6th International Conference on Computational Social Science IC^2S^2
July 17-20, 2020, MIT Media Laboratory Boston, MA USA

Which is your favourite music genre? A validity comparison of Facebook data and survey data

keywords: Facebook, survey, validity, music interest, algorithmically inferred interests

Social sciences' primary aim is to construct valid measurements. The validity of surveys has been criticised for a long time as they create an artificial environment while collecting data with some pre-specified purpose. In contrast, Facebook (FB) yields 'organic data,' that is, observational data of users' behaviors. However, validity is a concern in the case of FB, too, see e.g. the biases introduced by shifts in algorithms or fake self-representation conformed to social expectations.

In our study, we investigate whether there are transition paths between the two data sources and whether we can overcome some validity issues and operationalization-related questions by using the two information sources in parallel to cross-validate them by each other.

Because of the Cambridge Analytica scandal, FB practically cut off data access even for researchers, only contents of the public sites are possible to load [1]. We solved the accessibility problem by collaborating with the users instead of the company. [2,3] After informed consent was obtained, respondents were asked to download their FB profile archive that was anonymized on the spot to ensure confidentiality. One hundred fifty respondents took part in our study, the sample is non-probability quota sample, with quotas for age category and gender. All the respondents were Hungarian.

The collection of personal FB data archive was combined with a face-to-face survey. The topic we chose to study in this paper was music interest, which is a key indicator in cultural sociology, and whose 'digital trace' has its own relevance since the Internet has been the primary locus of music consumption.

In our analysis, we measured music interest in three different ways. (1) *Self-reported interest*. In the survey, we measured nine music genres, using a 1-7 scale of liking the given genre. (2) *Digitally expressed interest*. FB's page like data contains the name of the Facebook page. We used those pages which were categorized as music pages by FB. Our coders categorized these pages manually into genre categories that were identical to the typology what we had in the survey. (3) *Algorithmically inferred interest*. Here we used ads interest data. Facebook categorizes every user for sales for advertising. This is an algorithmic classification of the users based on their own likes, activities, and used keywords and also based on their friends' preferences. The algorithm is a black-box; we can only observe the result of the categorization. We could extract 7 music genres. We didn't find any interest group for the genre 'mulatós' that is a specific Hungarian music.

To our knowledge, there are no previous studies on the rate at which any population is categorized by Facebook to advertisers as interested in music genres, or on the relationship between self-reported interest, digitally expressed interest and ad-interest categories.

Specific genres were detected which show remarkable different pictures when measured these different ways. We found that people without music page like, have overall strong music preference, and they use Facebook actively, but they don't use page-like functionality.

Personal FB data archives also make it possible to analyse the dynamical patterns of people's behaviour. Our results showed that music page like increased linearly in the first five years of users' FB-usage, but the growth rate started to decrease afterward (see Figure 1). Using only the first three years of page-like data, we were able to estimate quite well the whole range of full-period music preference of the user.

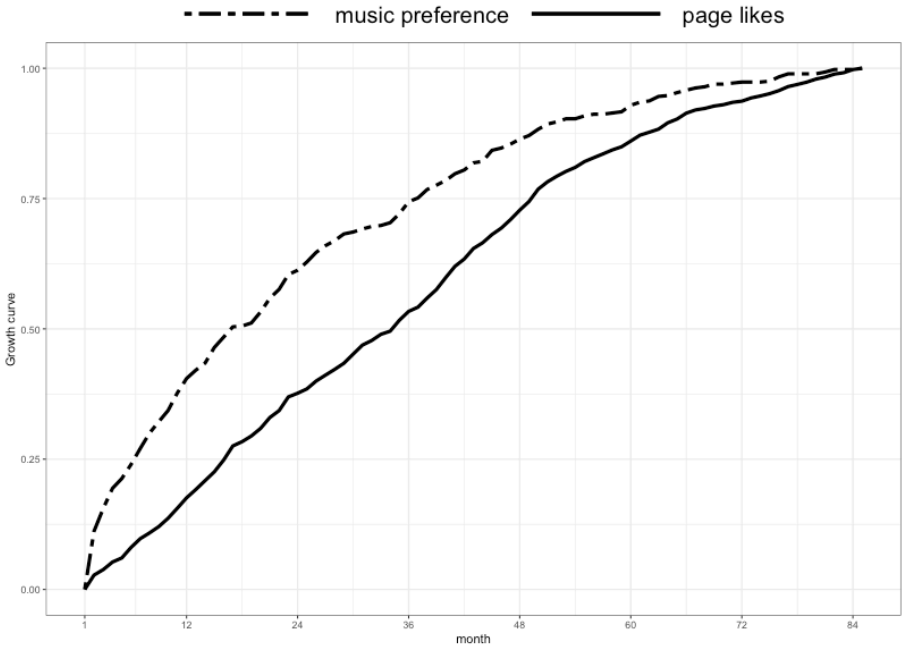

Figure 1. The Cumulative growth of Facebook page likes and ratio of preferred music genres

In this paper, we primaryly focused on methodological problems: validity issues and operationalization questions. However, our combined data source could be used also to better understand some important sociological phenomenon. Social desirability and self-representation in social media are one of these areas. Digital data - like social media data or search engine data - opens the possibility to examine topics we could not examine earlier or re-examine topics with new approaches. However, all the data sources have their own validity problems. Our paper adds a new contribution to this topic. We show the research potential of using alternative data sources together. Our methodological experiences add to the technical feasibility of such studies, while our substantive results provide important results about validity issues of different measurement methods. We show that we can overcome some of these issues by cross-validating the two data sources by each other. We also show that if the researcher gets access to a part of the users' social media data archive, the problem of the "leanness" of social media data is eliminated. We hope our data collection method and the presented validity approaches will initiate a future dialogue in digital data research in social sciences.

References

[1] D Freelon,. Political Communication, 35, no. 4: 665-668. (2018)

[2] C Marino et al., Computers in Human Behavior, 73, 541–546 (2017)

[3] K Thorson, K Cotter, M Medeiros, C Pak, Information, Communication & Society, 1-18 (2019)