https://doi.org/10.1140/epjds/s13688-019-0214-8

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Novelty and influence of creative works, and quantifying patterns of advances based on probabilistic references networks

Doheum Park[1], Juhan Nam[1] and Juyong Park[1,2*]

*Correspondence:
juyongp@kaist.ac.kr
[1] Graduate School of Culture Technology, Korea Advanced Institute of Science & Technology, Daejeon, Republic of Korea
[2] Sainsbury Laboratory, University of Cambridge, Cambridge, United Kingdom

Abstract

Recent advances in the quantitative, computational methodology for the modeling and analysis of heterogeneous large-scale data are leading to new opportunities for understanding human behaviors and faculties, including creativity that drives creative enterprises such as science. While innovation is crucial for novel and influential achievements, quantifying these qualities in creative works remains a challenge. Here we present an information-theoretic framework for computing the novelty and influence of creative works based on their generation probabilities reflecting the degree of uniqueness of their elements in comparison with other works. Applying the formalism to a high-quality, large-scale data set of classical piano compositions–works of significant scientific and intellectual value–spanning several centuries of musical history, represented as symbolic progressions of chords, we find that the enterprise's developmental history can be characterised as a dynamic process composed of the emergence of dominant, paradigmatic creative styles that define distinct historical periods. These findings can offer a new understanding of the evolution of creative enterprises based on principled measures of novelty and influence.

Keywords: Novelty; Influence; Computational creativity; Probabilistic reference; Network; Paradigm shift; Classical music

1 Main text

Stories of how creative enterprises–science, technology, and art being principal examples–have evolved are often filled with tales of revolutionary, triumphant "Eurekas" that usher in a new era: Einstein's theory of relativity, Kekulé's determination of the structure of benzene, Tesla's invention of the alternating current (AC) motor, and Brunelleschi's invention of the linear perspective in art are widely-cited examples [1]. But recent studies have discovered that in reality the evolution of a creative enterprise is driven by innovations–achievements based on new ideas and practice–on many 'scales' of significance [2–4], rather than only by those that become parts of a legend or folklore. In order to understand this important phenomenon properly, we must ask why innovations are so valued in human society, and what are the characteristic patterns of their emergence and impact. Recent scientific studies offer some clues: first, studies on human and animal brains have found their innate preference for new stimuli [5, 6]; and second, data-driven studies on citation networks and impacts of scientific papers and patents have shown that novelty is often a key feature in influential scientific knowledge and technological systems [7, 8]. This is also true in cultural creations such as music, where continual experimentations of musical elements (e.g. notes, chords, and rhythm) and compositional rules (e.g. modes, scales, forms, and tonality) have long been recognised as instrumental for innovation throughout history [9, 10]; As a scientific publication is composed of results from experiments and calculations put into words and figures, a musical composition is composed of results of experimentation put down into notes, chord transitions, tempo markings, etc. To achieve further progress in answering those questions quantitatively and to fully grasp the role of innovation in the advancement of a creative field, however, there still remains the challenge of how to quantify the novelty of a creative work. The ability to do so could be very useful in identifying novel works and how a creative form develops dynamically in time. In this work we introduce a foundational information-theoretic framework for computing the novelty of a creative work utilising its mathematical representation as a set of correlated elements and the past, prior works with which it is compared. We also show that it allows us to define the influence of a creative work on later ones, allowing us to identify the most followed or referenced works that could be understood as having laid the groundwork of the styles found among the works of a given time.

1.1 Model and formalism

In order to compute the novelty of a creative work, we first consider the fact that any new creative work—be it a scientific paper, a technological patent, or a musical composition—contains the familiar, 'conventional' elements that can be found in known older works, and the unfamiliar, 'novel' elements that have not [4, 7, 8]. Intuitively then a work that features a larger novel-to-conventional ratio of elements could be considered more novel, and vice versa. How can one tell if an element is conventional or novel? In some form of creative works, notably research publications and patents, the information is explicitly given in the form of a reference or a citation, represented as solid arrows in Fig. 1(A). The study of citation networks with creative works as nodes and the citations as directed edges with the adjacency matrix

$$\mathbf{A} \equiv \{A_{ij}\}, \tag{1}$$

with $A_{ij} = 1$ when paper i is cited by j and 0 otherwise, is a much-studied type in network science [7, 11–13]. A citation network defined in this manner, while straightforward and easy to visualise, possess conceptual issue that need to be discussed: Since it relies solely on self-reporting by the creator subject to human shortcomings such as faulty memory and bias, it could be incomplete and unreliable due to missing (unreported) citations, represented as dotted lines in Fig. 1(A). Also, such citation information, incomplete as it may be, is rarely provided in many other types of written works as well as most cultural works such as literature, music, and art. A more reliable and general method would be then to make a direct comparison between an older work and a new one to identify common elements. If there exists one, it would be an indication that the older work may have been referenced in the creation of the new one. We say 'may have been' because the shared element could have been taken from a different work (either known, or unknown to us

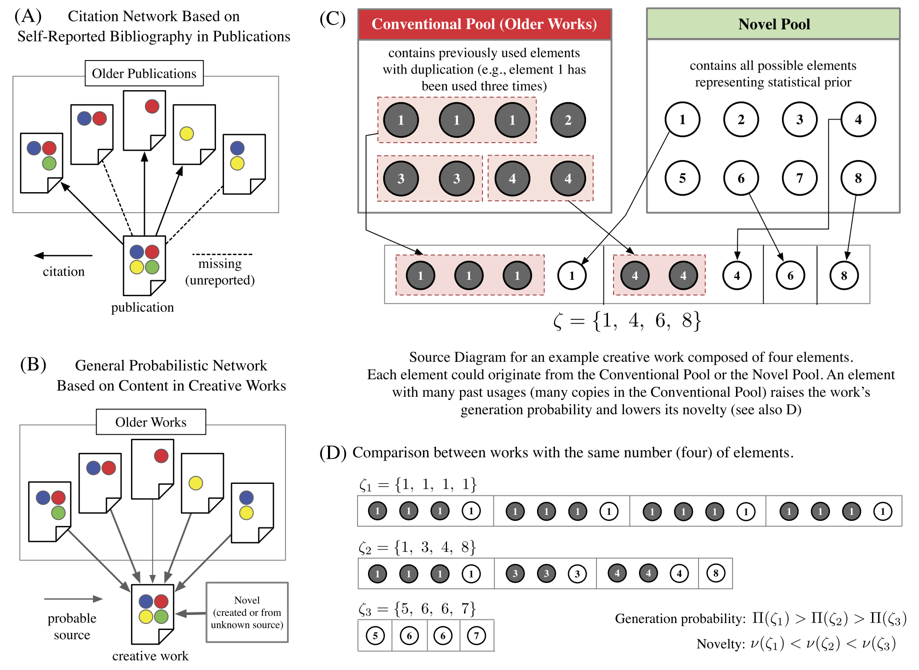
(A) Citation Network Based on
Self-Reported Bibliography in Publications
Older Publications
citation
publication
missing
(unreported)
(B) General Probabilistic Network
Based on Content in Creative Works
Older Works
probable
source
creative work
Novel
(created or from
unknown source)
(C)
Conventional Pool (Older Works)
contains previously used elements
with duplication (e.g., element 1 has
been used three times)
Novel Pool
contains all possible elements
representing statistical prior
1 1 1 2
3 3 4 4
1 2 3 4
5 6 7 8
1 1 1 1 4 4 4 6 8
$\zeta = \{1, 4, 6, 8\}$
Source Diagram for an example creative work composed of four elements.
Each element could originate from the Conventional Pool or the Novel Pool. An element
with many past usages (many copies in the Conventional Pool) raises the work's
generation probability and lowers its novelty (see also D)
(D) Comparison between works with the same number (four) of elements.
$\zeta_1 = \{1, 1, 1, 1\}$
$\zeta_2 = \{1, 3, 4, 8\}$
$\zeta_3 = \{5, 6, 6, 7\}$
Generation probability: $\Pi(\zeta_1) > \Pi(\zeta_2) > \Pi(\zeta_3)$
Novelty: $\nu(\zeta_1) < \nu(\zeta_2) < \nu(\zeta_3)$

Figure 1 Reference and Information Network formalism for Novelty of Creative Works. (**A**) A citation network for publications relies on self-reported bibliography connecting a new work to older works (solid arrows). Various factors as human error and bias may cause missing links (dotted), indicating loss of information. (**B**) A general probabilistic network is constructed by connecting two works that share a common element that represents a probabilistic reference. Unlike a citation network, therefore, no known source is omitted due to error or bias. Furthermore, since an element can be truly invented or come from unknown sources, the Novel Pool (statistical prior) is also considered. (**C**) Calculating the generation probability of a creative work given past works. The generation process is modeled as choosing elements from the Conventional Pool (CP) that contains previous used elements (with duplication, so that often used elements are common in the CP as well) or from the Novel Pool (NP) that represents 'inventing' the element or referencing an unknown source. Setting NP to contain a fixed number of each possible element corresponds to the Maximum A Priori estimator and additive Laplace smoothing, Eq. (6). The probability of choosing an element is proportional to its multiplicity in the combined pool. Consider an example work of length four, $\zeta = \{1,4,6,8\}$. Here element 1 is the most common (four copies total, three from CP and one from NP), whereas 6 and 8 are the least so (one from NP only). The total number of those elements in the pool represent the generation probability of ζ. (**D**) Illustration by comparison between three example creative works ζ_1, ζ_2, ζ_3 composed of the same number of elements (four). ζ_1 is composed of the most used element 1, making it the most conventional (least novel), whereas ζ_3 is composed of unused elements only, giving it the highest novelty.

presently), or been 'invented' by the creator oblivious to a previous usage. This indicates another shortcoming of the citation network that characterises a publication based solely on given relational information: The 'combination' of conventional and new elements—an essential trait of a creative work—cannot be properly considered, as it takes into account only the conventional elements in the work. The multiplicity of possible sources, known or unknown, suggests the need for a *probabilistic* model of reference that assigns the probabilities that an element in a work has been taken from a known older work, an unknown (lost) older work, or been invented. This is depicted in Fig. 1(B) where the set of known works Ω is considered a probable source of the generation of ζ, along with the other possibilities labeled as the 'novel' source (since they are new to us). We later show that the latter can be implemented mathematically as an uninformed prior.

Figure 1(C) shows in detail how an example work $\zeta = \{1,4,6,8\}$ can be originating from the known works (constituting the **Conventional Pool** of elements) or the novel sources (constituting the **Novel Pool** of elements). For instance, elements quite common in the conventional pool (such as 1 and 4) raise the conventionality (and lower the novelty) of ζ, whereas rarer or nonexistent elements (such as 6 and 8) raise its novelty (or lower its conventionality). This is consistent with the intuitive meaning of conventional and novel as being common and familiar (conventional) and rare and unfamiliar (novel). As further examples, in Fig. 1(D) we compare three hypothetical works ζ_1, ζ_2, and ζ_3 where the Green-to-Red ratio of elements represents each work's novelty $\nu(\zeta_1) < \nu(\zeta_2) < \nu(\zeta_3)$ which is in the opposite order of the generation probabilities $\Pi(\zeta_1) > \Pi(\zeta_2) > \Pi(\zeta_3)$ represented by the volume under the generation process of Fig. 1(B).

Formally, from our model of Fig. 1 we can represent the generation probability $\Pi_\Omega(\zeta)$ of a creative work $\zeta = \{e_1, e_2, e_3, \ldots, e_m\}$ as the probability of choosing its elements from the element pools given by

$$\Pi_\Omega(\zeta) = \prod_{i=1}^{m} \pi_\Omega(e_i), \tag{2}$$

where $\pi_\Omega(e_i)$ is the selection probability of element e_i. Since a smaller generation probability Π means that a work is less expected and therefore more novel, the novelty of the work is be a decreasing function of Π. We thus define the novelty $\nu(\zeta)$ as the log inverse Π_Ω, normalised by the work ζ's length m (we take log to mean $\log_{10}$ in this work):

$$\nu(\zeta) \equiv \frac{1}{m} \log \frac{1}{\Pi_\Omega(\zeta)} = \frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{\pi_\Omega(e_i)}. \tag{3}$$

This form shows a clear connection to information theory. In information theory, the log of inverse probability of an event is called its information content that measures the unexpectedness (degree of surprise) of an event [14] (measured in bits had we used log of base 2). Therefore, the novelty of a given work ζ is defined to be the average unexpectedness of its elements (the normalisation m^{-1} is necessary because without it any new work can be made arbitrarily novel by lengthening it), in agreement with its intuitive meaning. Note that Eq. (3) is not to be confused with the Shannon entropy $H(\Omega) = \sum_{i\in\Omega} p_i \log(1/p_i)$ of the ensemble Ω which is the mean information content (degree of surprise) from a random sample i chosen from the ensemble with probability p_i. The Shannon entropy is therefore a characteristic of the whole ensemble, whereas $\nu(\zeta)$, Eq. (3), is a characteristic of a specific work ζ.

Although we have above argued for the appeal of novel works to living beings and their value for progress, it is unlikely that novelty alone is a sign that the work is of any value; if it were, one could simply assemble elements not found in the older works (or ask a chimpanzee that has just completed composing a piece of 'literature' on a typewriter to play the piano now, practically to the same effect), and claim to have created the most valuable work. In addition to quantifying a work's difference from the past, it is necessary to gauge a work's value via how much it has influenced the posterity, in other words how much the later works have referenced it. To do so, we can again use Eqs. (2) and (3) to define the *influence* of an earlier work or a set of earlier works ω (for instance, the works by a specific creator) on a later work ζ. Intuitively, we can suspect influence of an earlier work when ζ shares common elements with it. We say only 'suspect' because, as before, the shared elements could have been taken from a different work or invented (i.e. from the novel pool) by ζ's creator. What we can be more certain of, however, is the *lack* of influence when ζ shares no element ω, meaning that all the elements of ζ can only be found in $\overline{\omega} \equiv \Omega - \omega$ or the novel pool. Mathematically, we can interpret the *discrepancy* between the full generation probability $\Pi_{\Omega}(\zeta)$ and the reduced one $\Pi_{\overline{\omega}}(\zeta)$ as ω's influence on ζ; if the two are identical, for instance, it means that ω has had no effect on ζ, i.e. no influence. In other words, influence is the share of the generation probability of ζ that ω is accountable for. More specifically, we define $\eta_{\omega}(\zeta)$ as

$$\eta_{\omega}(\zeta) \equiv \frac{1}{m} \log \frac{\Pi_{\Omega}(\zeta)}{\Pi_{\overline{\omega}}(\zeta)} \geq 0, \tag{4}$$

a form more consistent with Eq. (3). As stated above, the equality $\eta_{\omega}(\zeta) = 0$ (no influence) holds when ω shares no elements with ζ, i.e. has not contributed at all to its generation. Note that this is also a characteristic value for a specific set-work pair (ω, ζ) based on their composition.

With these quantitative measures of novelty and influence of creative works, we tackle the following important questions regarding the advancement of creative enterprises: How do the novelty and influence change over time? How do they correlate, i.e. does novelty lead to influence and later recognition? How do these characterize the evolution patterns of a creative field? We illustrate our methodology by analysing a representative example of a creative enterprise with a long history of high scientific and intellectual value, Western classical piano music.

2 Novelty and influence in classical piano music

We study Western classical piano music from the so-called common practice period (circa 1700–1900 CE) chosen for the following advantages: High scientific and cultural significance, widely credited for having produced many fundamental musical styles that are influential today; A rich body of musicological understanding available from traditional research that could be compared with new, alternative approaches such as ours; And the abundance of high-quality data. The availability of large-scale musical databases and advances in scientific, analytical methods continue to enable novel and interesting findings on their properties [10]. Recent examples include researches on the topology and dynamics of the networks of musicians for the discovery of human and stylistic factors in the creation of music [15–20] and stylometric analyses of music that lead to corroborations or fresh challenges to established musicological understanding [21–25].

Using our framework we start by computing the level of novelty in musical compositions and composers, and study how they relate to the known characteristics of music at a given point in history. We then compute their influence on later times and how it can be used to characterize the evolution of compositional styles throughout history. The first step in using the formalism of Eqs. (3) and (4) is representing music as a set of elements, in other words modeling. Since modeling a system is an abstraction process that necessarily leaves out some real features of the system, it is ideal to retain the most sensible, relevant ones that also suit the modeler's interests. For instance, for written works such as literature, scientific publications, etc., they could be words or groups of words such as the n-grams [7, 8, 26], and for paintings they could be colours and shapes [27, 28]. Here we model a musical composition as a temporally ordered set of simultaneously played

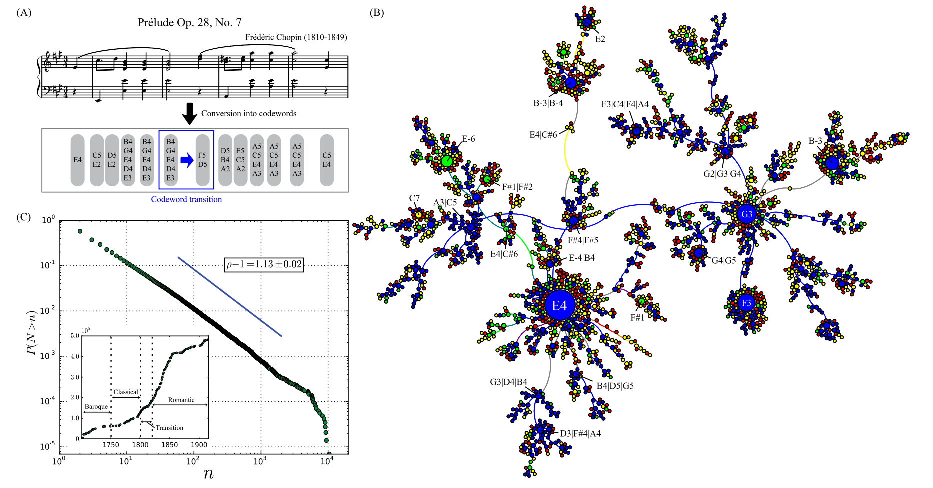

Figure 2 (**A**) A musical score can be converted to a sequence of codeword (simultaneously-played notes) transitions (blue box). (**B**) The backbone of the network of codeword transitions in our data. Only 2267 out of 144,183 codewords (1.5%) are shown. The node radius indicates the number of transitions into and out of the corresponding codeword, while the edge width indicates the number of the corresponding transition. The node colour indicates the period when the corresponding codeword first appeared (blue-Baroque, green-Classical, yellow-Transition, red-Romantic). (**C**) The cumulative distribution of the occurrences of the codewords and the cumulative number of unique codeword transitions ever used (inset). The distribution exhibits a highly-skewed, power law-like behavior with power exponent $\rho = 2.13 \pm 0.02$ established early in history (Fig. S1).

nodes or *codewords*. For the actual element we take the *codeword transition*, the bigram (2-gram) of codewords. They are shown in Fig. 2(A) with the beginning of one of Chopin's preludes as an example. While our methodology can be applied in a clear and straightforward manner to analysing musical compositions, we note that other aspects of music such as structure, tempo, instrumentation, etc. are also important in music. Our primary focus on codeword transitions here are based on the importance of harmony and melody in the Western classical music tradition [29] and the fact that for this paper we will be studying the piano, but for a more complete and useful modeling of music those elements will need to be incorporated in the future, and later we discuss some recent developments therein. We also note that our definition of a codeword retains all the original information on octaves and the keys in which the works were composed, resulting in a more complete and truthful representation than the one used in Ref. [22] where only the pitch class was considered (i.e. discarding the octave information; for instance, F4 and F5 were considered both F) and the keys were unified to the C scale.

Our data set consists of MIDI (Musical Instrument Digital Interface) files collected from Kunst der Fuge (www.kunstderfuge.com) and Classical Piano MIDI (www.piano-midi.de) archives of 900 classical piano works by 19 prominent composers from the common practice period spanning the Baroque (c. 1700–1750), Classical (c. 1750–1820), Classical-To-Romantic Transition (c. 1800–1820), and Romantic (c. 1820–1910) periods, featuring Johann S. Bach and Georg F. Handel of the Baroque era, and Maurice Ravel of the late Romantic era. The composers and their works are in the Additional file 1 (SI Dataset 2). The MIDI files were converted into musicXML format via MuseScore2 software and chordified using Music21, a python library toolkit for computer-aided musicology [30]. The `chordify` method in Music21 converts a multiple-part complex musical score into a series of simultaneous notes as visualised in Fig. 2(A). Since each codeword transition is a directed dyad, they can be collectively visualised as a network whose backbone is shown in Fig. 2(B). The cumulative distribution of the number of occurrences of the codewords is shown in Fig. 2(C), and approximates a power law with exponent $\rho = 2.13 \pm 0.02$, indicating significant disparities in popularity between codeword transitions. Although such a pattern is established early in history (Fig. S1), the number of unique codeword transitions ever used also constantly increases in time (inset of Fig. 2(C)), with the highest rate of increase observed during the Romantic period.

We now compute the novelty and influence of musical compositions. Writing a composition ζ as a sequence of codewords $\zeta = \{\gamma_1, \gamma_2, \dots, \gamma_m\}$ the generation probability of ζ is given by the first-order Markov chain

$$\varPi_\Omega(\zeta) = \pi_\Omega(\gamma_1)\pi_\Omega(\gamma_1 \to \gamma_2) \cdots \pi_\Omega(\gamma_{m-1} \to \gamma_m), \tag{5}$$

For π_Ω we employ the Maximum A Priori (MAP) estimator [31] commonly used in Markov chains, given as

$$\pi_\Omega(\gamma_i \to \gamma_j) = \frac{z(\gamma_i \to \gamma_j) + \alpha_0(\gamma_i \to \gamma_j)}{\sum_{\gamma \in \varGamma}(z(\gamma_i \to \gamma) + \alpha_0(\gamma_i \to \gamma))}, \tag{6}$$

where $z(\gamma_i \to \gamma_j)$ is the number of the $\gamma_i \to \gamma_j$ transition in the conventional pool Ω and $\alpha_0(\gamma_i \to \gamma_j)$ is the prior representing the novel pool in our scheme. The form can also be viewed as a type of additive Laplace smoothing. When α_0 is a constant it is also called the uninformed prior, and interpreting the prior as the novel pool allows us to make a graphical representation in Fig. 1(C) with $\alpha_0 = 1$ meaning the novel pool contains exactly one copy of each possible transition. $\varGamma$ is the codeword space. The probability of the first codeword $\pi_\Omega(\gamma_1)$ is similarly $\pi_\Omega(\gamma_1) = (z(\gamma_1) + 1)/(\sum(z(\gamma) + 1))$, where $z(\gamma_1)$ is the number of occurrences of γ_1 as the first codeword in Ω. Plugging this into Eq. (3), we obtain the novelty

$$\nu(\zeta) \equiv \frac{1}{m} \log \frac{1}{\varPi_\Omega(\zeta)} = \frac{1}{m}\left[\log \frac{1}{\pi_\Omega(\gamma_1)} + \sum_{k=1}^{m-1} \log \frac{1}{\pi_\Omega(\gamma_k \to \gamma_{k+1})}\right]. \tag{7}$$

2.1 Historical and psychological novelty

When computing the novelty of Eq. (7), we are free to choose Ω, the reference set of previous works that determine the conventional pool. A straightforward choice of Ω would be all known works that preceded ζ in history. This was aptly given the name **historical novelty** (H-novelty) in Artificial Intelligence (AI) research circles [1], and represents a given work's novelty within the entire history of the field up to its creation. Another interesting choice of Ω contains all the previous works by the very creator of ζ. The resulting novelty is named **psychological novelty** (P-novelty) [1] that represents, for instance, the degree of improvement in a new version of an algorithm or a machine over its previous versions. Applied to our data it would show how a composer evolves in compositional style against his own past works [32].

We show in Figs. 3(A) and (B) the cumulative distributions of the H- and P-novelties of the piano works in our data for each period. Of the four, the Classical compositions tend to score low in both novelties, showing that many past conventions were reused both historically and psychologically (see Fig. S2 for the H- and P-novelty scores of the pieces

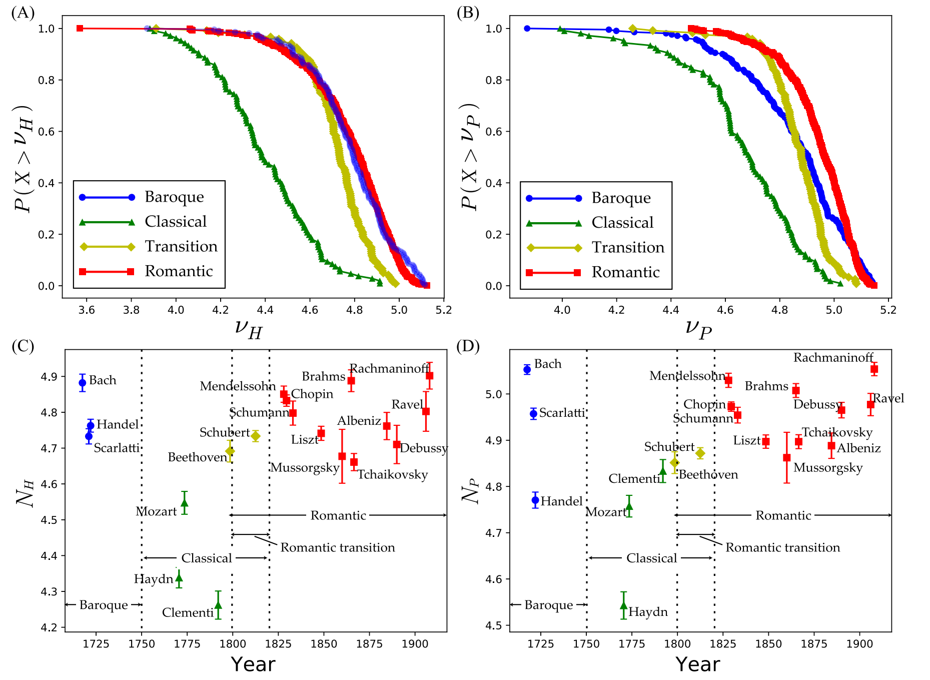

Figure 3 The H-novelty (left) and P-novelty (right) of the piano compositions (top) and the composers (bottom). (**A**) The cumulative distribution of the H-novelty scores ν_H of the works. The median and mean values are (4.80, 4.78) for the Baroque, (4.38, 4.40) for the Classical, (4.73, 4.729) for the Transition, and (4.82, 4.78) for the Romantic periods. (**B**) The cumulative distribution of the P-novelty scores ν_P of the works. The median and the means are (4.90, 4.86) for the Baroque, (4.69, 4.66) for the Classical, (4.88, 4.87) for the Transition, and (4.97, 4.94) for the Romantic periods. (**C**) & (**D**) The novelty N_H and N_P of the composers (defined as the mean of ν_H and ν_P of their works). A composer's position on the *x*-axis (year) is the midpoint between his birth and death years. One should note that the conventional pool of elements is smaller for baroque composers, which could skew the H-novelty to look higher. However, the P-novelty does not suffer from limitation, and Bach and Scarlatti still have high P-novelty values.

over time). The novelties of the composers (given by the average of their works') noted N_H and N_P are shown in Figs. 3(C) and (D). We note that the confidence in the high H-novelty of the Baroque composers should be low due to the much smaller conventional pool than other periods. The raised H-novelty of the Romantic composers, on the other hand, should be considered more impressive since it is achieved against the largest conventional pool. The high level of P-novelty shows Romantic composers having also actively introduced diverse and new codeword transitions throughout their careers. This is in clear agreement with the widely-accepted thesis that credits Romantic composers with having broken many accepted musical conventions and diligently conducting personal experimentation with new combinations of pitches [33]. The H- and P-novelties are generally positively correlated throughout, with the Spearman correlations equal to 0.820 ± 0.013 for the compositions and 0.827 ± 0.113 for the composers, respectively, meaning that pursuing novelty involved deviating from both the others and oneself (Fig. 4). The most notable outlier from this trend is Muzio Clementi (1752–1832) whose H-novelty is significantly lower than what his P-novelty would suggest, as shown in Fig. 4(B). This means that while he produced works distinct from his own earlier works (even more so than Handel, Mozart, and Haydn, and on par with Beethoven), they as a whole would sound conventional when compared with other composers'. This may be a quantitative corroboration

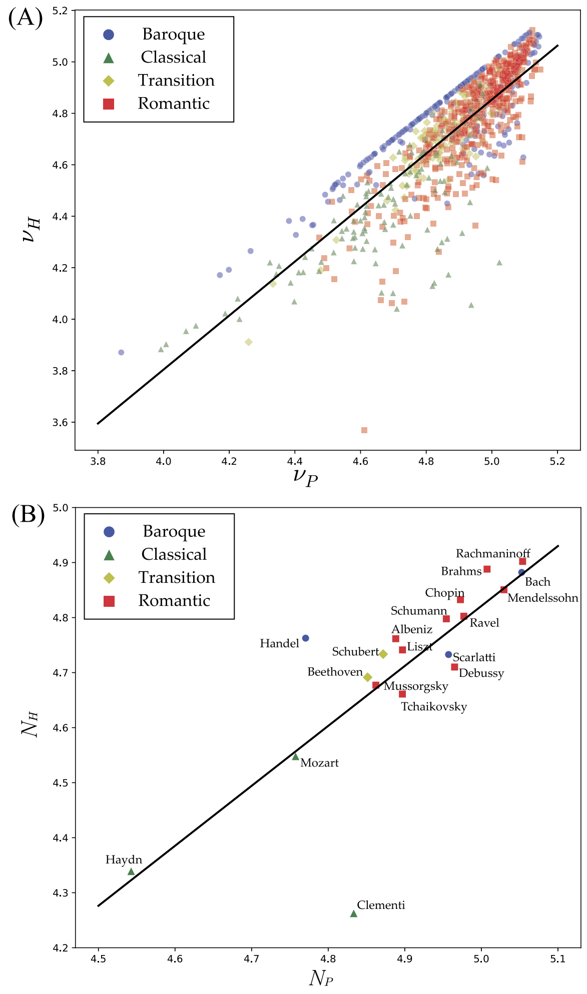

Figure 4 (**A**) Scatter plot of the H- and P-novelty scores of the piano works, with a high level of correlation (Spearman correlation 0.82 ± 0.01). (**B**) The H- and P-novelty scores of composers (Spearman correlation 0.83 ± 0.11). A notable outlier is Muzio Clementi who shows a significantly small H-novelty given P-novelty.

of the explanation behind the common assessment of Clementi that in his time his reputation rivaled Haydn's among his contemporaries, but languished for much of the 19th century and beyond [34]; The diversity of codeword transitions that he employed in his works (reflected in the high P-novelty) could have been the source of high reputation during his lifetime, but as time passed his works failed to distinguish themselves from others (reflected in the low H-novelty) and caused his loss in stature.

2.2 Influence and shifts in dominant styles

While novel achievements are indispensable for the progress and growth of a creative enterprise, our results above suggest that novelty alone would not cause one to be considered 'the greatest': Beethoven, for instance, stand among the lower half in computed novelty. This is in line with many recent research findings that a creative work's impact on its posterity does not depend solely on the degree of its novelty, and how it builds on tradition is also important [4, 7, 8]. Musical composition would be no exception: Past works exert influence on the future by serving not only as training material for new composers, but also by inspiring new works or lending themselves to be tweaked and transformed into new original works [1, 10]. Even mimicry or imitation, normally associated with subpar works lacking in originality and artistic value, can sometimes occur in renowned masters' works and gain recognition: Franz Liszt, a leading Romantic-era composer, admired Beethoven so much that in a famous deed of homage he transcribed Beethoven's complete symphony cycle into the piano [35] now considered a significant and influential achievement in its own right. These observations suggest that a sensible definition of 'influence' of a work would be the the degree to which it has been referenced by later works as in Eq. (4).

To compute $\eta_\omega(\zeta)$ of Eq. (4), the influence of composer ω on ζ, we start by rewriting $z(\gamma_i \to \gamma_j)$, the number of $\gamma_i \to \gamma_j$ transitions in Ω, in Eq. (6) as

$$z(\gamma_i \to \gamma_j) = z_\omega(\gamma_i \to \gamma_j) + z_{\overline{\omega}}(\gamma_i \to \gamma_j), \tag{8}$$

where z_ω is the number of instances of the transition used by ω, and $z_{\overline{\omega}}$ is that by all the other composers before ζ. Then $\Pi_\Omega(\zeta)$ becomes

$$\Pi_\Omega(\zeta) = \frac{(z_\omega(\gamma_1) + z_{\overline{\omega}}(\gamma_1) + 1)}{\sum_{\gamma\in\Gamma}(z(\gamma) + 1)} \times \frac{(z_\omega(\gamma_1 \to \gamma_2) + z_{\overline{\omega}}(\gamma_1 \to \gamma_2) + 1)}{\sum_{\gamma\in\Gamma}(z(\gamma_1 \to \gamma) + 1)} \times \cdots. \tag{9}$$

Eliminating all z_ωs in the numerator, we obtain

$$\Pi_{\overline{\omega}}(\zeta) = \frac{(z_{\overline{\omega}}(\gamma_1) + 1)}{\sum_{\gamma\in\Gamma}(z(\gamma) + 1)} \times \frac{(z_{\overline{\omega}}(\gamma_1 \to \gamma_2) + 1)}{\sum_{\gamma\in\Gamma}(z(\gamma_1 \to \gamma) + 1)} \times \cdots. \tag{10}$$

After computing the influences $\{\eta\}$ between all 7298 eligible composer–composition pairs (self-influences were excluded) we plot each composer's mean influence on the works created at any given time t (± 10 years for smoother curves), shown in Fig. 5. During the

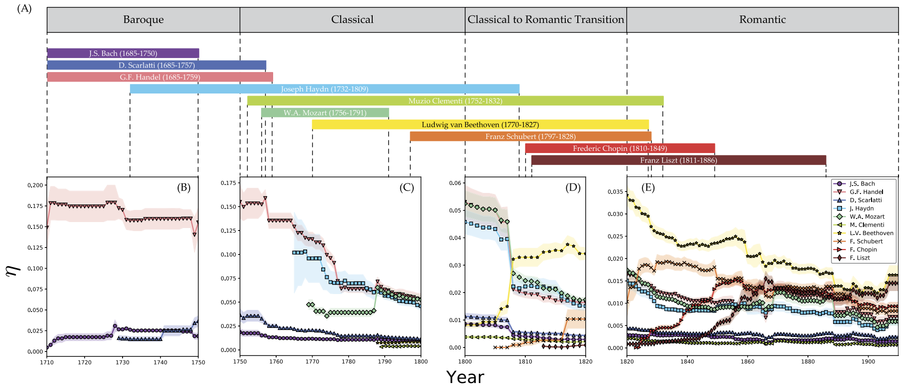

Figure 5 The influence of composers and shifts in paradigmatic, dominant styles. (**A**) The common period designations and the living years of ten major composers in our data. (**B**)–(**E**) The mean influence η of the composers on the works composed at a given time t (±10 years). Each period is distinguished by the emergence of newly dominant composers that indicate paradigmatic shifts in composition styles, and provide quantitative support for period designations. (**B**) During the Baroque period Handel exerts a dominant influence on other composers. (**C**) In the Classical period initially Scatlatti's influence increases, while Handel's influence begins to wane. Then Haydn and Mozart's influence rival Handel's. (**D**) The Classical-to-Romantic Transition period is characterised by Beethoven who overtakes the most influential ones from the previous period (Handel, Haydn, and Mozart). (**E**) The Romantic period also witnesses the emergence of newly highly influential composers such as Schubert, Chopin and Liszt.

Baroque period (B) Handel is the most influential, indicating that his codeword transitions were often also used at a later time by his contemporaries Bach and Scarlatti, whereas the opposite did not occur as frequently. More interesting patterns can be found when we observe the rise and fall of the composers' influences over time. Since a high influence means that later works share common elements, we can interpret such rise and fall of composers' influence as indicating shifts in compositional style, and providing a quantitative justification for the distinct period labels. Let us examine, as a start, the Baroque and the Classical periods in Figs. 5(B) and (C). While Handel maintains his dominant influence until around the mid-Classical period, we identify two notable patterns: First, Scarlatti overtakes Bach in influence shortly before the Classical period, in agreement with the well-acknowledged significance of Scarlatti on the Classical period [36]; Second, Haydn and Mozart emerge during the Classical period with a high influence, soon rivaling Handel's. Similar dynamics–the clear rise and emergence of a new leading influential figure and therefore dominant 'style', reminiscent of Kuhn's so-called *paradigm shift* [2]–are observed in subsequent periods. The Classical-to-Romantic transitional period (Fig. 5 D) is characterised by the emergence of Beethoven whose historical significance [37] is clearly shown. Beethoven's high influence in this period shows his younger contemporaries adopting his codewords more willingly than any other predecessor's (Figs. S6(C) and (D)) that continues well into the Romantic period. Also, from Eq. (4), we see that being referenced by a highly-novel composer leads to high influence, as high novelty means referencing uncommon elements, and so the one referenced is credited with more influence. This is likely why Beethoven, referenced by highly novel Romantic composers, has a high influence score. Then, through a similar mechanism, during the Romantic period new composers such as Schubert, Chopin, and Liszt rise in influence to rival or overtake Mozart and Beethoven (Fig. 5 E), befitting their reputation as of finally eclipsing those "classical sounds" and establishing many essential repertoire now permanently associated with the piano [37].

3 Discussion

This work presents a general mathematical framework for computing the novelty and influence of creative works based on the degree of shared elements between past and future works. Novelty measures how different a work is from the past, representing originality and unpredictability of generation. Influence measures how much a work has been referenced in the future, representing its success and impact as an inspiration for future creations. While originality and success are both important characteristics of meaningful creative works, they do not correlate perfectly. That Handel was less novel than Bach and many others but had more influence on Classical and Romantic composers (Figs. 3 and 5) is a good example. Similarly, Beethoven, Schubert, and Liszt were less novel than Mendelssohn and Schumann (Fig. 3), but eventually came to exert more influence and inspire more piano music to follow (Fig. S5). The separation between novelty and influence is particularly auspicious in the case of music from the Classical period (especially Mozart): Mozart is shown to have used fewer novel codewords per se and opted to use the conventions from the Baroque period, but his works nevertheless had enough high artistic value that he gained much influence in the future. This is another agreement between our findings and traditional musicology that characterises the Classical period as one that "values shared conventions, rational restraint and the playful exploitation of established constraints" [10]. This contrasts with the composers of the later Romantic period who introduced new elements at a faster pace (Fig. 3) and again corroborates the traditional musicological assessment of their "pursuit of the value of being individual, peculiar and original" [10, 37, 38].

We note that, while we employed the simplest first-order Markov model of codeword transitions to model music, the framework is general enough for a higher-order Markov model or related techniques such as Hidden Markov Model (HMM) and neural networks that have been previously applied to analysis of text and music [31, 39–44]. Higher-order Markov modeling shows a broad agreement with our main findings using first-order Markov, showing further robustness of our model and analysis (Figs. S7–13). One notable extra finding from higher-order Markov involves Debussy whose most significant innovation is believed to have been in the use of non-traditional scales. Also, since novelty and influence were developed using concepts from information theory, there could be many future possibilities for utilising other useful quantities such as cross entropy and others to define relationships between creative works systematically, which would further enable interesting theoretical advances.

We finally discuss the potential issues of using curated data such as ours and how they are addressed in our framework. One can justifiably point out that throughout the common practice era considered in our analysis there existed many active composers not included in our data set but who nevertheless likely referenced and influenced one another. Note that this situation is not unique to our data, but is becoming increasingly common in an era when interesting data are collected from many open real-world systems where one cannot easily expect them to be as complete or comprehensive as those from designed experiments conducted in highly-controlled laboratory environments. It is more important and practical, then, to deal with the situation by incorporating appropriate mechanisms in the methodology and understanding the nature of the data. First, in our formalism, incomplete data results in underrepresented or missing elements in the Conventional Pool. But our formalism addresses this issue via the Novel Pool that gives some weight to unknown or as yet unknowable cases via the uninformed prior, a well-established, unbiased method employed in statistics in the absence of usable information. Additionally, they can be updated whenever new information becomes available in a straightforward manner. Second, our data comprises works that are at the time of this study the most highly regarded, and most often studied and performed, implying that of all imaginable data sets of similar size it would be the most commensurate with the meaning of novelty and influence: Since they are by definition based on what is available for comparison, our data would be closer to a modern listener's true experience than others comprising obscure or less popular works, rendering it both desirable and useful given the conditions. Lastly, a potential temporal bias that gives the early composers advantage in the case of the H-novelty (but not the P-novelty) would need to be studied to fully solve the 'cold-start' by some type of temporal scaling or a continued discovery of earlier data.

4 Conclusion

The availability of a quantitative computational methodology confers the ability to confirm or challenge existing knowledge and understanding about a system in a statistically robust manner, and to find more detailed and advanced answers to both long-standing questions and new ones. In this paper we proposed a framework for quantifying the novelty of a creative work and the influence between those produced at different times based on the intrinsic compositions of the works. As an illustration of the methodology at work, we applied it to the development of classical music by using 900 classical piano compositions that cover the common practice era in the western musical tradition. For the intrinsic element of music we focused on the codeword transitions to measure the novelty and influence of the compositions and composers. From the use of codeword transitions over time, we found that commonly designated periods corresponded well to the emergence of newly influential composers indicating notable paradigmatic shifts in styles. In addition to a broad agreement with conventional understanding of the characteristic of periods and composers, an interesting finding was that being more novel, i.e. more willing to break from convention, did not necessarily translate to being influential on the posterity. This means while novelty is still necessary in a creative endeavor—high-novelty composers in our data set are undoubtedly universally recognised masters of the form themselves—it cannot account for all the creative, artistic qualities that facilitated those codeword transitions that were more widely transmitted to later generations.

This suggests a future research direction employing a more elaborate modeling of codeword transitions and other elements of music. Possibilities in the former category include the change of the number of notes [10], the tonality [45, 46], melody [47–49] and the chord progression [50], to name a few. Possibilities in the latter category include the rhythmic structure of music that is recently gaining increased attention [21, 32] and the global structure of a composition, given the common assertion that the most significant innovation in the piano music during the Classical period was the establishment of the sonata form [37] which may have little to do with the codeword transitions. To study such macroscopic structures, note-level features would be insufficient, and even higher-order Markov models are limited as they focus on local dependencies in sequential data [51]. A possible approach would be to summarise the codewords onto segment level (e.g., bar or phrase) and to find global features such as self-similar structures [52]. More recently, machine learning algorithms based on neural networks (e.g. LSTM-RNN and Deep Attention Networks) have shown promise in modeling long-term musical patterns [53]. We believe that such pattern-detection methods could be very useful in deeper future studies. An extension beyond the piano is also an obvious possibility, as many composers we considered were prolific in other forms including Haydn who is also very well known for his chamber music and symphonies [54]. Given the continued importance of music as a creative cultural form, analysis of modern-day composers would also be necessary and further illuminating.

Also, there are implications for the psychological study of novelty as well. It has been known in optimal theory of novelty that the positive acceptance (also called the “hedonic value”) of novelty follows the so-called Wundt curve that increases initially but decreases after a peak, indicating that too much novelty can be off-putting to humans [55, 56], which can potentially explain why some less novel composers can be well-received and become influential. If a more comprehensive dataset including less appreciated composers could be created, the theory could be studied comprehensively.

Given the generality of our methods, we envision our framework proposed here being useful in addressing many questions pertinent to the development of various cultural, creative fields and genres other than music as we have presented here. To do so, much effort will be needed find appropriate discretisation of elements depending on the form. In a narrative with long-term correlations of story elements, for instance, topic modeling could be more appropriate for dividing the text into related units than n-grams of words [57–59].

For pictures, on the other hand, adjacent pixel clusters or motifs could be considered as the elements [60]. We believe such principled scientific approach will permit new understanding of human creativity and the dynamics of the progress of intellectual and cultural products.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-019-0214-8.

Additional file 1. Supplementary information. (PDF 8.4 MB)

Acknowledgements

The authors thank Kyungmyeon Lee for helpful discussions.

Funding

This work was supported by the National Research Foundation of Korea (NRF-2016S1A2A2911945 and NRF-2016S1A3A2925033) and the BK21 Plus Postgraduate Organisation for Content Science.

Abbreviations

CP, Conventional Pool; NP, Novel Pool; MIDI, Musical Instrument Digital Interface; XML, eXtensible Markup Language; MAP, Maximum A Priori; AI, Artificial Intelligence; H-novelty, Historical novelty; P-novelty, Psychological novelty; HMM, Hidden Markov Model.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its Supplementary Information files.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DP, JP conceived and designed the experiments. DP, JP performed the experiments. DP, JN, JP analysed the data. DP, JP contributed reagents/materials/analysis tools. DP, JN, JP wrote and reviewed the manuscript. DP crawled data. All authors read and approved the final manuscript.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 9 June 2019 Accepted: 29 November 2019 Published online: 30 January 2020

References

1. Boden MA (2004) The creative mind: myths and mechanisms. Routledge, London
2. Kuhn TS (2012) The structure of scientific revolutions. University of Chicago Press, Chicago
3. Strumsky D, Lobo J (2015) Identifying the sources of technological novelty in the process of invention. Res Policy 44(8):1445–1461
4. Ackerman JS (1962) A theory of style. J Aesthet Art Crit 20(3):227–237
5. Bunzeck N, Düzel E (2006) Absolute coding of stimulus novelty in the human substantia nigra/vta. Neuron 51(3):369–379
6. Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. Neuron 58(6):967–973
7. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342(6157):468–472
8. Kim D, Cerigo DB, Jeong H, Youn H (2016) Technological novelty profile and invention's future impact. EPJ Data Sci 5(1):1
9. Meyer LB (1957) Meaning in music and information theory. J Aesthet Art Crit 15(4):412–424
10. Meyer LB (1989) Style and music: theory, history, and ideology. University of Chicago Press, Chicago
11. Price DdS (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306
12. Newman ME (2001) Scientific collaboration networks. I. Network construction and fundamental results. Phys Rev E 64(1):016131
13. Newman ME (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys Rev E 64(1):016132
14. MacKay DJ (2003) Information theory, inference and learning algorithms. Cambridge University Press, Cambridge
15. Gleiser PM, Danon L (2003) Community structure in jazz. Adv Complex Syst 6(04):565–573
16. Silva DdL, Soares MM, Henriques M, Alves MS, de Aguiar S, de Carvalho T, Corso G, Lucena L (2004) The complex network of the Brazilian popular music. Phys A, Stat Mech Appl 332:559–565

17. Bae A, Park D, Park J (2014) The network of western classical musicians. In: Complex networks V. Springer, Switzerland, pp 13–24
18. Park D, Bae A, Park J (2014) The network of western classical music composers. In: Complex networks V. Springer, Switzerland, pp 1–12
19. Park D, Bae A, Schich M, Park J (2015) Topology and evolution of the network of western classical music composers. EPJ Data Sci 4:2
20. Bae A, Park D, Ahn Y-Y, Park J (2016) The multi-scale network landscape of collaboration. PLoS ONE 11(3):0151784
21. Levitin DJ, Chordia P, Menon V (2012) Musical rhythm spectra from bach to joplin obey a 1/f power law. Proc Natl Acad Sci 109(10):3716–3720
22. Serrà J, Corral Á, Boguñá M, Haro M, Arcos JL (2012) Measuring the evolution of contemporary western popular music. Sci Rep 2:521
23. Liu L, Wei J, Zhang H, Xin J, Huang J (2013) A statistical physics view of pitch fluctuations in the classical music from bach to chopin: evidence for scaling. PLoS ONE 8(3):58710
24. Wu D, Kendrick KM, Levitin DJ, Li C, Yao D (2015) Bach is the father of harmony: revealed by a 1/f fluctuation analysis across musical genres. PLoS ONE 10(11):0142431
25. Mauch M, MacCallum RM, Levy M, Leroi AM (2015) The evolution of popular music: USA 1960–2010. R Soc Open Sci 2(5):150081
26. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235
27. Kao J, Jurafsky D (2012) A computational analysis of style, affect, and imagery in contemporary poetry. In: Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature, pp 8–17
28. Kim D, Son S-W, Jeong H (2014) Large-scale quantitative analysis of painting arts. Sci Rep 4:7370
29. Powell J (2010) How music works: the science and psychology of beautiful sounds, from beethoven to the beatles and beyond. Hachette, New York
30. Cuthbert MS, Ariza C (2010) Music21: a toolkit for computer-aided musicology and symbolic music data
31. Murphy KP (2002) Learning markov processes. The Encyclopedia of Cognitive Sciences
32. Margulis EH (2014) On repeat: how music plays the mind. Oxford University Press, New York
33. Kravitt EF (1992) Romanticism today. Music Q 76(1):93–109
34. Youngren W (1996) Finished symphonies. Atl Mon 227(5):104
35. Searle H (1980) Liszt, franz. In: The new grove dictionary of music and musicians, pp 28–74
36. Taruskin R (2009) Music in the seventeenth and eighteenth centuries: the Oxford history of western music. Oxford University Press, New York
37. Grout DJ, Palisca CV, Burkholder JP (2006) A history of western music, 7th edn. Norton, New York
38. Taruskin R (2010) Music in the nineteenth century: the Oxford history of western music. Oxford University Press, New York
39. Sutskever I, Martens J, Hinton GE (2011) Generating text with recurrent neural networks. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 1017–1024
40. Graves A (2013) Generating sequences with recurrent neural networks. arXiv preprint. arXiv:1308.0850
41. Kim Y, Jernite Y, Sontag D, Rush AM (2016) Character-aware neural language models. In: AAAI, pp 2741–2749
42. Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
43. Gers FA, Schmidhuber J, Cummins F (1999). Learning to forget: continual prediction with lstm
44. Mikolov T, Karafiát M, Burget L, Cernockỳ J, Khudanpur S (2010) Recurrent neural network based language model. In: Interspeech, vol 2, p 3
45. Cuddy LL, Lunney CA (1995) Expectancies generated by melodic intervals: perceptual judgments of melodic continuity. Atten Percept Psychophys 57(4):451–462
46. Temperley D (2007) Music and probability. MIT Press, Cambridge
47. Narmour E (1991) The top-down and bottom-up systems of musical implication: building on Meyer's theory of emotional syntax. Music Percept 9(1):1–26
48. Schellenberg EG (1996) Expectancy in melody: tests of the implication-realization model. Cognition 58(1):75–125
49. Zivic PHR, Shifres F, Cecchi GA (2013) Perceptual basis of evolving western musical styles. Proc Natl Acad Sci 110(24):10034–10038
50. Loui P, Wessel D (2007) Harmonic expectation and affect in western music: effects of attention and training. Atten Percept Psychophys 69(7):1084–1092
51. Pachet F (2017) A joyful ode to automatic orchestration. ACM Trans Intell Syst Technol 8(2):18
52. Dannenberg RB, Goto M (2008) Music structure analysis from acoustic signals. Springer, New York
53. Huang C-ZA, Vaswani A, Uszkoreit J, Shazeer N, Simon I, Hawthorne C, Dai AM, Hoffman MD, Dinculescu M, Eck D (2016) Music transformer. arXiv preprint. arXiv:1809.0428
54. Webster J, Feder G (2003) The new grove haydn. Oxford University Press, New York
55. Berlyne D (1969) Arousal, reward and learning. Ann NY Acad Sci 149(3):1059–1070
56. Berlyne D (1970) Novelty, complexity, and hedonic value. Percept Psychophys 8(5A):279–286
57. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562
58. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022
59. Min S, Park J (2016) Mapping out narrative structures and dynamics using networks and textual information. arXiv preprint. arXiv:1604.03029
60. Lee B, Kim D, Sun S, Jeong H, Park J (2018) Heterogeneity in chromatic distance in images and characterization of massive painting data set. PLoS ONE 13(9):0204430