ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

Lennon or McCartney? Can Statistical Analysis Solve an Authorship Puzzle?

Stylometry—the use of statistical techniques to determine authorship—is best known for identifying the Unabomber as Theodor Kaczynski and revealing that Shakespeare collaborated with Christopher Marlowe on the Henry IV play cycle. In textual analysis, it is not the unusual word choice that betrays the hidden voice, but the habitual—the recurring patterns of common words, such as prepositions, that mark the probable identity of one person alone.

It was a mutual Beatles passion—discovered at a conference on Prince Edward Island—that led Mark Glickman, senior lecturer in statistics at Harvard, and Jason Brown, professor of mathematics at Dalhousie University, to wonder whether a stylometric approach could answer the burning question: Lennon or McCartney?

As Glickman explains, for most Lennon-McCartney songs, it is well-known and well-documented which of the two wrote the song. However, a surprisingly large number of songs (or portions of songs) have disputed authorship. As an example, no one knows who wrote the music for "In My Life," a track from the 1965 album Rubber Soul, which is ranked 23 on *Rolling Stone's* The 500 Greatest Songs of All Time. Both Lennon and McCartney remembered differently. "So, we wondered whether you could use data analysis techniques to try to figure out what was going on in the song to distinguish whether it was by one or the other," says Glickman.

With help from former Harvard statistics student Ryan Song, Glickman and Brown "decomposed" each Beatles song from 1962 to 1966 into five representations. Each representation consisted of the frequency of occurrence of a set of musical features within each song. "The basic idea behind our approach," says Glickman, "is to convert a song, whose musical content is difficult to quantify in any direct way, into a set of different data structures that are amenable for establishing a signature of a song using a quantitative approach." Glickman continues, "Think of decomposing a color into its constituent components of red, green and blue with different weights attached. We're doing the same thing with Beatles songs, though with more than three components. In total, our method divides songs into a total of 149 constituent components."

"The first representation simply consists of the frequencies of different commonly played chords, along with aggregations of uncommon chords," says Glickman. "We were able to form

11 chord categories." Then, they characterized melodic notes—notes sung by the lead singer. Third, they recorded the frequencies of occurrence of chord transitions, that is, one chord followed by another chord. Again, certain uncommon chord transitions were aggregated into single categories. Fourth, they recorded the frequencies of consecutive melodic note pairs.

And then, finally, they decomposed songs into four-melodic note "contours." A contour, says Glickman, is a four-note melodic sequence categorized into a series of "ups," "downs" and "stays the same." In other words, if a four-note melodic passage involves four notes increasing in pitch, then the contour would be ("up," "up" "up") because each consecutive pair of notes involves an increase in pitch. Examining four-note contours, says Glickman, adds extra detail that can help distinguish styles of melodic composition.

The reason these five representations can serve as signatures of different musical compositional styles is because, as Glickman points out, there is something well-known about the Beatles' songwriting styles: Lennon typically wrote melodic lines that didn't vary much.

"Consider the Lennon song, 'Help!'" says Glickman. "It basically goes, 'When I was younger, so much younger than today,' where the pitch doesn't change very much. It stays at the same note repeatedly, and only changes in short steps. Whereas with Paul McCartney, you take a song like 'Michelle,' and it goes, 'Michelle, ma belle. Sont les mots qui vont très bien ensemble.' In terms of pitch, it's all over the place."

Their approach to infer unknown or disputed authorship from musical features can be understood in three steps. First, their model posits that each of the frequencies of the 149 musical features within a song depends on the song's author. For example, the "tonic" (the root chord of a song) is assumed to occur with one frequency in Lennon songs, but a possibly different frequency in McCartney songs. Second, they use a common tool in probability called "Bayes rule" to reverse the probability. In other words, starting with the frequency of the 149 musical features knowing a song's author, they determine a model for the probability Lennon or McCartney wrote a song given the frequency of the 149 musical features. This model was then trained using 70 Lennon-McCartney songs or song portions in which the authorship was truly known. Finally, as a third step, the results of this model were applied to Lennon-McCartney songs and song portions in which the authorship was disputed, which resulted in probability predictions for the songs of unknown authorship.

"So, the probability that 'In My Life' was written by McCartney is .018," says Glickman, "which basically means it's pretty convincingly a Lennon song." McCartney remembers differently. But "The Word," which Glickman thought was certain to be a Lennon song turned out, according to their model, to be almost certainly by McCartney.

Is there more to this exercise than a fun musical whodunnit? "Yes," says Glickman. "This technology can be extended. We can look at pop history and chart the flow of stylistic influence."

JSM Talk: Assessing Authorship of Beatles Songs from Musical Content: Bayesian Classification Modeling from Bags-of-Words Representations
http://ww2.amstat.org/meetings/jsm/2018/onlineprogram/AbstractDetails.cfm?abstractid=329336

For details, contact: Mark Glickman

Email: glickman@fas.harvard.edu

Tel: (617) 496-1505

Webpage: www.glicko.net

About JSM 2018

JSM 2018 is the largest gathering of statisticians and data scientists in the world, taking place July 28–August 2, 2018, in Vancouver. Occurring annually since 1974, JSM is a joint effort of the American Statistical Association, International Biometric Society (ENAR and WNAR), Institute of Mathematical Statistics, Statistical Society of Canada, International Chinese Statistical Association, International Indian Statistical Association, Korean International Statistical Society, International Society for Bayesian Analysis, Royal Statistical Society and International Statistical Institute. JSM activities include oral presentations, panel sessions, poster presentations, professional development courses, an exhibit hall, a career service, society and section business meetings, committee meetings, social activities and networking opportunities.
http://ww2.amstat.org/meetings/jsm/2018/index.cfm

About the American Statistical Association

The ASA is the world's largest community of statisticians and the oldest continuously operating professional science society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare. For additional information, please visit the ASA website at www.amstat.org.