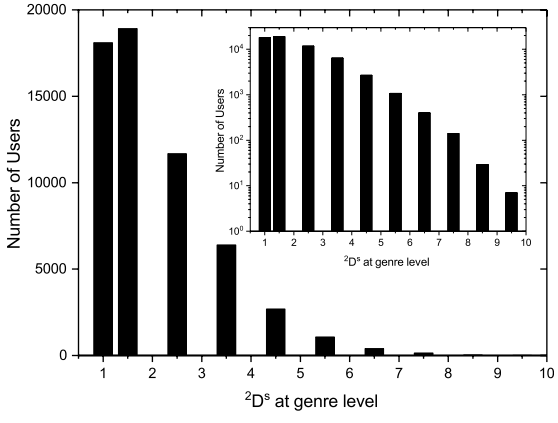


Fig. 2. The distribution of user's diversity at the genre level. The inset shows this distribution in semi-log plot. At x-axis, the bar at scale 1 shows the number of users with diversity at the lowest limit 1.0, and the bars at scale 1.5 and 2.5, respectively, corresponds to the range of diversity $(1, 2]$ and $(2, 3]$, and so on.

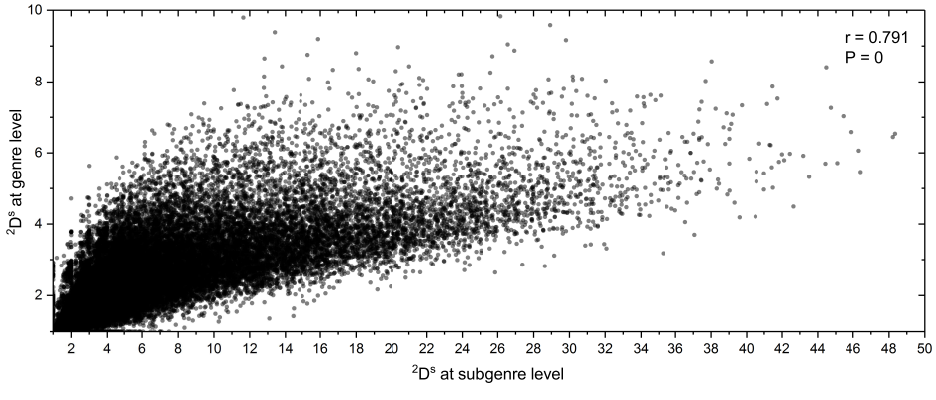


Fig. 3. The correlation between diversity values at the genre and subgenre levels.

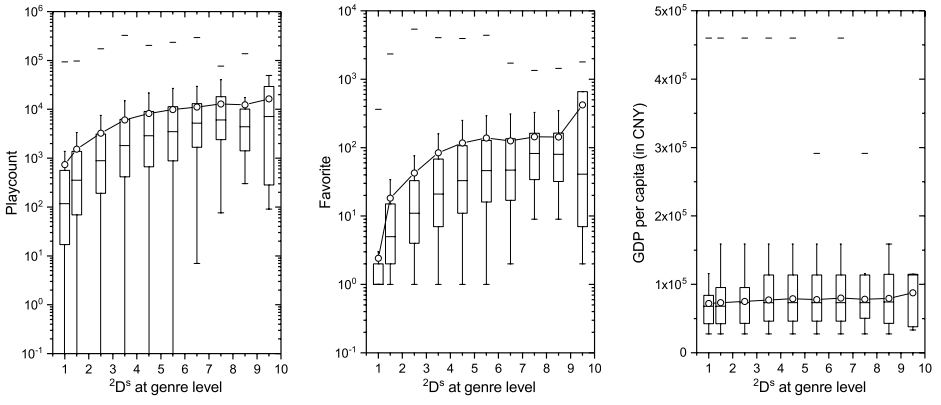
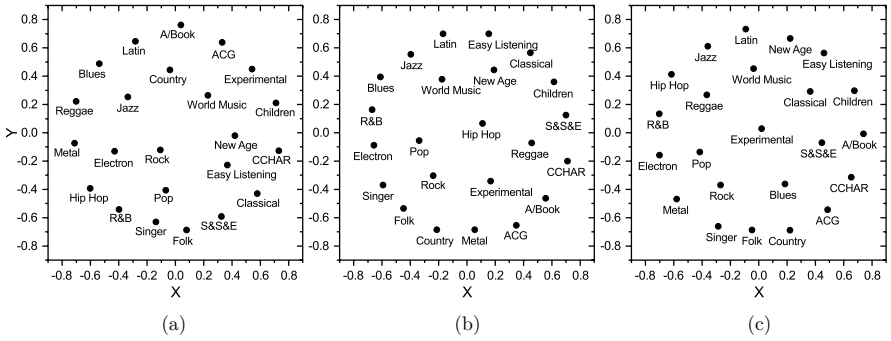


Fig. 4. Boxplot of the dependence between three variables of users and their diversity at the genre level. The meaning of tick labels in the x-axis is the same as that in Fig. 2. The three panels show the boxplot of user's playcount (panel (a)), number of favorite songs (panel (b)) and GDP per capita of the province of user living (panel (c)), respectively. At x-axis, the bar at scale 1 shows the number of users with diversity at the lowest limit 1.0, and the bars at scale 1.5 and 2.5, respectively corresponds to the range of diversity $(1, 2]$ and $(2, 3]$, and so on. The bars in each box from outside to inside, respectively, show the mild outliers, the upper and lower whiskers, the upper and lower quartiles, and the median. The circle-pointed line in each panel shows the trend of the average value.

Moreover, we dug out the effect of several intrinsical or external factors which would be dependent on users' diversity of music tastes. As shown in Figs. 4(a) and 4(b), users with higher diversity on music tastes would be more active in the online music society: they usually have higher playcount and more favorite songs. More interestingly, we collected the GDP per capita of each province in China in 2016, and found that the users with higher diversity may be living in the provinces with higher GDP (see Fig. 4(c)), implying the possible promoting effect of economic development on users' diversity of music tastes. Possibly, these results mean that people living in developed area would have more chance to contact various genres.

We further analyzed the distribution of distance between different genres using the three types of dissimilarity. By using the method of multidimensional scaling,[28] we mapped the genre distance for different distance definitions on a two-dimensional space, as shown in Fig. 5. The figure on the left is the co-consumption method used by Refs. 8, 13 and 14, but the scaling is a far cry from that in Ref. 14. From the figure, we concluded that measuring the disparity of genres by using co-consumption method is improper because users' consumption mode changes along with the change of researchers' choices of samples.

Comparing the cosine distance and jaccard distance based on the same data, we found that the distribution of each node measured by cosine distance is more even than that by jaccard distance. In high-dimensional space, cosine distance is almost invalid, i.e. vectors are nearly perpendicular. Since our data of album is in hundreds of thousands, it is clear that the disparity between genres shouldn't be measured by



(a) (b) (c)

Fig. 5. The genre maps using three different definitions of genre distances. Panel (a) represents cosine distance based on co-consumption.[14] Panel (b) represents cosine distance based on the information from albums. Panel (c) represents Jaccard distance based on the information from albums. Here CCHAR means Chinese Characteristic, A/Book means Audio Book. Singer means Singer-Songwriter, and S&S&E means Stage & Screen & Entertainment.

cosine distance. There also exists some notable inappropriateness in Fig. 5(b). For example, Blues and R&B are closer in Fig. 5(b) than in Fig. 5(c). Moreover, Reggae is closer to CCHAR (Chinese Characteristic) in Fig. 5(b), but closer to Hip Hop in Fig. 5(c) which is more reasonable. All of these results above confirm that Jaccard distance is a better metric to analyze the distance between different genres, and thus the efficiency of our novel diversity metric is convinced.

5. Conclusion

In this paper, we firstly proposed an improved measure of diversity of musical tastes based on the Hill-type true diversity. Comparing with the diversity metrics used in previous studies on the analysis of music-listening behaviors, our new metric not only efficiently considers all of the three aspects — variety, balance, and disparity — but also keeps higher discriminatory power. Using this metric of diversity, we further analyzed the music tastes of users on Xiami, one of the largest social music media in China. Our empirical findings are twofold: On the one hand, we found the dependence between the diversity and several intrinsical or external factors of users, such as user's activity, and the economic development level of users' living location, which is helpful in the understanding of influenced factors on human music tastes. On the other hand, we provided the map of musical social distance between genres, which reflects some information on the cultural structure in the social sense of the users in Chinese online musical society. In summary, these empirical findings confirm the efficiency of the improved diversity metric, and provide some deeper insights on musical sociology and culture popularity.[30]

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2018YFB1004901), the National Natural Science Foundation of China

(Nos. 11622538, 61673150 and 61673151), and the Zhejiang Provincial Natural Science Foundation of China (Nos. LGF18F030007 and LR16A050001). Correspondence should be addressed to Z. Pan (zgpan@hznu.edu.cn), X.-P. Han (xp@hznu.edu.cn) and L. Lü (linyuan.lv@gmail.com).

References

1. T. DeNora, *Music in Everyday Life* (Cambridge University Press, Cambridge, 2000).
2. T. J. Tighe and W. J. Dowling, *Psychology and Music: The Understanding of Melody and Rhythm* (Psychology Press, Taylor & Francis Group, New York and London, 2014).
3. F. Figueiredo, B. Ribeiro, C. Faloutsos, N. Andrade and J. M. Almeida, Mining online music listening trajectories, in *Int. Soc. Music Information Retrieval Conference(ISMIR)* (2016), p. 688.
4. N. K. Baym and A. Ledbetter, *Inform. Comm. Soc.* **12**, 408 (2009).
5. B. Zhang, Understanding user behavior in spotify, in *INFOCOM, 2013 Proc. IEEE* (2013), p. 220.
6. M. Pichl, E. Zangerle and G. Specht, Combining spotify and Twitter data for generating a recent and public dataset for music recommendation, in *Grundlagen von Datenbanken* (2014), p. 35.
7. L. Chen, W. Wu and L. He, How personality influences users' needs for recommendation diversity? in *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (2013), p. 829.
8. K. Farrahi, M. Schedl, A. Vall, D. Hauger and M. Tkalcic, Impact of listening behavior on music recommendation, in *Int. Soc. Music Information Retrieval Conference (ISMIR)* (2014), p. 483.
9. Z. Yang, J. Wang and M. Mourali, *J. Bus. Res.* **68**, 516 (2015).
10. J. M. Buldú, P. Cano, M. Koppenberger, J. A. Almendral and S. Boccaletti, *New J. Phys.* **9**, 172 (2007).
11. S. M. McNee, J. Riedl and J. A. Konstan, Being accurate is not enough: How accuracy metrics have hurt recommender systems, in *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (2006), p. 1097.
12. F. Morchen, A. Ultsch, M. Nocker and C. Stamm, Databionic visualization of music collections according to perceptual distance, in *Int. Soc. Music Information Retrieval (ISMIR)* (2005), p. 396.
13. N. Hurley and M. Zhang, *ACM T. Internet. Tech.* **10**, 14 (2011).
14. M. Park, I. Weber, M. Naaman and S. Vieweg, Understanding musical diversity via online social media, in *ICWSM* (2015), p. 308.
15. M. Zitt, *Measurement* **3**, 38 (2009).
16. A. E. Magurran, *Measuring Biological Diversity* (Wiley-Blackwell, Oxford, 2003).
17. R. Rousseau and P. Van Hecke, *Acta Biotheor.* **47**, 1 (1999).
18. C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Champaign, 1998).
19. N. Eagle, M. Macy and R. Claxton, *Science* **328**, 1029 (2010).
20. J. Bruggeman, arXiv:1011.0208.
21. A. Stirling, *J. R. Soc. Interface* **4**, 707 (2007).
22. Q. Zhou, R. Rousseau, L. Yang, T. Yue and G. Yang, *Scientometrics* **93**, 787 (2014).
23. L. Leydesdorff and I. Rafols, *J. Informetr.* **5**, 87 (2011).
24. T. Leinster and C. A. Cobbold, *Ecology* **93**, 477 (2012).

25. L. Jost, *Oikos* **113**, 363 (2006).
26. L. Jost, *Ecology* **88**, 2427 (2007).
27. L. Jost, *Ecol. Econ.* **68**, 925 (2009).
28. J. B. Kruskal and M. Wish, *Multidimensional Scaling* (Sage, Thousand Oaks, 1978).
29. L. Zhang, R. Rousseau and W. Glänzel, *J. Assoc. Inf. Sci. Tech.* **67**, 1257 (2016).
30. N. Askin and M. Mauskapf, *Am. Soc. Rev.* **82**, 910 (2017).