Media and Communication
Open Access Journal | ISSN: 2183-2439
Volume 8, Issue 3 (2020)

Computational Approaches to Media Entertainment Research

Editors

Johannes Breuer, Tim Wulf and M. Rohangis Mohseni

COGITATIO

Media and Communication, 2020, Volume 8, Issue 3
Computational Approaches to Media Entertainment Research

Published by Cogitatio Press
Rua Fialho de Almeida 14, 2º Esq.,
1070-129 Lisbon
Portugal

Available online at: www.cogitatiopress.com/mediaandcommunication

Table of Contents

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 147–152
DOI: 10.17645/mac.v8i3.3530

Editorial

New Formats, New Methods: Computational Approaches as a Way Forward for Media Entertainment Research

Johannes Breuer [1,*], Tim Wulf [2] and M. Rohangis Mohseni [3]

[1] Data Archive for the Social Sciences, GESIS—Leibniz Institute for the Social Sciences, 50667 Cologne, Germany; E-Mail: johannes.breuer@gesis.org
[2] Department of Media and Communication, LMU Munich, 80538 Munich, Germany; E-Mail: tim.wulf@ifkw.lmu.de
[3] Institute of Media and Communication Science, Ilmenau University of Technology, 98693 Ilmenau, Germany; E-Mail: rohangis.mohseni@tu-ilmenau.de

* Corresponding author

Submitted: 31 July 2020 | Published: 13 August 2020

Abstract
The rise of new technologies and platforms, such as mobile devices and streaming services, has substantially changed the media entertainment landscape and continues to do so. Since its subject of study is changing constantly and rapidly, research on media entertainment has to be quick to adapt. This need to quickly react and adapt not only relates to the questions researchers need to ask but also to the methods they need to employ to answer those questions. Over the last few years, the field of computational social science has been developing and using methods for the collection and analysis of data that can be used to study the use, content, and effects of entertainment media. These methods provide ample opportunities for this area of research and can help in overcoming some of the limitations of self-report data and manual content analyses that most of the research on media entertainment is based on. However, they also have their own set of challenges that researchers need to be aware of and address to make (full) use of them. This thematic issue brings together studies employing computational methods to investigate different types and facets of media entertainment. These studies cover a wide range of entertainment media, data types, and analysis methods, and clearly highlight the potential of computational approaches to media entertainment research. At the same time, the articles also include a critical perspective, openly discuss the challenges and limitations of computational methods, and provide useful suggestions for moving this nascent field forward.

Keywords
communication research; computational methods; computational social science; media entertainment

Issue
This editorial is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

With the rapid development of technology and the growing competition for the attention (and money) of the audience, the entertainment media landscape is constantly changing. The global spread of broadband internet, mobile devices, streaming platforms, and communication tools with which people can, for example, discuss entertainment content, have had an immense impact on the structure and use of entertainment media. These fundamental changes in the media entertainment landscape not only affect the everyday lives of people worldwide, but also create opportunities and challenges for research that looks at the use and effects of these media.

New media entertainment formats entail new research questions about, for example, the motivations and experiences of their users, and the effects that the use of these new formats can have on them. Answering

these questions may require revisions to existing theories or even completely novel theoretical approaches. In addition to that, studying new media entertainment formats may also necessitate the development of new methods for collecting and analyzing data. Besides the development or refinement of theories and methods, another important aspect associated with the emergence of new (digital) media formats and platforms is the huge amount of data that their usage generates, which is both a challenge and an opportunity for entertainment research.

For several decades now, most quantitative research on the content, use, and effects of media entertainment has been based on data from surveys, manual content analyses, or lab experiments. While there is no doubt that these studies have produced many important insights into media entertainment, the data they are based on have certain limitations. For example, several recent studies have shown that self-reports of media use tend to be unreliable (e.g., Araujo, Wonneberger, Neijens, & de Vreese, 2017; Scharkow, 2016). This is especially problematic if researchers are interested in very specific, rare, or socially undesirable forms of media entertainment. Experimental lab studies, on the other hand, tend to have relatively small samples and often occur in somewhat unnatural settings. Moreover, manual content analyses are not suitable for the large amounts of data that users of media entertainment generate (e.g., discussion threads on Reddit or tweets about a show, movie, or video game).

Parallel to the largely technology-driven developments in the entertainment landscape, the methodological portfolio of social-scientific research has also been substantially extended by the rise of computational social science which "leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale" (Lazer et al., 2009, p. 722). According to Hox (2017), two key identifying features of computational methodology are the use of "big data" (although the term is often defined differently and tends to be underdefined) and the use of analysis techniques that are suited for these kinds of data. These analysis methods typically belong to the areas of text mining and natural language processing, machine learning, and network analysis. Regarding the type of data used, especially for computational communication research, it is typically more precise to speak of digital trace data which can be roughly defined as "records of activity (trace data) undertaken through an online information system" (Howison, Wiggins, & Crowston, 2011) and can originate from various sources, including social media platforms, websites, or smartphone apps. These traces can be intentional, such as tweets or Reddit comments, or unintentional, such as information about users or their activity collected by a streaming platform (Hox, 2017). Given their expertise in analyzing the use, content, and effects of digital media, "communication scholars are in a uniquely strategic position to lead the development of the computational approaches that promise to offer novel and exciting insights" (Hilbert et al., 2019, p. 3932). Indeed, computational communication science is a distinct subdiscipline "that investigates the use of computational algorithms to gather and analyze big and often semi- or unstructured data sets to develop and test communication science theories" (Van Atteveldt, Margolin, Shen, Trilling, & Weber, 2019, p. 1; also see Domahidi, Yang, Niemann-Lenz, & Reinecke, 2019; Van Atteveldt & Peng, 2018). Computational communication research has seen a rapid growth over the last few years. One clear indicator of this is that the former interest group Computational Methods has become a full division of the International Communication Association in 2020. While most studies in this area have looked at topics related to information seeking, news consumption, or political communication, there has been relatively little research on entertainment media. This thematic issue seeks to address this gap.

The characteristics identified by Hox (2017) also apply to the articles included in this thematic issue: They use (big) digital trace data and advanced analysis methods to study various phenomena related to the use of different kinds of entertainment media. In addition, they combine different analysis methods and types of data, which is also typical of computational communication research (and computational social science in general). To illustrate the diversity of topics and approaches, we provide an overview of the media and data types as well as the analysis methods in Table 1. Interestingly, there is a striking overlap in the types of data and analysis methods that almost all large entertainment companies nowadays use to evaluate and improve their products (as well as to profile and better target users) and the new computational methods that researchers have started to use for entertainment research. This further highlights the practical relevance of computational approaches in entertainment research.

A key challenge for computational entertainment and communication research, and even computational social science in general, is the question of how to access digital trace data and what can be done with them. Researchers not only have to consider the (privacy) interests of the people whose data they collect and use, but also those of commercial companies as typically specified in their Terms of Service (Van Atteveldt, Strycharz, Trilling, & Welbers, 2019). Especially when it comes to the ideals of open science, the interests of the researchers who use the data and the commercial companies who control it can be conflicting (Breuer, Bishop, & Kinder-Kurlanda, in press). Against this background, we are especially excited that for several of the articles included in this thematic issue, the authors were able to make their analysis code and data available (see Table 1). Of course, using digital trace data also has other limitations and potential pitfalls. These include the common lack of individual level information about the users and relevant outcome variables (Stier, Breuer, Siegers, & Thorson, 2019) or potential biases (Sen, Flöck, Weller,

Table 1. Overview of the articles in this thematic issue, including focus, data types, and analysis methods.

Article title	Author(s)	Entertainment media focus	Data types	Analysis method(s)	Open materials
What Is Important When We Evaluate Movies? Insights from Computational Analysis of Online Reviews	Schneider, Domahidi, and Dietrich	Movies	Online movie reviews, self-reports from online surveys	Correlated topic models, qualitative content analysis	https://osf.io/pqnk6
A Graph-Learning Approach for Detecting Moral Conflict in Movie Scripts	Hopp, Fisher, and Weber	Movies	Movie scripts	Social network analysis, natural language processing	Script: https://github.com/EdinburghNLP/scriptbase Data: https://osf.io/rbdws
(A)synchronous Communication about TV Series on Social Media: A Multi-Method Investigation of Reddit Discussions	Unkel and Kümpel	Series	Reddit threads, self-reports from online surveys	Automated content analysis	https://osf.io/7v49t
Popular Music as Entertainment Communication: How Perceived Semantic Expression Explains Liking of Previously Unknown Music	Lepa, Steffens, Herzog, and Egermann	Music	Self-reports from surveys/experiments, audio data (music)/music information retrieval data	Regression, factor analysis, machine learning, algorithmic audio signal analysis	
A Computational Approach to Analyzing the Twitter Debate on Gaming Disorder	Schatto-Eckrodt, Janzik, Reer, Boberg, and Quandt	Video games	Tweets	Sentiment analysis, network analysis, topic models	https://osf.io/vzymj
Exploring the Effect of In-Game Purchases on Mobile Game Use with Smartphone Trace Data	Boghe, Herrewijn, De Grove, Van Gaeveren, and De Marez	Video games	Smartphone log data	Survival analysis	
Open-Source's Inspirations for Computational Social Science: Lessons from a Failed Analysis	Poor	Not applicable	Not applicable	Not applicable	

Weiss, & Wagner, 2019). It is reassuring and promising for the future of this young field to see that all contributions in this thematic issue are aware of these issues and explicitly address them.

The study by Schneider, Domahidi, and Dietrich (2020) compares insights from self-report measures with online movie reviews to capture how viewers evaluate movies. They used subjective movie evaluation criteria (SMEC), identified based on self-report data from online surveys, and related those to a correlated topic model that explores the underlying topics of openly available user reviews. The study found correspondences for three major SMEC categories (hedonism, narrative, and actors' performance) in the online reviews, with additional qualitative analyses revealing further occurrence of SMEC categories in the review texts.

The study by Hopp, Fisher, and Weber (2020) also looks at movies. Using a combination of social network analysis and natural language processing techniques, they were able to develop a method for detecting moral conflict in scripts of more than 80,000 movie scenes. Among other things, they found that moral conflict can be identified by changes in the structures of social networks of movie characters.

Unkel and Kümpel (2020) also used a combination of computational and traditional methodological approaches to study synchronous and asynchronous communication about a TV series on Reddit. Specifically, they examined the motives of using Reddit forums for communication before, while, and after watching new episodes of the final season of *Game of Thrones*. Combining automated content analyses of these threads with a survey among thread users, they found that different motives lead to using these thread types, and different thread types are associated with different forms of interactions.

The contribution by Lepa, Steffens, Herzog, and Egermann (2020) employed a set of computational and other methods to study popular music as entertainment communication. Using an existing dataset, they developed a model for predicting listener liking ratings for previously unknown songs and found that unknown music is liked more, the more it is perceived as emotionally and semantically expressive. In a second study, the authors developed and tested a machine learning model drawing on automatic audio signal analysis and found that it can predict significant proportions of variance in musical meaning decoding.

Schatto-Eckrodt, Janzik, Reer, Boberg, and Quandt (2020) made use of computational approaches for analyzing the debate about gaming disorder on Twitter around the time in 2018 when the World Health Organization (WHO) decided to include the addictive use of digital games (gaming disorder) as a diagnosis in the International Classification of Diseases. The authors used a combination of sentiment, network, and automated content analysis (topic models), and found that the debate was largely organic (i.e., not driven by spam accounts) and heavily impacted by the WHO decision.

The article by Boghe, Herrewijn, De Grove, Van Gaeveren, and De Marez (2020) also looks at digital games, although with a very different research question and methodological approach. They used smartphone data to explore the effect of in-game purchases on continual mobile game use. In a survival analysis with the log data, they found that, while making an in-game purchase initially decreases the risk of stopping to play a game, there is a reversal effect in the sense that previous in-game purchases negatively affect the chance of continued play at a later point in time.

Unlike the other articles, the final contribution to this thematic issue by Poor (2020) does not present empirical results but offers a critical meta-perspective on computational approaches to media entertainment research. Building on his own experiences, the author discusses how and why computational research can fail and what the young field of computational social science can learn from the long history of the open source (software) movement.

Overall, the articles in this thematic issue cover different topics and employ different (methodological) approaches to study media entertainment. Despite their differences, they all show the potential of computational approaches for media entertainment research, while at the same time also highlighting some of the challenges and potential limitations. What all of the articles clearly illustrate is that combinations of different methods (including computational as well as more traditional approaches) and data types (including digital trace data as well as other types, such as self-reports) represent a promising way of moving entertainment research forward. Hence, we believe that with this thematic issue we offer researchers in the field of (entertainment) communication a diverse portfolio of applications of computational methods for various research questions. We hope that this work will inspire entertainment research and guide the way to a more nuanced triangulation and diversity of methods used in this research area.

Acknowledgments

First of all, we would like to thank the authors for their wonderful contributions. We also want to thank the reviewers of this thematic issue. Their knowledge of a variety of computational methods and the subject matter of the articles was invaluable and greatly helped in improving the overall quality of the thematic issue. Finally, we would like to thank the editorial team at *Media and Communication* for their great organization of the whole process from the initial planning to the publication of this issue.

Conflict of Interests

The authors declare no conflict of interests.

References

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Boghe, K., Herrewijn, L., De Grove, F., Van Gaeveren, K., & De Marez, L. (2020). Exploring the effect of in-game purchases on mobile game use with smartphone trace data. *Media and Communication*, *8*(3), 219–230.

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (in press). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*.

Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). Outlining the way ahead in computational communication science: An introduction to the IJoC Special Section on "computational methods for communication science: Toward a strategic roadmap." *International Journal of Communication*, *13*, 3876–3884.

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., . . . Zhu, J. J. H. (2019). Computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication*, *13*, 3912–3934.

Hopp, F. R., Fisher, J. T., & Weber, R. (2020). A graph-learning approach for detecting moral conflict in movie scripts. *Media and Communication*, *8*(3), 164–179.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767–797. https://doi.org/10.17705/1jais.00282

Hox, J. J. (2017). Computational social science methodology, anyone? *Methodology*, *13*(Supp. 1), 3–12. https://doi.org/10.1027/1614-2241/a000127

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lepa, S., Steffens, J., Herzog, M., & Egermann, H. (2020). Popular music as entertainment communication: How perceived semantic expression explains liking of previously unknown music. *Media and Communication*, *8*(3), 191–204.

Poor, N. (2020). Open-source's inspirations for computational social science: Lessons from a failed analysis. *Media and Communication*, *8*(3), 231–238.

Scharkow, M. (2016). The accuracy of self-reported internet use: A validation study using client log data. *Communication Methods and Measures*, *10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Schatto-Eckrodt, T., Janzik, R., Reer, F., Boberg, S., & Quandt, T. (2020). A computational approach to analyzing the Twitter debate on gaming disorder. *Media and Communication*, *8*(3), 205–218.

Schneider, F. M., Domahidi, E., & Dietrich, F. (2020). What is important when we evaluate movies? Insights from computational analysis of online reviews. *Media and Communication*, *8*(3), 153–163.

Sen, I., Flöck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A total error framework for digital traces of humans. *arXiv.org*. Retrieved from https://arxiv.org/abs/1907.08228

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review.* Advance online publication. https://doi.org/10.1177/0894439319843669

Unkel, J., & Kümpel, A. S. (2020). (A)synchronous communication about tv series on social media: A multimethod investigation of Reddit discussions. *Media and Communication*, *8*(3), 180–190.

Van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019). A roadmap for computational communication research. *Computational Communication Research*, *1*(1), 1–11. https://doi.org/10.5117/CCR2019.1.001.VANA

Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, 12(2/3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Towards open computational communication science: A practical roadmap for reusable data and code. *International Journal of Communication*, *13*, 3935–3954.

About the Authors

Johannes Breuer (PhD) is a Senior Researcher at the Data Archive for the Social Sciences at GESIS—Leibniz Institute for the Social Sciences in Cologne, Germany, where his work focuses on the topics of data linking and using digital trace data for social science research. His other research interests include the use and effects of digital media, computational methods, and open science. More information: https://www.johannesbreuer.com

Tim Wulf (PhD) is a Post-Doctoral Researcher at the Department of Media and Communication at LMU Munich in Germany. His research interests include the effects of media-induced nostalgia, media psychological perspectives on video games and video game streaming, and persuasion through narrative media. More information: https://www.tim-wulf.de

M. Rohangis Mohseni (Dr.rer.nat) is a Post-Doctoral Researcher of the Media Psychology and Media Design research group at Ilmenau University of Technology in Germany. His research interests include electronic media effects and moral behavior. His latest publications address gendered hate speech on YouTube (*SCM*, 2020), digital aggression (*merz*, 2020), and mobile learning (*Handbuch Bildungstechnologie*, 2020). More information: http://www.rmohseni.de

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 153–163
DOI: 10.17645/mac.v8i3.3134

Article

What Is Important When We Evaluate Movies? Insights from Computational Analysis of Online Reviews

Frank M. Schneider [1,*], Emese Domahidi [2] and Felix Dietrich [1]

[1] Institute for Media and Communication Studies, University of Mannheim, 68159 Mannheim, Germany;
E-Mails: frank.schneider@uni-mannheim.de (F.M.S.), fedietri@mail.uni-mannheim.de (F.D.)
[2] Institute for Media and Communication Science, TU Ilmenau, 98693 Ilmenau, Germany;
E-Mail: emese.domahidi@tu-ilmenau.de

* Corresponding author

Submitted: 14 April 2020 | Accepted: 15 July 2020 | Published: 13 August 2020

Abstract
The question of what is important when we evaluate movies is crucial for understanding how lay audiences experience and evaluate entertainment products such as films. In line with this, subjective movie evaluation criteria (SMEC) have been conceptualized as mental representations of important attitudes toward specific film features. Based on exploratory and confirmatory factor analyses of self-report data from online surveys, previous research has found and validated eight dimensions. Given the large-scale evaluative information that is available in online users' comments in movie databases, it seems likely that what online users write about movies may enrich our knowledge about SMEC. As a first fully exploratory attempt, drawing on an open-source dataset including movie reviews from IMDb, we estimated a correlated topic model to explore the underlying topics of those reviews. In 35,136 online movie reviews, the most prevalent topics tapped into three major categories—Hedonism, Actors' Performance, and Narrative—and indicated what reviewers mostly wrote about. Although a qualitative analysis of the reviews revealed that users mention certain SMEC, results of the topic model covered only two SMEC: Story Innovation and Light-heartedness. Implications for SMEC and entertainment research are discussed.

Keywords
entertainment media; IMDb; movie evaluation; movie reviews; topic modeling; self-reports

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany), and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

When Louis Leon Thurstone (1930) developed *A Scale for Measuring Attitude Toward the Movies* in the context of the Payne Fund Studies (1929–1932), it was one of the first attempts to measure interindividually different attitudes in movie effects research. Back then, social scientists saw movies as a social problem, in particular, for child and youth development (cf. Wartella & Reeves, 1985). Nowadays, the entertainment and film industry are booming (Hennig-Thurau & Houston, 2019). More than ever before, communication scholars dedicate themselves to learning about how watching movies influences entertainment experiences, what (positive) consequences follow, or how predispositions towards entertainment media shape movie selection and use (Raney & Bryant, 2020). Here, subjective movie evaluation criteria (SMEC) play a crucial role in evaluating movies before, during, and after exposure (Schneider, Welzenbach-Vogel, Gleich, & Bartsch, in press) and can help to predict specific individual movie evaluations (Schneider, 2012a). SMEC are conceptualized as "mental representations of important attitudes towards specific film features" (Schneider, 2017, p. 71). To mea-

sure SMEC and address the question of what is important when viewers evaluate a movie, the SMEC scales have been developed and validated (Schneider, 2012a, 2017). This, however, has been largely based on factor-analytical examination of self-report data. As subjective criteria may best predict subjective choices, processing, and effects, such a methodological approach makes good sense. Nevertheless, support for the construct validity of SMEC could be strengthened if distinct methodological approaches arrive at similar conclusions from different angles. Moreover, it might be interesting to learn from viewers' written evaluative responses to movies in a more natural environment, aside from a scientific setting (e.g., an online survey). Such an approach may be more unobtrusive and less prone to issues with mental accessibility or social desirability. The tremendous opportunities that movie users have to express themselves online coupled with today's computing power provide opportunities for the computational analyses of online users' movie reviews which could help further examine the construct validity of SMEC and explore the movie write-ups of lay audiences. Thus, in the present article, we follow an exploratory approach as we are interested in what online movie reviewers write about, and if such online reviews provide insight into underlying SMEC, which they might have applied to evaluate a movie.

2. Theoretical Background and State of Research

2.1. A Brief Overview of Subjective Movie Evaluation Criteria

Whereas Thurstone (1930, see beginning of Section 1) was interested in attitudes towards movies in general, SMEC aim at examining the standards lay audiences use to assess movie features. Although several criteria had been suggested, most of them were not validated or applied to TV shows or specific target groups only (Schneider, 2017). To address these shortcomings, previous research developed and validated scales for measuring and examining the structure of SMEC (for details see Schneider, 2012a, 2017). The procedure comprised open-ended data from an online-survey, studies including a modified structure formation technique, focus groups, and quantitative content analysis of criteria categorization, pretesting and revising of the item pool, exploratory and confirmatory factor analyses, and latent state-trait analyses. Results provided evidence for the content, structural, and substantive validity as well as for the reliability of the SMEC scales. Moreover, the nomological network of SMEC was explored (external validity by examining correlations with related constructs like film genre preferences and personality traits). The eight dimensions that emerged during this process and have been validated are as follows: Story Verisimilitude (SV), which reflects correspondence to (contemporary) reality (e.g., Gunter, 1997; Valkenburg & Janssen, 1999); Story Innovation (SI), which reflects the originality of the story (e.g., Greenberg & Busselle, 1996); Cinematography (CI), which reflects cinematic techniques (e.g., Gunter, 1997); Special Effects (FX), which also reflects cinematic techniques but focuses more on the technical aspects (e.g., Neelamegham & Jain, 1999; Rössler, 1997); Recommendation (RE), which reflects external resources for film evaluation (e.g., Neelamegham & Jain, 1999); Innocuousness (IN), which reflects a lack of potentially unpleasant characteristics (e.g., Nikken & van der Voort, 1997; Valkenburg & Janssen, 1999); Lightheartedness (LH), which reflects amusement and escapism (e.g., Greenberg & Busselle, 1996; Valkenburg & Janssen, 1999); Cognitive Stimulation (CS), which reflects the viewer's cognitive processes such as cogitation or learning (e.g., Himmelweit, Swift, & Jaeger, 1980; Nikken & van der Voort, 1997). Whereas the first four dimensions (SV, SI, CI, FX) summarize film-inherent elements, RE refers to film-external features, and the final three dimensions concern anticipated effects of use (IN, LH, CS).

In addition, some studies investigated the predictive power of LH for the evaluation of a comedy show (*My Name is Earl*; Burtzlaff, Schneider, & Bacherle, in press, Study 1), as well as the predictive power of IN, LH, and CS for the appreciation and enjoyment of movies in general as well as for specific genres (Schneider, 2012b). This makes the notion of SMEC particularly interesting for broader entertainment research. For instance, traditionally, film-specific evaluations have been mainly examined in light of entertainment experiences like enjoyment and pleasure (e.g., Vorderer, Klimmt, & Ritterfeld, 2004). They correspond to an evaluation criterion such as LH: Whereas when movie viewers enjoy a movie, their overall judgment about a movie can be more positive when they rate LH as highly important. However, recent advances go beyond hedonic consumption and advocate a more nuanced view of audience responses, reflecting a sense of meaning and growth, self-transcendence, or aesthetic and artistic quality (e.g., Oliver & Bartsch, 2010; Oliver & Raney, 2011; Oliver et al., 2018; Vorderer & Reinecke, 2015; Wirth, Hofer, & Schramm, 2012). Similarly, this gain in complexity is also reflected in various evaluation criteria that supplement criteria that refer to effects of use (e.g., CS) with criteria that focus more on the features of the movie (e.g., CI). SMEC might be better and more fine-grained predictors of film-specific evaluations than genre preferences and also emphasize content-related aspects (compared, e.g., to a user-centred approach; e.g., Swanson, 1987; Wolling, 2009). Thus, in sum, preliminary findings underscore the usefulness of SMEC for current entertainment research on movies and may help to understand the role of stable criteria in explaining audience responses before, during, and after movie exposure.

2.2. Previous Research on Online Movie Reviews

Although online movie reviews have been extensively researched in the last decades, this has been done al-

most exclusively in the domain of marketing studies (e.g., when investigating effects of online word of mouth on box-office success; e.g., Chintagunta, Gopinath, & Venkataraman, 2010; Eliashberg, Jonker, Sawhney, & Wierenga, 2000). Besides, more recently, online movie databases such as the International Movie Database (IMDb), Yahoo! Movies, or Douban have been among platforms subjected to computational analyses (e.g., Bader, Mokryn, & Lanir, 2017; Simmons, Mukhopadhyay, Conlon, & Yang, 2011; Yang, Yecies, & Zhong, 2020; Zhuang, Jing, & Zhu, 2006). In the following paragraphs, we give a brief overview of large-scale studies dealing with online movie reviews. For a more comprehensive summary, see Table S1 in the Supplementary Material. Please note that Table S1 and some parts of the remainder of this article include technical language of computational science. We refer the interested reader to comprehensible and introductory texts for communication scientists such as Günther and Quandt (2016) or Maier et al. (2018).

2.2.1. Predicting Box-Office Success

A wide range of management/economics studies have attempted to predict a movie's box-office success through statistical models, often including samples of online movie reviews (e.g., Hu, Shiau, Shih, & Chen, 2018; Hur, Kang, & Cho, 2016; Lee, Jung, & Park, 2017). Most of these models incorporated the online movie review's sentiment as well as other factors, some of which are also specific to reviews, such as the writing style, use of certain words (bag of words approach), or the length of review (Yu, Liu, Huang, & An, 2012). Other characteristics of online movie reviews such as its rated helpfulness (Lee & Choeh, 2018), the movie's numeric 'star'-rating (Hur et al., 2016), or the genre of movie (Lee & Choeh, 2018) were also often used but referred not directly to the online movie review's content.

2.2.2. Predicting Sentiments

The second line of research, focusing on methodological aspects of computational methods, attempted to establish, complement, or modify algorithms for the mining of online movie reviews, especially for sentiment analysis (e.g., Liu, Yu, An, & Huang, 2013; Parkhe & Biswas, 2016; Yang et al., 2020). As online movie reviews are relatively easy to scrape (e.g., from IMDb) and by nature mostly positive or negative (and rarely neutral), they provide good examples to develop and test classification algorithms. Most of these studies are situated within the fields of computational linguistics and computer science.

2.2.3. Other Computational Approaches

A few studies do not fit either of the previous categories (e.g., Bader et al., 2017; Otterbacher, 2013; Simmons et al., 2011; Yang et al., 2020). See Table S1 in the Supplementary Material for more details.

Three studies are particularly interesting concerning the aim of the present article. Drawing on emotion theory, Bader et al. (2017) created emotional signatures of movies and their genres based on emotions toward or elicited by a film that were extracted from its online reviews on IMDb. Their results imply that emotional evaluations also manifested themselves in online reviews and can help to cluster entertainment-related concepts such as movie genres. Moreover, as emotional and affective states are also related to SMEC (e.g., LH and IN), the appearance of words representing emotions in online movie reviews may help detect those criteria. In other words, it seems likely that SMEC that rely on affective evaluations are reflected in movie reviews. An 'emotional' approach to movies is also common in entertainment research (e.g., Bartsch, 2012; Soto-Sanfiel & Vorderer, 2011). Whereas these findings concern criteria as anticipated effects of viewing, two other studies focused more on film-inherent features as criteria. For instance, in an often-cited article, Zhuang et al. (2006) mined feature–opinion pairs within online movie reviews, based on a movie feature–opinion list. The feature part of this list contained names of movie-related individuals such as directors or leading actors and feature words of six movie elements, which were not derived from theory but were somewhat related to evaluation criteria—for instance, visual effects (partially refers to CI), FX, and screenplay (refers to SI). Thus, these findings support the idea that it is not only users' personal experiences (e.g., emotions) which play a role in movie reviews, but also movie-related features that are deemed necessary to achieve artistic and aesthetic quality. Lastly, using computer-aided content analysis, Simmons et al. (2011) found that storyline—among four other movie elements—was strongly related to the overall film grade. However, a deeper analysis revealed that what they called 'storyline' included statements about CI, action, humour, and entertainment, and thus represented a rather fuzzy concept. Disentangling what lies behind the storyline may hint at effects of use and film-inherent features as they are reflected by SMEC.

In sum, regarding entertainment theory and SMEC, all previous perspectives on online movie reviews show several weaknesses: First, although some studies refer to features or criteria, those criteria are not based on theoretical assumptions. Moreover, research has so far heavily relied on lexical databases or dictionaries, which implies that criteria are directly observable within the online reviews. Given the theoretical assumption that SMEC are latent constructs (Schneider, 2012a, 2017), methodological approaches that take this assumption into account could be more appropriate (e.g., Amplayo & Song, 2017, combined a multi-level sentiment classification with bi-term topic modeling).

To our best knowledge, on the one hand, no studies that computationally analyzed online movie reviews

have yet done so against the background of concepts related to entertainment theory. On the other hand, particularly within the field of entertainment theory, communication researchers have rarely used online review platforms to address research problems, even though the opportunities to assess digital traces of audience reactions seem easily available on a large scale and allow conclusions to be drawn about personal characteristics from online behaviour such as liking (Kosinski, Stillwell, & Graepel, 2013). For instance, if entertainment experiences are conceptualized as media effects (for recent overviews, see Raney & Bryant, 2020; Raney, Oliver, & Bartsch, 2020), responses to movie exposure and evaluative judgments of movies—both can be expressed as written online reviews—may indicate underlying evaluative factors (Schneider et al., in press).

2.3. The Present Research

Unlike previous studies that focused on predicting box-office success or sentiments from online movie reviews, we are interested in what online movie reviewers write about and if those online reviews provide insight into underlying SMEC, which the writers might have applied to evaluate a movie. In doing so, we try to figure out whether text mining of online movie reviews' content can support findings from self-report data and how analyses from such methodologically different approaches can contribute to construct validity. Until now, we are not aware of any similar attempt.

As the SMEC development has been a data-driven, inductive process, we decided against a confirmatory approach in favour of an exploratory, inductive, and unsupervised approach (i.e., topic modeling).

3. Method

Our sample is based on an open-source dataset including movie reviews from IMDb and their positive or negative sentiment classification (Maas et al., 2011). The dataset consists of 25,000 positive and 25,000 negative movie reviews. Additionally, 50,000 unlabeled reviews are provided. Only up to 30 reviews per movie are included in order to avoid a high number of correlated reviews. After downloading the data, we decided to focus on the positive and negative reviews only because these sentiments might be a sign that users expressed their SMEC. All data management, cleaning, and analysis was performed using *R 3.5.3* (R Core Team, 2020) and *RStudio 1.2.5033* (RStudio Team, 2019).

Before analysis, we opted for an extensive data preprocessing as recommended in the literature (Maier et al., 2018; Manning, Raghavan, & Schütze, 2008). First, we excluded all duplicate reviews from the dataset. Afterwards, we implemented common data preprocessing steps to delete text that provided no relevant information for automatic text analysis, such as cleaning of HTML tags and links and deleting numbers and whitespace via the *textclean* package in *R* (Rinker, 2018). To improve the quality of our dataset and to reduce the number of possible features, we deleted common stopwords via a combination of different stopword-lists and implemented lemmatization (Manning et al., 2008) via the *spacyr* (*library ldaR* wrapper; Benoit & Matsuo, 2020). Online movie reviews are of varying quality as users employ, for instance, internet slang as opposed to formal writing. To enhance the quality of the data and reduce internet slang, we automatically removed internet slang via *textclean* package in *R* (Rinker, 2018) and, based on part-of-speech tagging via *spacyr*, we selected only verbs, nouns, adjectives, and adverbs for further analysis. We deleted the most common words—'movie' and 'film'—because they are very general in our context but occurred more than 60,000 times in the corpus and thus three times more often than any other word. We implemented term frequency–inverse document frequency (tf–idf) weighting in order to determine how relevant a word in a given document is—that is, how often a word occurs in a document in relation to how often the word occurs in other documents of the corpus (Manning et al., 2008). In the following, we removed words that had a low tf–idf score (tf–idf < 0.050) and, thus, were not important for our analysis.

For data analysis, we employed topic modeling, an unsupervised machine learning approach to infer latent topics from a large sample size (Maier et al., 2018). Given the characteristics of our sample and theoretical assumptions (i.e., topics are likely to be correlated), we estimated a Correlated Topic Model (CTM) based on the movie reviews (Blei & Lafferty, 2009) using the *topicmodels* package in *R* (Grün & Hornik, 2011). To select the number of topics, we estimated 21 topic models from $k = 10$ to $k = 70$ via *ldatuning 1.0.0* package (Murzintcev & Chaney, 2020) and selected the 38-topic model as the best fitting model to our data. Then, we estimated a set of ten separate 38-topic CTMs with different initial parameters and selected from this set the best model regarding log-likelihood (Grün & Hornik, 2011) as our final model. We selected the topic with the highest probability per online movie review and with a minimum probability (gamma) of 0.02. The best fitting CTM included 38 topics for 41,434 online movie reviews. All scripts for data cleaning and analysis can be accessed via OSF (https://osf.io/pqnk6).

To allow for succinct presentation whilst ensuring coverage of the most important topics in the dataset, we focus on the most frequent topics in our sample with at least 600 reviews per topic. For all topics discovered in the dataset, please see the topic distribution and the top words for all topics in OSF. Based on a qualitative assessment of the top words of each topic, we organized the remaining 14 topics ($N = 35{,}136$) in three broad categories (see Table 1). Furthermore, drawing on the material (i.e., evaluation terms and criteria) used during the development of the SMEC scales (Schneider, 2012a, see Appendices A and B; Schneider, 2017, see item wording),

Table 1. CTM (k = 38, max. 1 topic/review, probability ≥ 0.02) with 14 manually selected topics with at least 600 reviews per review merged into three thematically overlapping topic categories, sorted alphabetically and by aggregated frequencies (*N* = 35,136).

k	Label	*n*	Top-10 words
9	AP: Acting 1	7,320	role, act, guy, character, script, Hollywood, excellent, kill, family, write
1	AP: Acting 2	1,847	actor, episode, watch, woman, lot, script, star, hard, character, funny
15	AP: Acting 3	776	cast, kid, act, guy, stupid, comedy, life, actor, character, effect
13	HE: Comedy 1	3,054	life, comedy, love, kid, excellent, waste, dvd, series, plot, recommend
2	HE: Comedy 2	1,213	comedy, performance, write, script, bad, family, love, pretty, story, original
14	HE: Fun 1	7,166	laugh, performance, fun, book, bad, family, funny, kid, dvd, bit
35	HE: Fun 2	6,838	funny, family, episode, script, series, lot, character, laugh, story, people
25	HE: Fun 3	1,724	funny, hour, rent, horror, laugh, series, story, write, money, shoot
6	HE: Fun 4	1,225	laugh, funny, people, story, love, effect, lot, write, pretty, minute
12	HE: Fun 5	998	funny, watch, bad, people, scene, pretty, plot, friend, enjoy, hard
23	HE: Fun 6	739	bad, funny, awful, story, family, recommend, rent, original, watch, Hollywood
36	HE: Fun 7	729	laugh, bad, recommend, book, character, watch, song, scene, comment, time
11	NA: Story & Plot 1	824	story, performance, bad, people, plot, funny, dvd, kid, recommend, music
10	NA: Story & Plot 2	683	story, actor, enjoy, bad, plot, role, horror, play, happen, waste

Notes: k = index of topic in initial solution; topic numbers reflect the original topic numbers as assigned by the model. To ensure the reproducibility of our results we report these numbers here. The 15 top terms per topic are available on OSF (https://osf.io/pqnk6).

we closely inspected those randomly selected reviews that had the highest gammas ($\gamma_{\text{min}} = 0.02$) and marked to which SMEC they referred (see Table 2).

4. Results

4.1. Correlated Topic Model

To answer the question of what online movie reviewers write about, we grouped the 14 topics into three categories for better interpretation (see Table 1). First, it is striking that most of the discovered topics concern funniness and comedy (labeled as 'Hedonism' [HE] category). Although the topics in these categories have nuanced meanings, on a general level, all of them relate to the presence or absence of hedonic and pleasurable kinds of media consumption. This fits into traditional lines of research that assumed enjoyment to be at the heart of entertainment (for a recent overview, see Raney & Bryant, 2020). Moreover, the HE category also reflects audience reactions. Broadly speaking, this fits the subjective movie evaluative criterion LH well. A second set of topics is broadly related to the acting of the cast and summarized in the category 'Actors' Performance' (AP). Although aspects of how well actors play their characters is not included in the final version of the SMEC scales, items that tapped into this category were part of the construction process (see Table B1, Items 47–51, in Schneider, 2012a, e.g., Item 47 reflects the general performance of actors). The third category, 'Narrative' (NA), comprises topics concerning story and plot. It relates to the subjective movie evaluation criterion SI. Both AP and NA refer to what has often been argued to be the most important elements for movie choice or evaluation (e.g., Linton & Petrovich, 1988; Neelamegham & Jain, 1999).

Taken together, online movie reviewers mostly write about whether or not they enjoyed a movie, about the APs, and about the quality of the movie's NA.

4.2. Additional Qualitative Exploratory Results

Our initial focus lay on the topic model. During interpreting, labeling, and summarizing, it became clear that some SMEC may not have emerged as topics because they were not prevalent. Nonetheless, descriptions related to these SMEC were not totally absent from the data. Based on material from previous research (e.g., criteria that participants named in open-question tasks, content of items in the initial item pool and in the final SMEC scales, and content of cards during modified structure formation technique; Schneider, 2012a, 2017), meaningful words and phrases were qualitatively checked, interpreted, and marked using superscripts. To illustrate this, we describe two examples in Table 2. They provide deeper insight into how SMEC are applied when writing online movie reviews. For instance, the second example refers to SMEC such as SV, RE, or CI as well. These examples are particularly interesting with regard to SMEC because many of the criteria that have been previously described by Schneider (2012a, 2017) can be discovered in these reviews.

Table 2. Two examples of randomly selected reviews with $\gamma \geq 0.02$ for each topic (k).

Review
Yesterday I finally satisfied my curiosity and saw this movie. My knowledge of the plot was limited to about 60 seconds of the trailer, but **I had heard some good critics [5]** which caused my expectations to increase. As I saw the movie, those untied pieces had been combined in **a story that was becoming quite intriguing, with some apparently inexplicable details [2]**. But in the end, everything is disclosed as a simple succession of events of bad luck, "sorte nula" in Portuguese. Above everything, I felt that the **story made sense, and everything fits in its place, properties of a good script [2]**. I must also mention the **soundtrack, which helps the creation of an amazing environment [9]**. And if you think of the **resources Fernando Fragata used to make this film, I believe it will make many Hollywood producers envious… [10]** Movie Title: *Sorte Nula* (2004) Path in IMDb dataset: aclImdb/test/pos/11479_8.txt Topic k = 1; $\gamma = 0.028$
On October of 1945, the American German descendant Leopold Kessler (Jean-Marc Barr) arrives in a post-war Frankfurt and his bitter Uncle Kessler (Ernst-Hugo Järegård) gets a job for him in the Zentropa train line as a sleeping car conductor. While travelling in the train learning his profession, he sees the destructed occupied Germany and meets Katharina Hartmann (Barbara Sukowa), the daughter of the former powerful entrepreneur of transport business and owner of Zentropa, Max Hartmann (Jørgen Reenberg). Leopold stays neutral between the allied forces and the Germans and becomes aware that there is a terrorist group called "Werewolves" killing the sympathizers of the allied and conducting subversive actions against the allied forces. He falls in love for Katharina, and sooner she discloses that she was a "Werewolf." When Max commits suicide, Leopold is also pressed by the "Werewolves" and need to take a position and a decision. "Europa" is an **impressive and anguishing Kafkanian story [2]** of the **great Danish director Lars von Trier. Using an expressionist style that recalls Fritz Lang and alternating a magnificent black & white cinematography with some coloured details [3]**, this movie **discloses a difficult period of Germany and some of the problems this great nation had to face after being defeated in the war. Very impressive the action of the occupation forces destroying resources that could permit a faster reconstruction of a destroyed country [1]**, and the corruption with the Jew that should identify Max. Jean-Marc Barr has a **stunning performance [11]** in the role of a man that wants to stay neutral but is manipulated everywhere by everybody. The **hypnotic narration of Max Von Sydow is another touch of class [11]** in this **awarded film[5]**. My vote is nine. Movie Title: *Europa* (1991) File path: aclImdb/train/pos/130_9.txt Topic k = 1; $\gamma = 0.028$

Notes: k = index of topic; γ = the probability of a given review to be associated with the topic k (please note that we report here only the topic with the highest probability for the respective review); file path = path to the respective file in the IMDb dataset (Maas et al., 2011); bold with superscript indicates relation to SMEC, see interpretation below; italics indicate that text summarizes only content. Interpretation of superscripts (Schneider, 2017, unless indicated otherwise):

[1] refers to film-inherent features, SMEC: SV
[2] refers to film-inherent features, SMEC: SI
[3] refers to film-inherent features, SMEC: CI
[4] refers to film-inherent features, SMEC: FX (not mentioned in these examples)
[5] refers to film-external features, SMEC: RE
[6] refers to (anticipated) effects of use, SMEC: IN (not mentioned in these examples)
[7] refers to (anticipated) effects of use, SMEC: LH (not mentioned in these examples)
[8] refers to (anticipated) effects of use, SMEC: CS (not mentioned in these examples)
[9] refers to film-inherent features: 'soundtrack' was mentioned as a criterion during the SMEC development and part of the initial item pool (see Schneider, 2012a, Appendices A and B)
[10] refers to film-peripheral features: 'production' was mentioned as a criterion during the SMEC development (see Schneider, 2012a, Appendix A)
[11] refers to film-inherent features: 'performance of actor' was mentioned as a criterion during the SMEC development and part of the initial item pool (see Schneider, 2012a, Appendices A and B)

5. Discussion and Conclusion

We started this exploratory journey by asking what online movie reviewers write about and whether those online reviews provide insights into underlying SMEC. To address these questions, we applied correlated topic modeling to a large IMDb dataset.

We found 14 most prevalent topics in 35,136 online movie reviews that tapped into three major categories—HE, AP, and NA—and indicated what reviewers mostly

wrote about. A more detailed qualitative analysis of the reviews revealed that users do indeed mention certain SMEC, for example, SV, SI, CI, or RE. However, the focus of the online movie reviews as revealed by the topic model remains on the three overarching topic categories that only cover two SMEC: SI and LH.

Another finding is that top words in almost every topic represent affective reactions. This comes as no surprise because affective responses often represent the valence of a judgment and play an important role in movie evaluation (Schneider et al., in press). However, affective words in a written online movie review reflect not only evaluative judgments but also motivations of the writers. For instance, writing online reviews also fulfills an approval utility for the reviewers, enabling them to enhance themselves by signaling "a kind of connoisseurship or a level of social status that can become important to one's selfconcept" (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004, p. 43). IMDb quantifies this approval, for example, through ranking reviews by their rated helpfulness or the prolificacy of the reviewer. In general, if reviews contained positive emotional content, readers considered them as more helpful (Ullah, Zeb, & Kim, 2015). Further motivations that can lead to affective elements in reviews are concern for other consumers (e.g., intending to warn them) or the venting of negative feelings (Hennig-Thurau et al., 2004).

Besides these contributions of the present research, there are some limitations. Most of them concern the IMDb reviews and the specific dataset we used (Maas et al., 2011). First—and perhaps most problematic for automatic text mining—online movie reviews on IMDb vary in many aspects that may have introduced noise to our approach. Most crucial is the fact that critiques of a movie and summaries of its content are inextricably interwoven (for a review that contains a large part of content summary, see e.g., the second movie review in Table 2). Second, the IMDb dataset that we used comprises movies with a wide range of quality. Whereas most participants in the SMEC studies had specific and typical movies in mind when answering the items, the database we drew on also largely included mediocre and rare exemplars. Reviewers may have applied different criteria to qualitatively diverse movies. Some preliminary evidence supports this possibility. For instance, individuals named different criteria depending on whether they had to think about good, bad, or typical exemplars of a dramatic movie (Vogel & Gleich, 2012, Study 2). Second, some of the reviews dealt with TV shows or documentaries (e.g., The 74th Annual Academy Awards or Wrestling matches). These media types are not covered by SMEC. As this information was not available in the original dataset, it was not possible to exclude non-movie media types. To deepen our knowledge about this issue and get more details, we gathered meta-data of the respective items via OMDb API (this newly created dataset may also be helpful for future research and is available via OSF: https://doi.org/10.17605/OSF.IO/KA5D8). We found that 92% of the reviews in Maas et al.'s (2011) dataset actually referred to movies, rendering this limitation marginal. Third, the dataset included up to 30 reviews per movie. Thus, some plots and their descriptions could be overrepresented in the sample. However, given this very large dataset including 50,000 reviews and over 13,000 movies, this should not lead to an imbalance.

Movie evaluation criteria frequently appeared in online movies reviews. The number of criteria mentioned easily exceeded the eight SMEC dimensions as can be seen in the two examples in Table 2. However, they provide some support for content validity. Thus, another way to start developing items to measure SMEC could have been based on online movie reviews. The latent semantic variables, or topics, comprehensively summarized the content of the reviews and, using three broad categories, can be described as HE, AP, and NA. These categories resemble some of the SMEC (i.e., SI and LH), showing partial support for their construct validity but not for others (e.g., SV or CS).

Based on the conceptual framework of SMEC, we were interested in what users write about in online movie reviews and whether this could provide some insights into movie evaluation criteria from a different perspective than traditional self-report. However, after inspecting and interpreting the results of the topic models, we found that some criteria were more prevalent than others. This is perhaps also due to some slightly different goals of the research projects: Whereas the construction of the SMEC scales aimed to identify interindividual differences in what criteria viewers use when they evaluate movies, the present article examined what users write about in online movie reviews and what the most important topics are. Thus, reporting SMEC and applying them while writing about movies have a great deal of common ground but can, nevertheless, also lead to deviations. In short, we did not start with the idea that an unsupervised machine learning approach to movie reviews would result in exactly the same eight criteria that had previously been found in SMEC research based on self-reports. Nevertheless, we were hoping for some unsupportive or supportive insights into movie evaluation criteria.

Although it is hardly possible to explicitly state a priori hypotheses or expectations and test those against the results of a topic model, we think that our findings may spark interest in further assessing the usefulness of computational approaches to additionally explore previous research findings from a different angle or, if possible, to incorporate such procedures during scale development.

Future research could test several alternative computational methods to shed light on the specific SMEC that we could not find on the level of topics and broader categories and to further explore online movie reviews from different angles (for a concise overview, see Günther & Quandt, 2016). For instance, rule-based text extraction can help to refine an initial dataset by eliminating non-evaluative parts such as content summaries (e.g.,

Simmons et al., 2011). Building and validating a reliable movie criteria dictionary or using supervised machine learning to classify movie criteria based on manually labeled text could be another tool for computational SMEC research. The results of our study might be useful to plan such future analysis. However, this needs considerable effort and is probably not yet advisable because the SMEC construct itself is, as outlined in the introduction of this article, in need of further validation beyond the field of self-reports. To resolve this dilemma, future research endeavors that could be more deductive or supervised may draw on specific wordings of the SMEC scale items or on the preliminary coding scheme that has been developed during the qualitative phases of the SMEC construction (Schneider, 2012a). This information may then help to provide a gold standard for coders.

Besides choosing between unsupervised or supervised approaches, the predictive value of applied models could gain more attention in future. Although often examined outcome variables such as box-office success are often the focus of media economists but not of communication processes or effects research, a question such as how well can detected topics predict the evaluation of a movie on quantitative measures (e.g., star rating), follow-up communication (e.g., sharing or recommending a movie), or consumer choice (e.g., selecting the next movie) should matter to entertainment scholars. Moreover, the predictive validity can be used to compare different models and approaches and improve them (e.g., Amplayo & Song, 2017). Our newly created dataset provides the opportunity to engage in some of these analyses (e.g., using topics to predict box-office success, different types of ratings, or genre classification) that were beyond the scope of this article.

And what about entertainment research in general? Movies as entertainment fare have a long research tradition (e.g., Günther & Domahidi, 2017). Nowadays, it seems that economists, film studios, and online streaming providers—behind closed doors—have done much more applied work about movies than entertainment scholars have. This also becomes obvious when we take a look at the relevant marketing literature. For instance, Hennig-Thurau and Houston (2019) recently published an approximately 900-page book called *Entertainment Science* and summarize the field from an economist's perspective, while only marginally touching on recent advances in entertainment theory made by communication scholars and media psychologists (as summarized, e.g., in Vorderer & Klimmt, in press). On a macro level, a data-scientific and computational approach may bring these different disciplines closer together and recognize each other's achievements more thoroughly. It may not only be scholarly work (e.g., Taneja, 2016) that benefits but also entertainment industries that could learn from media and communication studies. If they interface with each other better, analyzing Big Data against a social-scientific background may help to improve recommender systems and user experiences within online review platforms, video streaming portals, or mixed-media channels. Although there are some notable but rare exceptions (e.g., Oliver, Ash, Woolley, Shade, & Kim, 2014), most entertainment researchers have not taken full advantage of the digital traces or responses that are publicly available online. Utilizing these data and applying computational methods to address open questions or supplement previous research could be a crucial factor for advancing both movie evaluation research and entertainment theory.

Acknowledgments

The publication of this article was funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the University of Mannheim, Germany.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Amplayo, R. K., & Song, M. (2017). An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews. *Data & Knowledge Engineering*, *110*, 54–67. https://doi.org/10.1016/j.datak.2017.03.009

Bader, N., Mokryn, O., & Lanir, J. (2017). Exploring emotions in online movie reviews for online browsing. In G. A. Papadopoulos, T. Kuflik, F. Chen, C. Duarte, & W.-T. Fu (Eds.), *IUI'17: Companion of the 22nd International Conference on Intelligent User Interfaces: March 13–16, 2017, Limassol, Cyprus* (pp. 35–38). New York, NY: The Association for Computing Machinery.

Bartsch, A. (2012). Emotional gratification in entertainment experience: Why viewers of movies and television series find it rewarding to experience emotions. *Media Psychology*, *15*(3), 267–302. https://doi.org/10.1080/15213269.2012.693811

Benoit, K., & Matsuo, A. (2020). spacyr: Wrapper to the 'spaCy' 'NLP' library [Computer software]. Retrieved from https://cran.r-project.org/web/packages/spacyr/index.html

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In A. N. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering and applications* (pp. 71–94). Boca Raton, FL: CRC Press.

Burtzlaff, J., Schneider, F. M., & Bacherle, P. (in press). Kamera ab! Einfluss der Beobachtungssituation auf das Rezeptionserleben [And action! The effect of being observed on reception processes]. In J. Vo-

gelgesang, J. Matthes, C. Schieb, & T. Quandt (Eds.), *Beobachtungsverfahren in der Kommunikationswissenschaft* [Observational methods in communication science]. Cologne: Herbert von Halem.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944–957. https://doi.org/10.1287/mksc.1100.0572

Eliashberg, J., Jonker, J.-J., Sawhney, M. S., & Wierenga, B. (2000). MOVIEMOD: An implementable decision support system for pre-release market evaluation of motion pictures. *Marketing Science*, *19*(3), 226–243. https://doi.org/10.1287/mksc.19.3.226.11796

Greenberg, B. S., & Busselle, R. W. (1996). Audience dimensions of quality in situation comedies and action programmes. In S. Ishikawa (Ed.), *Quality assessment of television* (pp. 169–196). Luton: University of Luton Press/ John Libbey Media.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(13). https://doi.org/10.18637/jss.v040.i13

Gunter, B. (1997). An audience-based approach to assessing programme quality. In P. Winterhoff-Spurk & T. H. A. van der Voort (Eds.), *New horizons in media psychology: Research cooperation and projects in Europe* (pp. 11–34). Wiesbaden: Westdeutscher Verlag.

Günther, E., & Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. *International Journal of Communication*, *11*, 3051–3071.

Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism*, *4*(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52. https://doi.org/10.1002/dir.10073

Hennig-Thurau, T., & Houston, M. B. (2019). *Entertainment science: Data analytics and practical theory for movies, games, books, and music*. Cham: Springer.

Himmelweit, H. T., Swift, B., & Jaeger, M. E. (1980). The audience as a critic: A conceptual analysis of television entertainment. In P. H. Tannenbaum (Ed.), *The entertainment functions of television* (pp. 67–106). Hillsdale, NJ: Erlbaum.

Hu, Y.-H., Shiau, W.-M., Shih, S.-P., & Chen, C.-J. (2018). Considering online consumer reviews to predict movie box-office performance between the years 2009 and 2014 in the US. *Electronic Library*, *36*(6), 1010–1026. https://doi.org/10.1108/EL-02-2018-0040

Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and independent subspace method. *Information Sciences*, *372*, 608–624. https://doi.org/10.1016/j.ins.2016.08.027

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lee, J. H., Jung, S. H., & Park, J. (2017). The role of entropy of review text sentiments on online WOM and movie box office sales. *Electronic Commerce Research and Applications*, *22*, 42–52. https://doi.org/10.1016/j.elerap.2017.03.001

Lee, S., & Choeh, J. Y. (2018). The interactive impact of online word-of-mouth and review helpfulness on box office revenue. *Management Decision*, *56*(4), 849–866. https://doi.org/10.1108/MD-06-2017-0561

Linton, J. M., & Petrovich, J. A. (1988). The application of the consumer information acquisition approach to movie selection: An exploratory study. In B. A. Austin (Ed.), *Current research in film: Audiences, economics and law* (Vol. 4; pp. 24–45). Norwood, NJ: Ablex.

Liu, Y., Yu, X., An, A., & Huang, X. (2013). Riding the tide of sentiment change: sentiment analysis with evolving online reviews. *World Wide Web*, *16*, 477–496. https://doi.org/10.1007/s11280-012-0179-z

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Portland, OR: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1015

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2/3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge University Press.

Murzintcev, N., & Chaney, N. (2020). ldatuning: Tuning of the latent Dirichlet allocation models parameters (Version 1.0.0) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/ldatuning/index.html

Neelamegham, R., & Jain, D. (1999). Consumer choice process for experience goods: An econometric model and analysis. *Journal of Marketing Research*, *36*(3), 373–386. https://doi.org/10.2307/3152083

Nikken, P., & van der Voort, T. H. A. (1997). Children's views on quality standards for children's television programs. *Journal of Educational Media*, *23*(2/3), 169–188. https://doi.org/10.1080/1358165970230206

Oliver, M. B., Ash, E., Woolley, J. K., Shade, D. D., & Kim, K.

(2014). Entertainment we watch and entertainment we appreciate: Patterns of motion picture consumption and acclaim over three decades. *Mass Communication and Society*, *17*(6), 853–873. https://doi.org/10.1080/15205436.2013.872277

Oliver, M. B., & Bartsch, A. (2010). Appreciation as audience response: Exploring entertainment gratifications beyond hedonism. *Human Communication Research*, *36*(1), 53–81. https://doi.org/10.1111/j.1468-2958.2009.01368.x

Oliver, M. B., & Raney, A. A. (2011). Entertainment as pleasurable and meaningful: Identifying hedonic and eudaimonic motivations for entertainment consumption. *Journal of Communication*, *61*(5), 984–1004. https://doi.org/10.1111/j.1460-2466.2011.01585.x

Oliver, M. B., Raney, A. A., Slater, M. D., Appel, M., Hartmann, T., Bartsch, A., . . . Das, E. (2018). Self-transcendent media experiences: Taking meaningful media to a higher level. *Journal of Communication*, *68*(2), 380–389. https://doi.org/10.1093/joc/jqx020

Otterbacher, J. (2013). Gender, writing and ranking in review forums: A case study of the IMDb. *Knowledge and Information Systems*, *35*, 645–664. https://doi.org/10.1007/s10115-012-0548-z

Parkhe, V., & Biswas, B. (2016). Sentiment analysis of movie reviews: Finding most important movie aspects using driving factors. *Soft Computing*, *20*, 3373–3379. https://doi.org/10.1007/s00500-015-1779-1

R Core Team. (2020). R: A language and environment for statistical computing (Version 3.5.3) [Computer software]. Retrieved from https://www.r-project.org

Raney, A. A., & Bryant, J. (2020). Entertainment and enjoyment as media effect. In M. B. Oliver, A. A. Raney, & J. Bryant (Eds.), *Media effects: Advances in theory and research* (4th ed.; pp. 324–341). New York, NY: Routledge.

Raney, A. A., Oliver, M. B., & Bartsch, A. (2020). Eudaimonia as media effect. In M. B. Oliver, A. A. Raney, & J. Bryant (Eds.), *Media effects: Advances in theory and research* (4th ed.; pp. 258–274). New York, NY: Routledge.

Rinker, T. W. (2018). textclean (Version 0.9.3) [Computer software]. Retrieved from https://github.com/trinker/textclean

Rössler, P. (1997). Filmkritiker und Publikum: Diskrepanzen und Übereinstimmungen. Ergebnisse einer Befragung von Filmrezensenten und Kinogängern [Film critics and the public: Discrepancies and agreements. Results of a survey of film critics and moviegoers]. *Media Perspektiven*, *28*, 133–140.

RStudio Team. (2019). RStudio: Integrated development for R (Version 1.2.5033) [Computer software]. Retrieved from https://rstudio.com

Schneider, F. M. (2012a). *Measuring subjective movie evaluation criteria: Conceptual foundation, construction, and validation of the SMEC scales* (Unpublished Doctoral dissertation). Universität Koblenz–Landau, Landau, Germany. Retrieved from https://nbn-resolving.de/urn:nbn:de:hbz:lan1-7813

Schneider, F. M. (2012b). *The importance of being challenged: Subjective movie evaluation criteria and entertainment experiences with challenging movies*. Paper presented at the 62nd Annual Conference of the International Communication Association, Phoenix, AZ, USA.

Schneider, F. M. (2017). Measuring subjective movie evaluation criteria: Conceptual foundation, construction, and validation of the SMEC scales. *Communication Methods and Measures*, *11*(1), 49–75. https://doi.org/10.1080/19312458.2016.1271115

Schneider, F. M., Welzenbach-Vogel, I. C., Gleich, U., & Bartsch, A. (in press). How do people evaluate movies? Insights from the associative–propositional evaluation model. In P. Vorderer & C. Klimmt (Eds.), *The Oxford handbook of entertainment theory.* New York, NY: Oxford University Press.

Simmons, L. L., Mukhopadhyay, S., Conlon, S., & Yang, J. (2011). A computer aided content analysis of online reviews. *Journal of Computer Information Systems*, *52*(1), 43–55.

Soto-Sanfiel, M. T., & Vorderer, P. (Eds.). (2011). Entertainment=Emotion [Special issue]. *Journal of Media Psychology*, *23*(1).

Swanson, D. L. (1987). Gratification seeking, media exposure, and audience interpretations: Some directions for research. *Journal of Broadcasting & Electronic Media*, *31*(3), 237–254. https://doi.org/10.1080/08838158709386662

Taneja, H. (2016). Using commercial audience measurement data in academic research. *Communication Methods and Measures*, *10*(2/3), 176–178. https://doi.org/10.1080/19312458.2016.1150971

Thurstone, L. L. (1930). A scale for measuring attitude toward the movies. *Journal of Educational Research*, *22*, 89–94.

Ullah, R., Zeb, A., & Kim, W. (2015). The impact of emotions on the helpfulness of movie reviews. *Journal of Applied Research and Technology*, *13*(3), 359–363. https://doi.org/10.1016/j.jart.2015.02.001

Valkenburg, P. M., & Janssen, S. C. (1999). What do children value in entertainment programs? A cross-cultural investigation. *Journal of Communication*, *49*(2), 3–21. https://doi.org/10.1111/j.1460-2466.1999.tb02790.x

Vogel, I. C., & Gleich, U. (2012). *"…and the good guy dies in the end"—Viewers' mental representations of emotionally challenging movies*. Paper presented at the 62nd Annual Conference of the International Communication Association, Phoenix, AZ, USA.

Vorderer, P., & Klimmt, C. (Eds.). (in press). *The Oxford handbook of entertainment theory*. New York, NY: Oxford University Press.

Vorderer, P., Klimmt, C., & Ritterfeld, U. (2004). Enjoyment: At the heart of media entertainment. *Communication Theory*, *14*(4), 388–408. https://doi.org/

10.1111/j.1468-2885.2004.tb00321.x
Vorderer, P., & Reinecke, L. (2015). From mood to meaning: The shifting paradigm in entertainment research. *Communication Theory*, *25*(4), 447–453. https://doi.org/10.1111/comt.12082
Wartella, E., & Reeves, B. (1985). Historical trends in research on children and the media: 1900–1960. *Journal of Communication*, *35*(2), 118–133. https://doi.org/10.1111/j.1460-2466.1985.tb02238.x
Wirth, W., Hofer, M., & Schramm, H. (2012). Beyond pleasure: Exploring the eudaimonic entertainment experience. *Human Communication Research*, *38*(4), 406–428. https://doi.org/10.1111/j.1468-2958.2012.01434.x
Wolling, J. (2009). The effect of subjective quality assessments on media selection. In T. Hartmann (Ed.), *Media choice: A theoretical and empirical overview* (pp. 85–101). New York, NY: Routledge.
Yang, J., Yecies, B., & Zhong, P. Y. (2020). Characteristics of Chinese online movie reviews and opinion leadership identification. *International Journal of Human-Computer Interaction*, *36*(3), 211–226. https://doi.org/10.1080/10447318.2019.1625570
Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*(4), 720–734. https://doi.org/10.1109/TKDE.2010.269
Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. In P. S. Yu, V. Tsotras, E. Fox, & B. Liu (Eds.), *CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. CIKM 2006 Workshops; November 6—11, 2006, Arlington, VA, USA* (pp. 43–50). New York, NY: ACM Press.

About the Authors

Frank M. Schneider (PhD, University of Koblenz–Landau, 2012) has been a Postdoctoral Researcher at the Institute for Media and Communication Studies, University of Mannheim, Germany, since 2013. From 2014–2015, he held an Interim Professorship in Communication Science at the University of Hohenheim. His research interests include digital communication, entertainment, political communication, communication processes and effects, and research methods.

Emese Domahidi (PhD, University of Münster, 2015) has been an Assistant Professor for Computational Communication Science at the TU Ilmenau, Germany, since 2017. Her main research interests include (cognitive) biases in digital media, well-being and social consequences of online media use, and computational methods for communication science.

Felix Dietrich is an MA Student and Research Assistant at the Institute for Media and Communication Studies, University of Mannheim, Germany. His research interests include digital media and entertainment, online privacy, political communication, and computational social science.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 164–179
DOI: 10.17645/mac.v8i3.3155

Article

A Graph-Learning Approach for Detecting Moral Conflict in Movie Scripts

Frederic René Hopp, Jacob Taylor Fisher and René Weber *

Media Neuroscience Lab, Department of Communication, University of California Santa Barbara, Santa Barbara, CA 93106, USA; E-Mails: fhopp@ucsb.edu (F.R.H.), jacobtfisher@ucsb.edu (J.T.F.), renew@ucsb.edu (R.W.)

* Corresponding author

Submitted: 14 April 2020 | Accepted: 24 June 2020 | Published: 13 August 2020

Abstract
Moral conflict is central to appealing narratives, but no methodology exists for computationally extracting moral conflict from narratives at scale. In this article, we present an approach combining tools from social network analysis and natural language processing with recent theoretical advancements in the Model of Intuitive Morality and Exemplars. This approach considers narratives in terms of a network of dynamically evolving relationships between characters. We apply this method in order to analyze 894 movie scripts encompassing 82,195 scenes, showing that scenes containing moral conflict between central characters can be identified using changes in connectivity patterns between network modules. Furthermore, we derive computational models for standardizing moral conflict measurements. Our results suggest that this method can accurately extract moral conflict from a diverse collection of movie scripts. We provide a theoretical integration of our method into the larger milieu of storytelling and entertainment research, illuminating future research trajectories at the intersection of computational communication research and media psychology.

Keywords
computational narratology; entertainment; eMFD; graph learning; MIME; moral conflict; movie scripts; network science

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

> Art is fire plus algebra. (Jorge Luis Borges)

Humans invest a large amount of their time and money engaging with fictional narratives. These narratives may be described as "cohesive and coherent stor[ies] with an identifiable beginning, middle, and end that provide information about scene, characters, and conflict; raise unanswered questions or unresolved conflict; and provide resolution" (Hinyard & Kreuter, 2007, p. 778). Fictional narratives on TV, like *Game of Thrones*, attract more than 44 million views per episode (Everett & Crockett, 2019), and screenwriters create up to 15,000 movie scripts per year (Eliashberg, Elberse, & Leenders, 2006). Successful movies can rake in billions of dollars. *Avengers: Endgame* achieved a record-breaking, worldwide gross of nearly three billion dollars (Box Office Mojo, 2019). Though, not all narratives are created equal. A majority of movies produced for entertainment underperforms, leaving producers and studios with a loss, and one of the main reasons for underperformance is a bad story (Eliashberg et al., 2006). Clearly, decision makers in the film industry face a critical decision when "choos[ing] among thousands of scripts to decide which ones to turn into movies" (Eliashberg, Hui, & John Zhang, 2007, p. 881).

But what exactly makes a 'good' story? Attempts to answer this question date back to the beginnings of storytelling itself (László, 2008), and communication and media psychology research have since formulated many different theories to explain narrative enjoyment (e.g.,

Oliver, 1993; Tamborini et al., 2013; Vorderer, Klimmt, & Ritterfeld, 2004; Zillman & Cantor, 1977). Some assert that a good narrative is all about the appeal and complexity of the characters (McKee, 2005), while others argue that a good story is mainly about a sequence of events (Franzosi, 2010). Recent evidence suggests that appealing stories emphasize interactions between the traits of characters, the events that befall them, and the specific sequence in which character traits and events are intertwined (Cron, 2012; Tamborini & Weber, 2020). In evolutionary media psychology, the fictional creation of narrative character and event chains has been hypothesized as an evolved principle to organize humans' cognitive machinery. According to this view, good stories are "attended to, valued, preserved, and transmitted because the mind detects that such bundles of representations have a powerfully organizing effect on our neurocognitive adaptations, even though the representations are not literally true" (Tooby & Cosmides, 2001, p. 21).

In this article, we assume that 'conflict,' and in particular 'moral conflict,' are crucial organizing principles permeating universally appealing narratives. In narrative psychology, conflict patterns are argued to undergird all fictional stories (László, 2008), elevating "the strength of any exciting character and story…without [conflict], characters don't have drive, desire, or desperation…there is no story, just words" (Ballon, 2014, pp. 49–51). Narratology further distinguishes external and internal conflict, with the former describing a character's goals and resistance from the environment (including interactions with other characters) and the latter occurring within a character and involving their internal needs (for an overview, see Szilas, Estupiñán, & Richle, 2018). Moral conflict—violating moral norms in order to uphold others (Tamborini, 2011, 2013)—has been related to the moral values of characters and to the ethical dimension of the story as a whole (Altman, 2008; McKee, 2005; Truby, 2007). In addition, moral conflict is regarded an especially salient component of cross-culturally appealing narratives (Knop-Huelss, Rieger, & Schneider, 2019; Lewis, Grizzard, Choi, & Wang, 2019; Lewis, Tamborini, & Weber, 2014; Tamborini, 2013; Weber, Popova, & Mangus, 2013). Yet, empirical assessments of the kinds of moral conflict that are present in existing, fictional narratives are absent, with only a few studies simulating agent-based moral conflicts in digital interactive storytelling (e.g., Battaglino & Damiano, 2014). This lack of knowledge hinders media and communication scholars to develop an inventory of moral conflict in narratives and stymies further research relating moral conflict and narrative appeal. As a first step toward filling this gap, we herein address two interlocking research questions:

RQ1: How can moral conflict be conceptualized and mathematized?

RQ2: How can complex narratives be abstracted to guide the detection of moral conflict?

To address these questions, we propose a scalable, computational approach to identify moral conflict in story lines of movie scripts. This approach considers narratives as an evolving network of relationships between characters, leveraging the community structure of this network to reveal important points in the narrative's arc. Using a total of 894 movie scripts encompassing 82,195 scenes, we show that scenes in which characters from different network modules interact with one another often contain morally-charged conflict that pushes the narrative forward.

2. Moral Conflict and Narrative Appeal

Moral conflict permeates human culture and history, manifested in early philosophical discussions between Plato and Socrates on the relative priority of repaying one's debts over protecting others from harm; in ancient literature, such as Aeschylus's *Agamemnon*, where the protagonist has to decide between saving his daughter or leading Greek troops to Troy; in modern philosophy, including Jean-Paul Sartre's tail of a student who is torn between his personal devotion to his mother and the attempt to contribute to the defeat of an unjust aggressor; and in recent debates on algorithmic judgment during moral dilemmas (Awad et al., 2018). Traditionally, moral conflict has been defined as conflict between moral requirements: Moral reasons for adopting an alternative such that "it would be morally wrong not to adopt that alternative if there were no moral justification for not adopting it" (Sinnott-Armstrong, 1988, p. 12). When these requirements conflict, a moral dilemma arises if neither requirement overrides the other (i.e., is "stronger overall in some morally relevant way," Sinnott-Armstrong, 1988, p. 20). According to this view, moral conflict can only be resolved if one moral requirement overrides the other, otherwise there exists a moral dilemma.

Mounting evidence suggests that moral conflict motivates a variety of actions (Weber & Hopp, in press) and is essential for the construction, processing, and evaluation of fictional narratives (Altman, 2008; Eden, Daalmans, & Johnson, 2017; Lewis et al., 2014). Stories featuring moral conflict are processed and appraised in a slower, more rational fashion, compared to stories that do not feature conflict, and are therefore processed fast and intuitively (Lewis et al., 2014). In the parlance of entertainment research, exposure to morally conflicting stories elicits higher levels of appreciation, an entertainment experience characterized by thought-provoking, deeper, and meaningful insights into life (Knop-Huelss et al., 2019; Lewis et al., 2014, 2019; Oliver & Bartsch, 2011). But why are audiences often drawn toward stories that prominently feature moral conflict, despite the cognitively effortful process of appraising and resolving these dilemmas? The Model of Intuitive Morality and Exemplars (MIME; Tamborini, 2011, 2013; for the latest update of the model see Tamborini & Weber, 2020) provides an answer to this question.

2.1. The Model of Intuitive Morality and Exemplars

The MIME describes short- and long-term reciprocal relationships between individuals' moral intuitions and their mediated and non-mediated environments (Tamborini, 2011, 2013). The MIME draws on Moral Foundations Theory (Graham et al., 2013) to conceptualize these moral intuitions. Moral Foundations Theory postulates the existence of five universal, innate moral sensibilities: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. The intuitive nature of these moral intuitions is described as "the sudden conscious appearance of a moral judgment, including an affective valence (good–bad, like–dislike) without any conscious awareness of having gone through steps of weighing evidence of inferring a conclusion" (Haidt, 2001, p. 818). Accordingly, the MIME suggests that media users experience an automatic, pleasurable response when their moral intuitions are reinforced, and an aversive reaction when these intuitions are violated, a prediction that has accumulated much evidence (e.g., Grizzard, Lewis, Lee, & Eden, 2011; Lewis et al., 2014; Tamborini et al., 2013; Weber, Tamborini, Lee, & Stipp, 2008).

The MIME asserts that individuals only exert these fast, pre-conscious evaluations in scenarios wherein a foundation is either upheld (e.g., care) or violated (e.g., harm) by character actions in a dominantly salient fashion (i.e., a moral foundation is distinctively upheld or violated). Conversely, within-foundation conflict (WFC) would arise if a foundation is both upheld and violated by a character's actions. For example, a character may need to enact (physical or emotional) harm in order to care for the (physical or emotional) well-being of another character. In contrast, between-foundation conflict (BFC) arises in situations in which different moral foundations are placed into conflict (Tamborini, 2011, 2013). For example, a character may physically harm innocent people (violating the care foundation) in order to be loyal to his/her group or country (upholding the loyalty foundation). To derive an affectively valenced response from these moral conflicts, individuals must decide whether the violation of one (or multiple) foundations as means of upholding another is justifiable in the scope of their individual moral preferences (also called moral intuition salience). While moral intuitions can exert chronic influence on moral judgments, stories may contain particular characters (e.g., highly moralized entities, such as Jesus or Ghandi) that may bias audiences' moral judgments in view of the character's general exemplification of particular foundations. For instance, a police officer may exemplify the fairness and care foundations in a story and at the same time may violate the authority foundation (subversion) by breaking the law (a 'bad cop' story). The initial exemplification may lead to a biased judgment of the police officer's motivation for violating another foundation.

3. Detecting Moral Conflict in Narratives

Attempts to computationally extract moral conflict from textual corpora are sparse. This is partially due to the latent, context-dependent nature of moral language that complicates moral information extraction (Weber et al., 2018) and missing operationalizations for moral conflict measurements. Recently, progress has been made toward the automated extraction of moral information from text (for an overview, see Hopp, Fisher, Cornell, Huskey, & Weber, in press). For example, Sagi and Dehghani (2014) combined dictionary-based content analysis with latent semantic analysis to identify moral rhetoric in news articles discussing the World Trade Center before and after the terror attacks on September 11th, blog posts and comments surrounding the Ground Zero Mosque, and speeches during the abortion debate in the US senate. Building on this work, Hopp et al. (in press) recently introduced the extended Moral Foundations Dictionary (eMFD) to computationally identify morally charged words in text. Compellingly, the eMFD is based on a large, crowd-sourced content-analysis study (Weber et al., 2018) and has been shown to outperform previous, word-count based moral extraction procedures.

We herein seek to build on this work to detect and mathematize moral conflict in narratives. To accomplish this aim, we theorize moral conflict as a set of hypotheses about the moral content and social network structure of narratives. We start with the simple assumption that moral conflict is likely to arise when characters from different groups, communities, or factions collide. Screenwriters frequently describe different communities of characters alternately, thereby constructing simultaneously evolving storylines (Weng, Chu, & Wu, 2007). Thus, as a story progresses, communities of characters come into focus as particular characters more frequently interact with each other. More specifically, the underlying cognitive process through which narrative consumers assign interacting characters to communities likely follows a stochastic, graph-learning process (Lynn & Bassett, 2019): The continuous registration of characters' spatiotemporal dependency leads viewers to develop expectations about characters' social relations, including their group affiliation, and hence expect particular characters to be more likely to 'flock together.' Based on this rationale and in line with narrative theory (Altman, 2008; Booker, 2004), we predict that narrative events in which characters from different communities collide highlight discrepancies across characters' group-based, moral motivations and thereby presage moral conflict (Tamborini & Weber, 2020).

To test these predictions, a computational translation is required that abstracts narratives into social network representations. In what follows, we make the bold claim that reducing a story to its main characters and their dialogue achieves a level of abstraction that enables computational learning of moral conflict.

3.1. Narratives as Social Networks of Characters in Movie Scripts

To examine characters and their interactions, mounting research draws on network science to conceptualize a narrative as an evolving social network in which nodes correspond to characters and edges between characters denote some dimension of their interaction (Ding & Yilmaz, 2010; Gleiser, 2007; Mac Carron & Kenna, 2012; Skowron, Trapp, Payr, & Trappl, 2016; Tran & Jung, 2015; Weng et al., 2007). This network representation of narratives has been fruitful in a variety of areas, from understanding character relations (Ding & Yilmaz, 2010; Park, Oh, & Jo, 2012) to identifying character types (Gleiser, 2007; Skowron et al., 2016) as well as learning leading roles, hidden communities, and storylines (Weng et al., 2007).

Due to their pre-structured format, movie scripts have been especially prominent for computationally constructing social networks of characters. Properly formatted movie scripts allow the distinction between several structural screenplay features such as scenes and stage directions, action descriptions, characters, and dialogue (Figure 1). Structural elements can subsequently be harnessed when computationally parsing a script, allowing one to extract, for example, which characters appear in the same scene or appear next to each other in dialogue. We utilize these structural features for computationally learning moral conflict patterns that permeate movie scripts.

4. Method

All data, code, analyses, and supplemental materials (SM) are made available at https://osf.io/rbdws/?view_only=24b117f22708457ca91f43ea2a4a6803

4.1. Movie Script Collection

We acquired scripts from the *ScriptBase J* dataset (Gorinski & Lapata, 2018), which contains 917 movie scripts. Each movie script is formatted in extended markup language (XML), with tags corresponding to particular script elements (e.g., speech, speaker, scene description, etc.). We wrote a custom Python script to query the XML tree for each script to extract the scene number, stage direction, speaker, and dialogue for each script. 63 movie scripts had to be discarded due to formatting errors, resulting in a total of 894 movie scripts spanning 82,195 scenes.

4.2. Character Dialogue Networks

To create the character dialogue network for each movie (Figure 3a), we divided the corresponding script into individual scenes. We created a node in the network for each unique character that has at least one line of dialogue in the script. We incremented the weight of the edge between two characters by one each time the two characters had lines of dialogue immediately adjacent to one another in the scene. For example, consider a scene containing the set of characters {*i*, *j*, *k*}, and the sequence of lines [*i*, *j*, *k*, *i*, *k*, *i*]. The network for this scene would contain one node for each character, and would contain three edges: an edge of weight one between *i* and *j*; an edge of weight one between *j* and *k*; and an edge of weight three between *i* and *k*. The final edge weight between two characters thus reflects how often the two characters appeared next to each other across the entire movie script.

Modularity is commonly used to detect sub-units or communities within a network. The extraction of these communities is based on computations that partition networks into classes that maximize the density of links between nodes inside that class compared to links between classes. Applied to our character dialogue networks (Figure 3b), we expect modularity to reveal communities among characters such that characters assigned to the same module are more likely to interact with each other in contrast to characters from different modules. To extract topological information about the characters and their relationships, we calculated network modularity using the Louvain modularity maximization algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

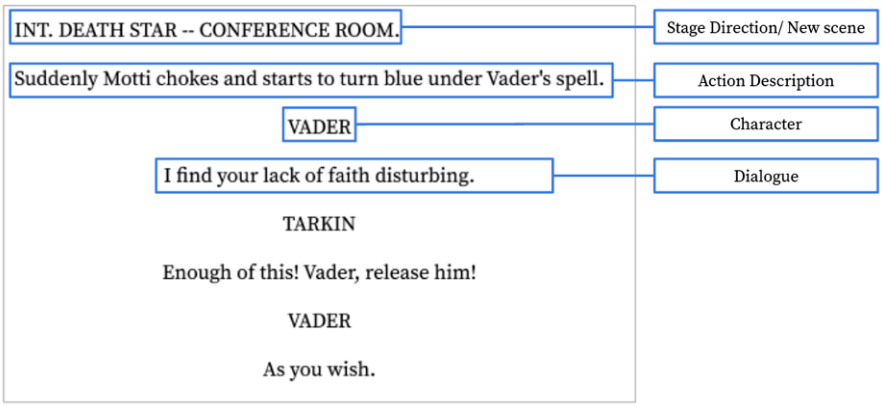


Figure 1. Elements and syntactic structure of a movie script excerpt from *Star Wars Episode IV: A New Hope* (Lucas, 1977).

4.3. Moral Conflict Scoring

4.3.1. Detection of Cross-Module and Within-Module Scenes

According to our network rationale for detecting conflict, we assume that the majority of conflicts—both moral and non-moral—occur between characters that belong to *different* communities as indexed by modularity. Hence, for each movie script, we label each scene as either a 'cross-module' scene if this scene contains characters from different modules or 'within-module' scene if this scene solely features characters from the same module (Figure 3c). Furthermore, we assert that the dialogue across character modules contains language cues that reveal: (a) whether the conflict has a moral or non-moral basis; and (b) how and which moral foundation(s) are conflicted.

4.3.2. Moral Content Extraction

To extract moral content from characters' dialogue, we utilized the extended eMFD (Hopp et al., in press; Weber et al., 2018). The eMFD contains word lists in which each word is assigned a vector of five probabilities, denoting the likelihood that this word belongs to any one of the five moral foundations. These probabilities were derived from crowd-sourced content annotations and have been shown to improve moral signal detection compared to extant moral content classification procedures.

In turn, we tokenized each character's dialogue and assessed which words in the dialogue are contained in the eMFD (Figure 3d). We then obtained the moral foundation probabilities for each detected word along with whether the given word denotes that the given foundation was 'upheld' (positive valence, e.g., care) or whether it was 'violated' (negative valence, e.g., harm). Finally, all probabilities for words in the upheld and violated categories were averaged, returning a vector of ten scores per scene—two (upheld versus violated) per foundation.

4.3.3. Moral Conflict Measurement

We operationalized moral conflict as a situation in which one foundation (or perhaps more than one) is violated in order to uphold the same foundation or other foundations (Tamborini, 2011, 2013). Thus, moral conflict can occur in two ways: WFC describes situations in which the same foundation is simultaneously upheld and violated, whereas BFC can occur in contexts where some moral foundations are upheld while others are violated.

Furthermore, we assert that the degree of moral conflict is a function of: (1) the magnitude to which a foundation is upheld and violated; and (2) the polarity (i.e., difference) of simultaneously upholding and violating a foundation. Magnitude signals the general moral relevance of a particular event (e.g., killing might have a stronger moral relevance than hitting), whereas polarity denotes the degree to which conflicting moral requirements are incomparable, such that one (or multiple) moral requirement cannot override each other as "neither is stronger than, weaker than, or equal in strength to the other" (Sinnott-Armstrong, 1988, p. 58; see also Tamborini, 2013). To better capture this relationship, we can visualize the above model on a two-dimensional grid, where one axis reflects the degree to which a foundation is upheld, with the other axis capturing the degree to which the same (or another) foundation is violated (see Figure 2). Accordingly, the degree of moral conflict on this grid is a function of: (a) the magnitude of the up-

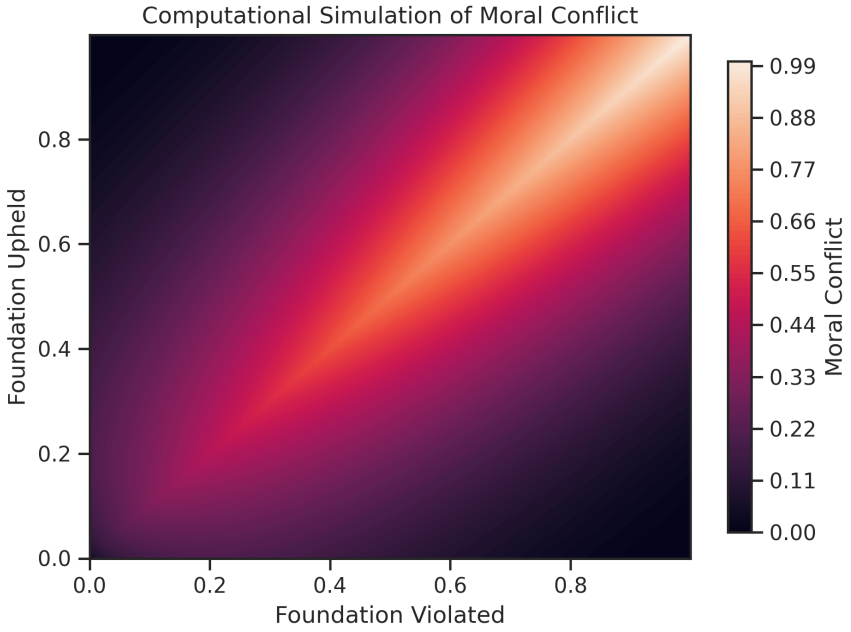


Figure 2. Computational model of moral conflict. Source: Authors.

held and violated foundation, where a higher magnitude results in moving further to the top right of the grid and increasing the degree of moral conflict; and (b) the polarity with which the moral requirements diverge from each other, with higher polarity indicating a greater distance to the diagonal of the grid. Mathematically, this relationship can be expressed as the Equation 1 shows below:

$$Moral\ Conflict = \sqrt{\frac{f_{i,u} + f_{i,v}}{2}} \times \left(1 - |f_{i,u} - f_{i,v}|\right)^2$$

Here, $f_{i,u}$ reflects the degree to which foundation *i* has been upheld and $f_{i,v}$ reflects the degree to which the same foundation *i* (in case of a WFC) has been violated. The first term expresses the average magnitude of upholding and violating a foundation: There will be no WFC if $f_{i,u}$ =0 or $f_{i,v} = 0$ or both $f_{i,u}$ and $f_{i,v} = 0$ (i.e., either only upholding or only violating a foundation, or neither upholding nor violating a foundation). Likewise, there will be maximum WFC if $f_{i,u} = 1$ and $f_{i,v} = 1$ (i.e., simultaneous, maximum upholding and violating a foundation). The second term integrates the polarity between virtue and vice: When a foundation is both maximally upheld and maximally violated ($f_{i,u} = 1$ and $f_{i,v} = 1$) the polarity between upheld and violated is maximized. In this scenario, there should not be any regularization of the moral conflict score and hence $f_{i,u} - f_{i,v} = 0$. However, if a foundation is relatively more upheld or violated, this should cause a decrease in moral conflict as either the upholding or the violation of the foundation becomes overridingly salient (Sinnott-Armstrong, 1988; Tamborini, 2011, 2013).

Building on this formula, we can express within- and between-foundation conflict in a straightforward manner. For WFC, the degree to which the same foundation is both upheld and violated can simply be computed by entering the virtue ($f_{i,u}$) and vice ($f_{i,v}$) score of the same foundation (e.g., care and harm; fairness and cheating; etc.) into Equation 1. Hence, five WFC scores can be computed that reflect the degree to which any one or all foundations are internally conflicted.

Analogously, we can compute the degree of BFC by entering the virtue ($f_{i,u}$) score of one foundation (e.g., care) and the vice ($f_{j,v}$) score of a different foundation (e.g., cheating). To obtain the average degree to which a particular foundation (e.g., care) is conflicted with all other foundations, the following Equation 2 can be applied:

$$BFC = \frac{\sum_{1 \le j \le 5; j \ne i} \sqrt{\frac{f_{i,u} + f_{j,v}}{2}} \times \left(1 - |f_{i,u} - f_{j,v}|\right)^2}{4}$$

In this equation, $f_{i,u}$ reflects the degree to which foundation *i* has been upheld (e.g., care) and $f_{j,v}$ reflects the degree to which a different foundation *j* has been violated (e.g., cheating). Accordingly, for every foundation, four conflict scores are calculated to measure the degree to which a single foundation has been upheld while violating any of the remaining four foundations. Averaging across these sums captures the mean degree to which a particular foundation has been upheld while violating all other foundations. WFC and BFC can also be expressed in terms of a conflict matrix *C* (see Table 1), where each cell reflects a computed moral conflict score. The diagonal of this matrix denotes the previously introduced WFC, whereas values on the off diagonal capture BFC. Summing the diagonal of this matrix produces the total WFC, whereas summing the upper and lower triangular elements of C produces the total BFC.

5. Results

We first provide a detailed, small-scale evaluation of our moral conflict detection algorithm across a set of three diverse movie scripts. Thereafter, we present results of a large-scale validation of our algorithm, scaling it up to a total of 894 screenplays.

5.1. Character Networks and Modularity

We first evaluated the computational construction of character networks and assignment of modularity

Table 1. Moral conflict matrix representing WFC and BFC.

i, *j*	Harm	Cheating	Betrayal	Subversion	Degradation
Care	$\sqrt{\frac{f_{i,u} + f_{i,v}}{2}} \times \left(1 - \lvert f_{i,u} - f_{i,v} \rvert\right)^2$ *	$\sqrt{\frac{f_{i,u} + f_{j,v}}{2}} \times \left(1 - \lvert f_{i,u} - f_{j,v} \rvert\right)^2$	.	.	.
Fairness	$\sqrt{\frac{f_{i,u} + f_{j,v}}{2}} \times \left(1 - \lvert f_{i,u} - f_{j,v} \rvert\right)^2$	.*			.
Loyalty	.		.*		
Authority	.			.*	
Sanctity	.				.*

Note: Cells with * represent WFC, and the ones without represent BFC.

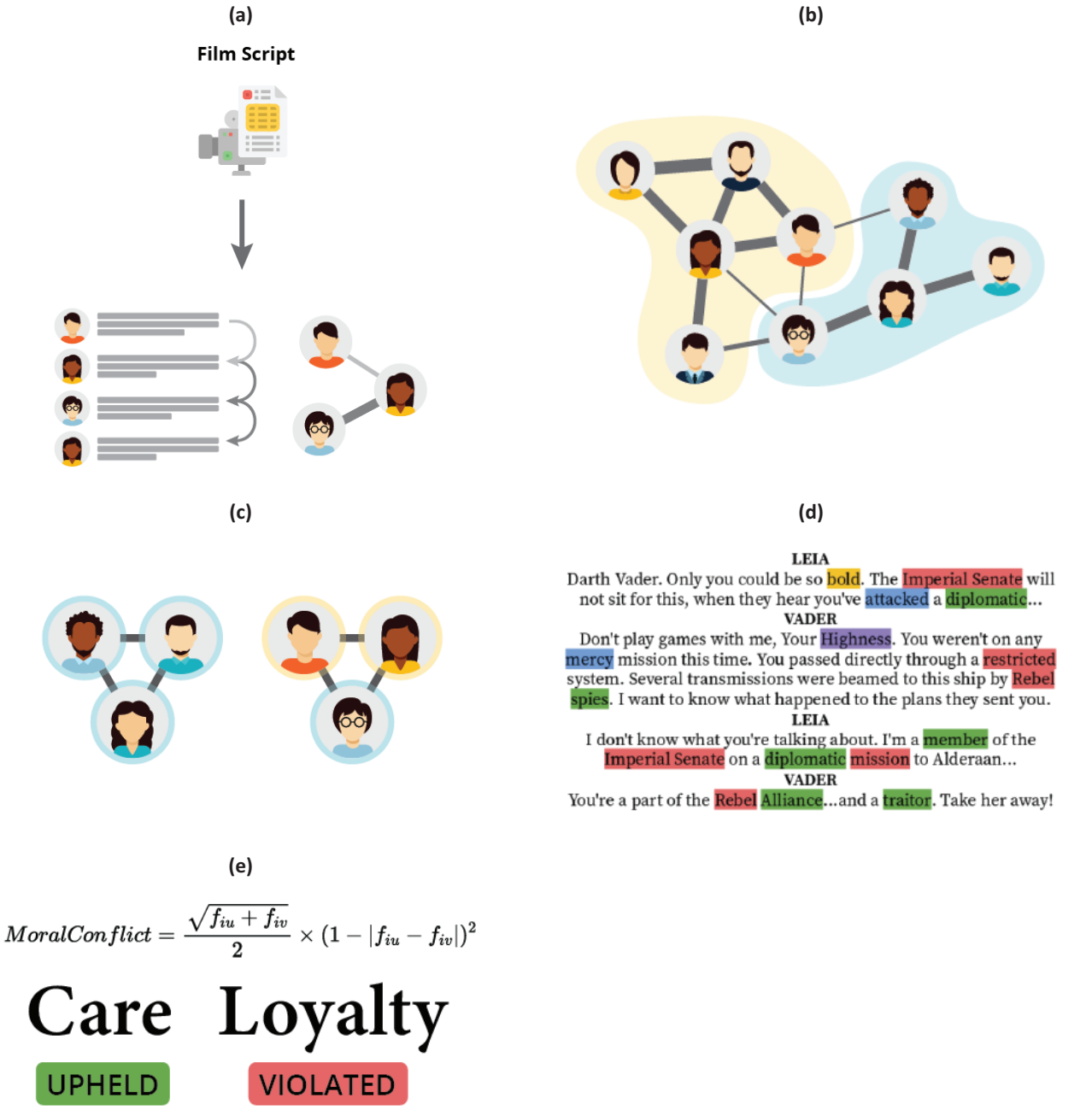


Figure 3. Flowchart of the Moral Conflict Detection Pipeline: **(a)** Network construction; **(b)** Modularity maximization; **(c)** Extracting within-module (left) and cross-module (right) scenes; **(d)** Coding moral content; **(e)** Calculating moral conflict. Source: Authors.

classes across three popular movie scripts. Figure 4 demonstrates that character networks provide a high-level summary of each movie, containing main characters and their relative interactions. Corroborating previous findings (Gleiser, 2007; Weng et al., 2007), modularity maximization performs well as a means of partitioning the character networks into different modules, such that characters belonging to the same module more frequently interact with each other compared to characters from a different module. For example, in *Star Wars Episode IV*, the largest module consists of the main hero (*Luke*) and his closest allies (*Ben*, *Han*, *Leia*, and *Threepio*), whereas the second largest module contains the main antagonist (*Vader*) along with his imperial allies (e.g., *Tarkin*, *Motti*, *Tagge*; Figure 4a). Similar 'good versus evil' communities are detected in *The Matrix* (Figure 4b). However, as the character network for *The Lord of the Rings: The Fellowship of the Ring* (Figure 4c) reveals, modules do not always reflect the social networks around heroes and villains, but rather capture which characters are co-present throughout alternately evolving storylines (Weng et al., 2007).

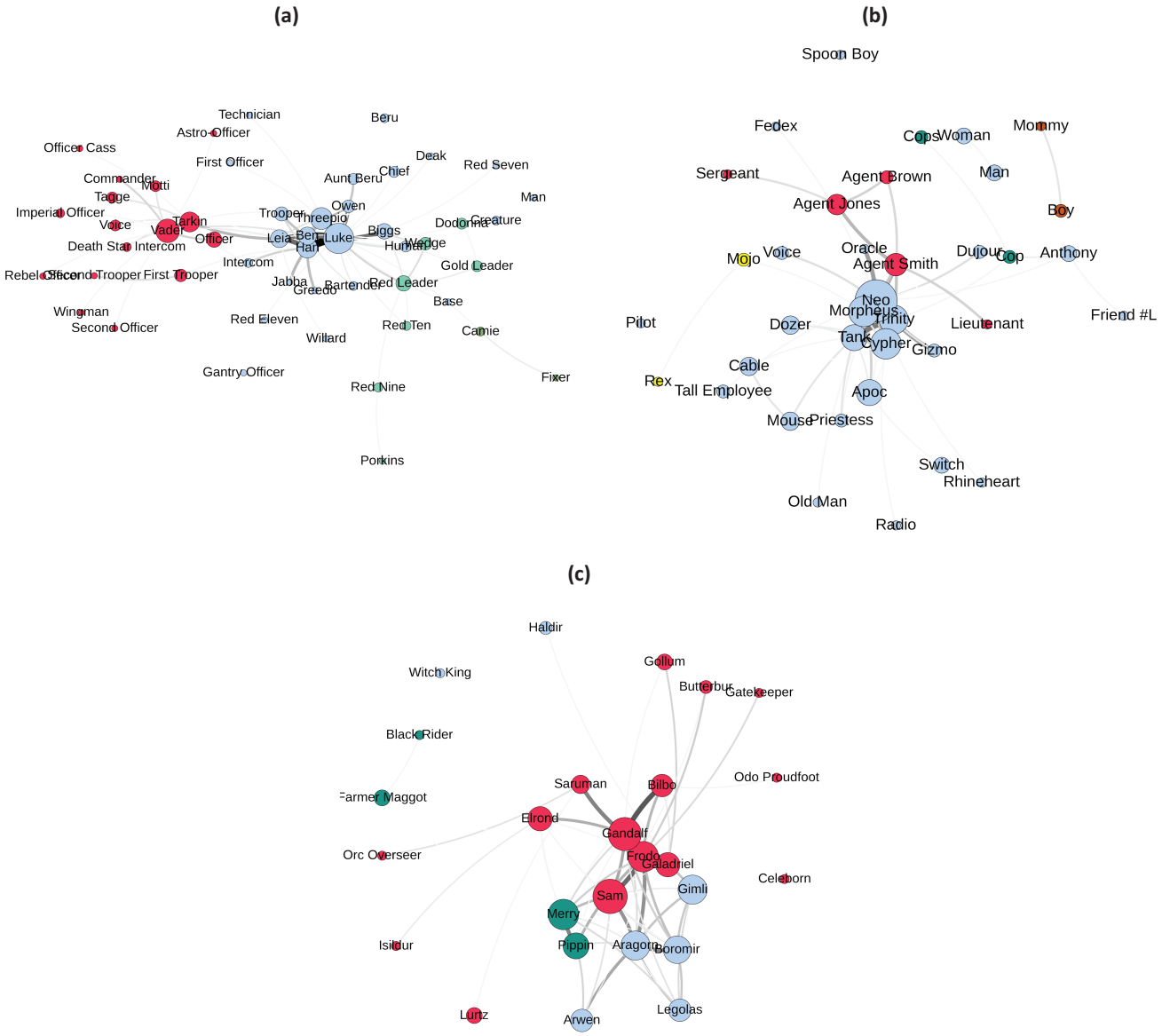


Figure 4. Character dialogue networks: **(a)** *Star Wars: Episode IV*: *A New Hope*; **(b)** *The Matrix*; **(c)** *The Lord of the Rings: The Fellowship of the Ring*. Notes: Node size reflects degree centrality, edge weight captures total co-occurrence in dialogue, node color illustrates modularity class. Networks are laid out using the force-directed Fruchterman-Reingold algorithm. See SM Sections 1 and 2 for scree plots and character dialogue adjacency matrices. Source: Authors.

5.2. Case-Study Evaluation of Moral Conflict

5.2.1. Star Wars Episode IV

A total of 20 cross-module scenes (Figure 5b) were detected for which we extracted the moral content across dialogues and calculated the degree of moral conflict (Figure 5b). Overall, the results are promising in that cross-module scenes indeed frequently contain moral conflict: For example, in scene 16, *Vader* and his allies are accusing *Leia* of passing through a restricted system and being part of the *Rebel Alliance*, which both are indicators of violating authority. Yet, *Leia* stresses the authority of the *Imperial Senate* to punish *Vader* for interfering in a diplomatic mission. Accordingly, as indexed by the scene's corresponding conflict matrix, there is strong WFC between authority and subversion. Moreover, in scene 89, *Leia* faces the BFC of whether to reveal the location of the hidden rebel base (betrayal, but following the authoritative order of *Tarkin*), or keeping loyal to her alliance at the cost of being partially responsible for the killing (harm) of innocent people should she not cooperate.

5.2.2. The Matrix

In *The Matrix*, a total of 13 cross-module scenes were identified (Figure 6A). In scene 19, for example, the highest WFC occurs within the loyalty foundation, denoting the emphasis on *Neo's* double-life as a program writer

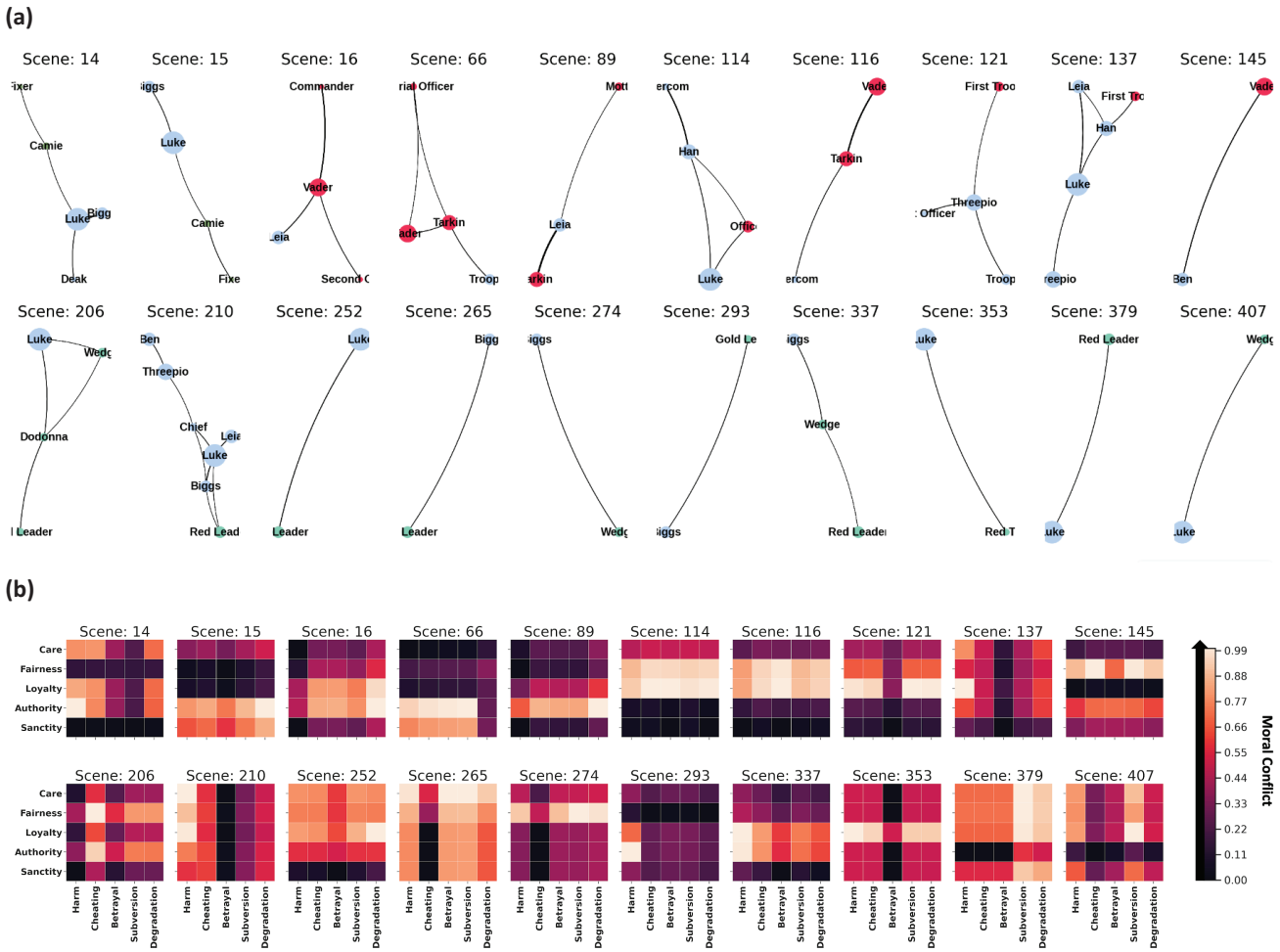


Figure 5. Moral conflict in *Star Wars Episode IV*: **(a)** Cross-module scenes and character interaction networks; **(b)** Moral conflict matrices. Source: Authors.

and a computer hacker who is "guilty of virtually every computer crime we have a law for" (Wachowski & Wachowski, 1999). Likewise, *Agent Smith* stresses that he "believe[s] [Neo] want[s] to do the right thing" (Wachowski & Wachowski, 1999) yet *Neo* emphasizes that he is denied his rights for not having the opportunity to make a phone call.

When evaluating the representation of moral conflict in scene 56, it becomes apparent that harm and degradation are most salient and conflicted: *Agent Smith* multiple times emphasizes the deal between him and *Cypher*, "I'll get you what you want" (Wachowski & Wachowski, 1999). While not explicitly mentioned in the dialogue, this scene clearly contains moral conflict: Deciding whether *Cypher* should betray his allies in return to being loyal to upholding the deal with *Agent Smith*.

5.2.3. The Lord of the Rings: The Fellowship of the Ring

In *The Lord of the Rings: The Fellowship of the Ring*, a total of 35 cross-module scenes were identified (Figure 7). Remarkably, the majority of cross-module scenes were high in (moral) conflict: The secret council meeting in *Rivendell* (scene 94), *Boromir*'s attempt to take the ring from *Frodo* (scene 133), the breaking of the fellowship and the death of *Boromir* (scene 136), and the appearance of the *Crebain* in the *Eregion Hills* (scene 98). For instance, in scene 94, the main conflict revolves around the appropriate manner in dealing with the one ring, which induced the highest conflict, both within and across fairness (e.g., who should be given the ring), loyalty, (e.g., who can be trusted with the ring), and authority (e.g., who decides what happens with the ring). Likewise, in scene 133, moral conflict emerges from *Boromir*'s attempt to steal the ring from *Frodo* (cheating), although assuring that he is no thief (fairness). Furthermore, *Boromir* accuses *Frodo* of betrayal, which explains the high degree of conflict surrounding the loyalty foundation.

5.3. Large-Scale Validation of Moral Conflict Detection Algorithm

The previous section demonstrated convergent validity of our moral conflict algorithm across three selected

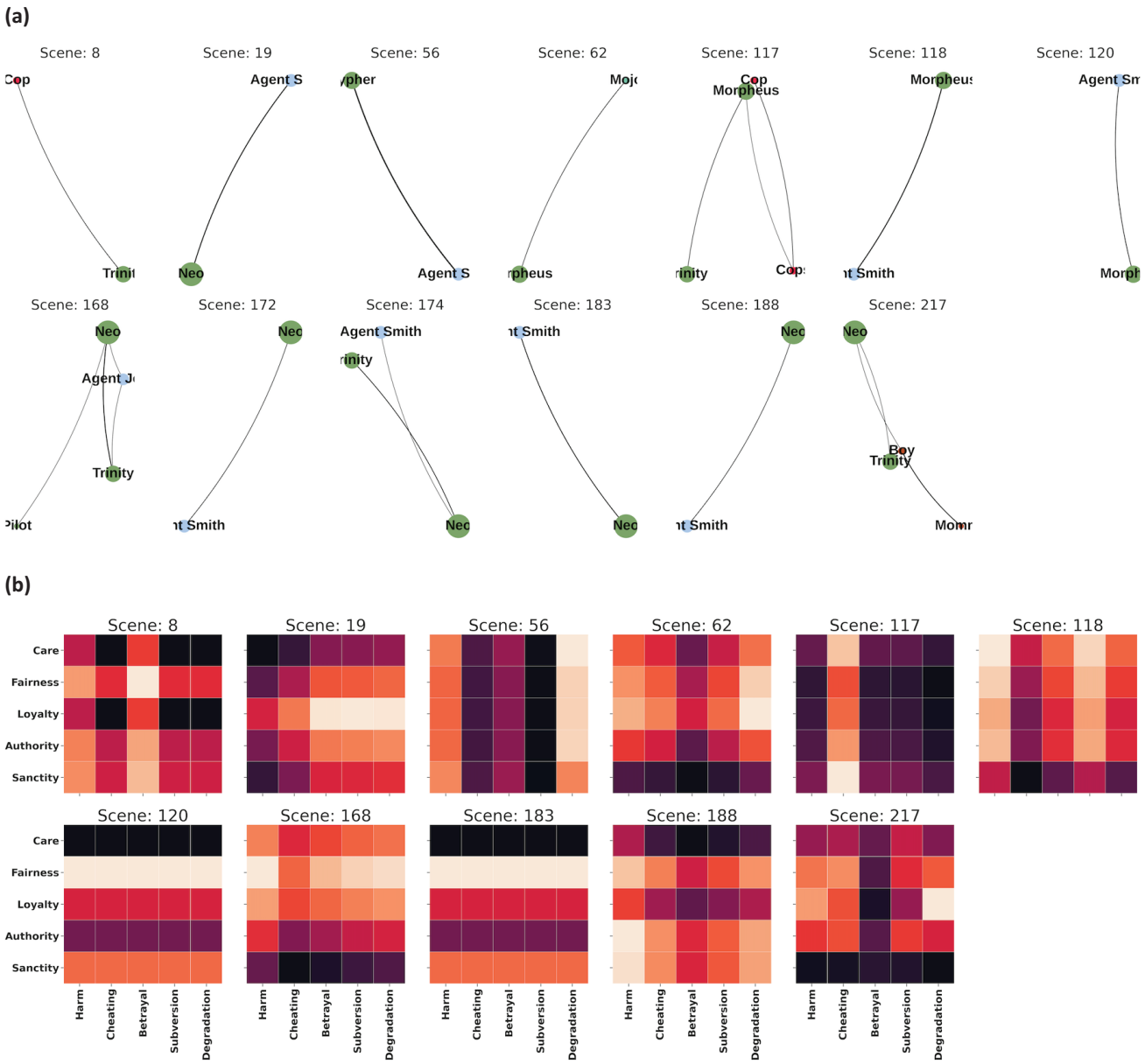


Figure 6. Moral conflict in *The Matrix*: **(a)** Cross-module scenes and character interaction networks; **(b)** Moral conflict matrices. Notes: Two cross-module scenes (172 and 174) were discarded for moral conflict computations as no moral content was identified. Source: Authors.

movies. To corroborate these findings across a larger set of screenplays, we scaled our algorithm to 894 movie scripts, containing a total of 82,195 scenes. First, we found a much higher frequency of within-module scenes (N = 69,042) compared to cross-module scenes (N = 13,153), although the ratio of cross-module to within-module scenes varies across movies (see Figure 3 in the SM). Furthermore, this result can be expected for two reasons: First, the modularity maximization algorithm assigns characters to modules that do not interact with each other frequently. As such, scenes in which characters from different modules are co-present will be rare. Second, screenplay writers frequently describe different communities of characters alternately, such that there will be a higher representation of scenes featuring characters from the same module compared to scenes in which characters from different modules are present.

Next, we tested the hypothesis that cross-module scenes are more moralized than within-module scenes. Compellingly, we indeed found that across moral foundation categories, cross-module scenes contain on average more moralized language as identified by the eMFD compared to within-module scenes (see Figure 8a). In a final step, we tested whether cross-module scenes are higher in moral conflict than within-module scenes. As illustrated in Figure 8b and Figure 8c, we again found support for our prediction that cross-module scenes are consistently higher in moral conflict than within-module scenes. Interestingly, we observed highest BFC for the loyalty foundation, which suggests that characters are

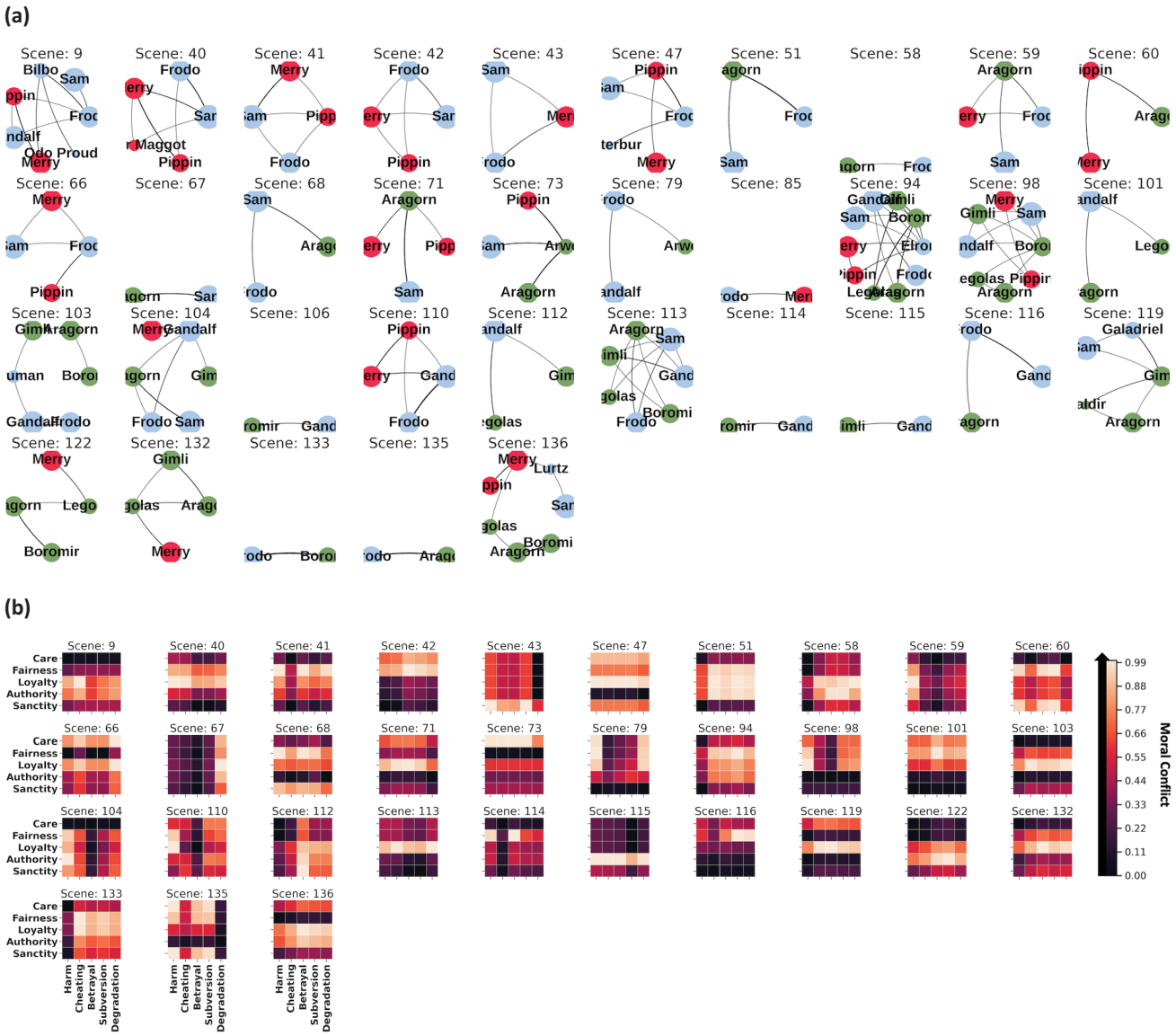


Figure 7. Moral conflict in *The Lord of the Rings: The Fellowship of the Ring*: **(a)** Cross-module scenes and character interaction networks; **(b)** Moral conflict matrices. Notes: Two cross-module scenes (85 and 106) were discarded for moral conflict computations as no moral content was identified. Source: Authors.

more concerned with upholding their in-group loyalty at the cost of violating other moral foundations. In contrast, we found that WFC is highest for the care foundation, indicating that characters frequently face situations in which they, for example, have to commit an act of (physical or emotional) harm in one context in order to uphold (physical or emotional) care in another. Likewise, we observed that BFC is higher than WFC across the loyalty and authority foundations, whereas WFC is higher in care and fairness foundations.

6. Discussion

This article introduced a computational approach for extracting moral conflict from movie scripts. While moral conflict is an important factor for story development (Ballon, 2014) and narrative appeal (Eden et al., 2017; Knop-Huelss et al., 2019; Lewis et al., 2014), no tools exist for the large-scale, standardized extraction of moral conflict from narratives. Our suggested approach provides a first step towards the automated, computational assessment of moral conflict in narrative texts. We found that network structures among characters serves as a useful heuristic for identifying movie scripts' main characters and their interactions. Furthermore, we demonstrated that modularity maximization is a promising method to extract communities of characters within a story. Lastly, we showed that scenes in which characters from different network modules co-occur are more moralized and higher in moral conflict than scenes in which characters from the same network module are co-present. In order to measure the degree and kinds of moral conflict that permeate the dialogue of rival characters within identified conflict scenes, we derived two different metrics:

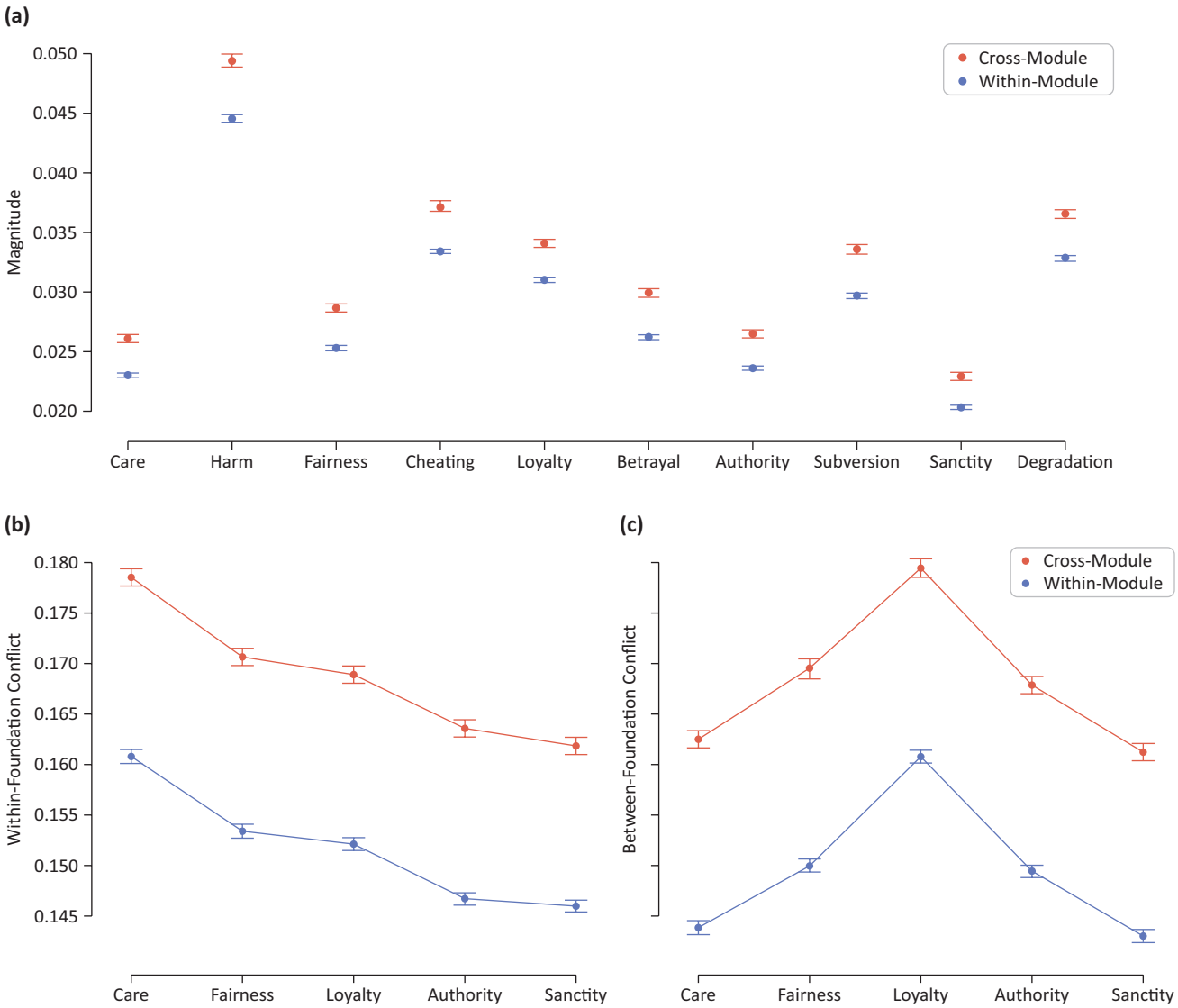


Figure 8. Moral content across within-module and cross-module scenes: **(a)** Representation of moral foundations; **(b)** Representation of within-foundation conflict; **(c)** Representation of between-foundation conflict. Notes: $N = 894$ screenplays. Error bars represent 99% confidence intervals estimated with 1,000 bootstrap iterations. Outliers above 95% were removed. Source: Authors.

First, we operationalized WFC as the degree to which distinct moral foundations are simultaneously upheld and violated; second, we defined BFC as the degree to which a single moral foundation is upheld, while other moral foundations are being violated. Our results indicate that these measures demonstrate acceptable validity as indicated by their alignment with expressed moral sentiment and conflict in characters' dialogue.

6.1. Implications

The herein introduced methodology is inspired by recent research synergizing structural and content features of narratives (e.g., Skowron et al., 2016) to push the envelope of computational entertainment research. Specifically, our finding that cross-module scenes are richer in moral content and moral conflict compared to within-module scenes is compelling for two reasons: First, it emphasizes the attentional capture of moral cues (Gantman & Van Bavel, 2015) for assisting viewers in directing their attention towards important plot points; second, the collision of characters from different network modules likely highlights characters' group-identity and reinforces motivations for coalition building and group cohesion. Hence, to emphasize distinctiveness between group affiliations, screenwriters may attribute characters in cross-module scenes with more moralized dialogue in order to highlight group differences and evoke moral conflict. This leads to the interesting future prediction whether: a) Character network modules are more homogenous in their moral language use; and b) whether this similarity increases during cross-module scenes where group-identity becomes hyper-salient.

Moreover, our herein developed approach paves the way for testing various central predictions of the MIME at an unprecedented scale. First, if moral conflict is indeed a central predictor of narrative appeal, then our developed conflict metrics should correlate with viewership and story evaluations, including movie performance as indexed by box office sales or film critic ratings (Weber, Hopp, & Fisher, 2020). Second, our approach may illuminate which kinds of moral conflict are more likely to engage particular morality subcultures: communities among audiences with distinct moral sensitivities (Mastro, Enriquez, & Bowman, 2013). Ongoing research demonstrates that subgroups among audiences differ in their moral sensitivities and hence tend to enjoy different moral content patterns (Bowman, Jöckel, & Dogruel, 2012). Accordingly, future work may examine how violating particular moral norms in order to uphold others is enjoyed in some morality subcultures but not in others. Accordingly, we envision that the evaluation of a story's moral conflict pattern may be of assistance to screenwriters during the script creation process and furthermore inform decision-makers in the film industry during content marketing (see e.g., The Moral Narrative Analyzer [MoNA], https://mona.medianeuroscience.org; Dramatica, https://dramatica.com; StoryFit, https://www.storyfit.com).

6.2. Limitations and Future Directions

Although our findings are promising, they have limitations. First, our approach does currently not allow for the tracing of the dynamic development of moral conflict related to specific characters, as it considers moral conflict as a holistic concept that is expressed across multiple characters within particular scenes. This limitation largely arises from the fact that the edges in our character networks are undirected and hence do not allow for an assessment of 'who is talking to whom, with what message, and with what effect.' Put differently, characters following each other in dialogue do not necessarily interact with each other, although we assume that dialogue co-occurrence is a more precise proxy for interaction compared to the commonly used scene co-occurrence (see e.g., Gorinski & Lapata, 2018; Weng et al., 2007). A consideration of the directed edges between particular characters may enable more in-depth analyses of moral conflicts between particular characters. As such, we are currently collecting human-annotated data to more precisely detect speaker–addressee relationships across movie script dialogue. Second, we herein applied modularity maximization to detect potential conflicts between different communities of characters. Hence, our approach does not account for potential conflicts that may arise *within* a given community. Future work may thus experiment with more fine-grained modularity parametrization to 'modularize within modules.' Relatedly, modularity maximization might not always extract modules that resemble classical 'good versus evil' splits as illustrated in *Star Wars Episode IV* or *The Matrix*. As such, future work may extend the herein introduced approach to detect stereotypical personas of film, such as heroes and villains (Bamman, O'Connor, & Smith, 2013). Analogously, we herein focus on external moral conflicts as expressed in dialogue between characters. Yet, recent advancements in the MIME (Tamborini & Weber, 2020) suggest that moral conflicts may also arise within characters, expressed as the collision of individuals' egoistic motivations and altruistic concerns for others. Moreover, our use-case of structured movie scripts reflects a 'best-case' scenario for computationally extracting moral conflict from text corpora. Yet, we think that many of the herein introduced steps can be mapped onto more unstructured texts to identify moral conflict in other media formats. For example, news articles may be promising to understand how moral conflict evolves in public discourse and dynamically shapes and is shaped by unfolding sociopolitical events (Hopp, Fisher, & Weber, 2020).

Perhaps most importantly, we encourage future researchers to further validate the herein established conflict computations. First, permutation tests in which words of a script are randomly assigned to scenes may corroborate our findings if our moral conflict matrices consistently deviate from null models. Likewise, it would be helpful to randomly permute characters to scenes in order to examine whether detected modules are preserved or destroyed. In addition, we suggest contrasting the computationally derived moral conflict measures against a crowd-sourced, manually annotated dataset. Doing so may reveal, among other things: (a) Whether humans attribute cross-module scenes higher conflict ratings compared to within-module scenes; or (b) whether algorithmically derived moral conflict matrices correlate with humans' judgment of conflicted moral foundations.

We are optimistic that future applications and enhancements of the computational algorithms can serve to help media psychology researchers—as well as industry professionals—decipher and learn the types of moral conflict that permeate stories, and how these conflicts influence how we engage with and respond to narratives in our everyday lives.

Acknowledgments

Contract grant sponsors: John Templeton Foundation (to R.W.), contract grant number: 61292. We extend our thanks to Cole Hawkins and Isaac Mackey for helpful discussions on moral conflict computations.

Conflict of Interests

The authors declare no conflict of interests.

References

Altman, R. (2008). *A theory of narrative*. New York, NY: Columbia University Press.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64.

Ballon, R. (2014). *Blueprint for screenwriting: A complete writer's guide to story structure and character development*. Mahwah, NJ: Routledge.

Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning latent personas of film characters. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (pp. 352–361). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P13-1035.pdf

Battaglino, C., & Damiano, R. (2014). A character model with moral emotions: Preliminary evaluation. In M. A. Finlayson, J. C. Meister, & E. G. Bruneau (Eds.), *5th workshop on computational models of narrative* (pp. 24–41). Wadern: Dagstuhl Publishing. https://doi.org/10.4230/OASIcs.CMN.2014.24

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Booker, C. (2004). *The seven basic plots: Why we tell stories*. London: Bloomsbury Academic.

Bowman, N. D., Jöckel, S., & Dogruel, L. (2012). A question of morality? The influence of moral salience and nationality on media preferences. *Communications*, *37*(4), 345–369.

Box Office Mojo. (2019). Avengers: Endgame. *Box Office Mojo*. Retrieved from https://www.boxofficemojo.com/title/tt4154796/?ref_=bo_se_r_1

Cron, L. (2012). *Wired for story: The writer's guide to using brain science to hook readers from the very first sentence*. Berkeley, CA: Ten Speed Press.

Ding, L., & Yilmaz, A. (2010). Learning relations among movie characters: A social network perspective. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer vision: ECCV 2010* (pp. 410–423). Heidelberg: Springer. https://doi.org/10.1007/978-3-642-15561-1_30

Eden, A., Daalmans, S., & Johnson, B. K. (2017). Morality predicts enjoyment but not appreciation of morally ambiguous characters. *Media Psychology*, *20*(3), 349–373. https://doi.org/10.1080/15213269.2016.1182030

Eliashberg, J., Elberse, A., & Leenders, M. A. (2006). The motion picture industry: Critical issues in practice, current research, and new research directions. *Marketing Science*, *25*(6), 638–661. https://doi.org/10.1287/mksc.1050.0177

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881–893. https://doi.org/10.1287/mnsc.1060.0668

Everett, J. A. C., & Crockett, M. J. (2019). What Game of Thrones reveals about moral decision-making. *Scientific American*. Retrieved from https://blogs.scientificamerican.com/observations/what-game-of-thrones-reveals-about-moral-decision-making

Franzosi, R. (2010). *Quantitative narrative analysis*. Thousand Oaks, CA: Sage.

Gantman, A. P., & Van Bavel, J. J. (2015). Moral perception. *Trends in Cognitive Sciences*, *19*(11), 631–633. https://doi.org/10.1016/j.tics.2015.08.004

Gleiser, P. M. (2007). How to become a superhero. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(9), P09020. https://doi.org/10.1088/1742-5468/2007/09/P09020

Gorinski, P., & Lapata, M. (2018). What's this movie about? A joint neural network architecture for movie content analysis. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long papers)* (pp. 1770–1781). Stroudsburg, PA: Association for Computational Linguistics.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *2013*(47), 55–130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Grizzard, M., Lewis, R., Lee, A., & Eden, A. L. (2011). Predicting popularity of mass market films using the tenets of disposition theory. *International Journal of Arts and Technology*, *4*(1), 48–60. https://doi.org/10.1504/IJART.2011.037769

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior*, *34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Hopp, F. R., Fisher, J., Cornell, D., Huskey, R., & Weber, R. (in press). The extended moral foundations dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*. Advance online publication. Retrieved from https://psyarxiv.com/924gq

Hopp, F. R., Fisher, J., & Weber, R. (2020). Dynamic transactions between news frames and sociopolitical events: An integrative, hidden Markov model approach. *Journal of Communication*, *70*(3), 335–355. https://doi.org/10.1093/joc/jqaa015

Knop-Huelss, K., Rieger, D., & Schneider, F. M. (2019). Thinking about right and wrong: Examining the ef-

fect of moral conflict on entertainment experiences, and knowledge. *Media Psychology*. Advance online publication. https://doi.org/10.1080/15213269.2019.1623697

László, J. (2008). *The science of stories: An introduction to narrative psychology*. Abingdon: Routledge.

Lewis, R. J., Grizzard, M. N., Choi, J. A., & Wang, P. L. (2019). Are enjoyment and appreciation both yardsticks of popularity? *Journal of Media Psychology: Theories, Methods, and Applications*, *31*(2), 55–64. https://doi.org/10.1027/1864-1105/a000219

Lewis, R. J., Tamborini, R., & Weber, R. (2014). Testing a dual-process model of media enjoyment and appreciation. *Journal of Communication*, *64*(3), 397–416. https://doi.org/10.1111/jcom.12101

Lucas, G. (Director). (1977). *Star Wars: Episode IV: A new hope* [Screenplay]. USA: Lucasfilm and Twentieth Century Fox.

Lynn, C. W., & Bassett, D. S. (2019). Graph learning: How humans infer and represent networks. *arXiv.org*. Retrieved from https://arxiv.org/abs/1909.07186

Mac Carron, P., & Kenna, R. (2012). Universal properties of mythological networks. *EPL (Europhysics Letters)*, *99*(2), 28002. https://doi.org/10.1209/0295-5075/99/28002

Mastro, D., Enriquez, M., & Bowman, N. D. (2013). Morality subcultures and media production: How Hollywood minds the morals of its audience. In R. Tamborini (Ed.), *Media and the moral mind* (pp. 99–116). New York, NY: Routledge.

McKee, R. (2005). *Story: Substance, structure, style, and the principles of screenwriting*. London: Methuen Publishing.

Oliver, M. B. (1993). Exploring the paradox of the enjoyment of sad films. *Human Communication Research*, *19*(3), 315–342. https://doi.org/10.1111/j.1468-2958.1993.tb00304.x

Oliver, M. B., & Bartsch, A. (2011). Appreciation of entertainment. *Journal of Media Psychology*, *2011*(23), 29–33. https://doi.org/10.1027/1864-1105/a000029

Park, S. B., Oh, K. J., & Jo, G. S. (2012). Social network analysis in a movie using character-net. *Multimedia Tools and Applications*, *59*(2), 601–627. https://doi.org/10.1007/s11042-011-0725-1

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review*, *32*(2), 132–144. https://doi.org/10.1177/0894439313506837

Sinnott-Armstrong, W. (1988). *Moral dilemmas*. Oxford: Basil Blackwell.

Skowron, M., Trapp, M., Payr, S., & Trappl, R. (2016). Automatic identification of character types from film dialogs. *Applied Artificial Intelligence*, *30*(10), 942–973. https://doi.org/10.1080/08839514.2017.1289311

Szilas, N., Estupiñán, S., & Richle, U. (2018). Automatic detection of conflicts in complex narrative structures. In R. Rouse, H. Koenitz, & M. Haahr (Eds.), *11th International Conference on Interactive Digital Storytelling, ICIDS 2018*. Cham: Springer.

Tamborini, R. (2011). Moral intuition and media entertainment. *Journal of Media Psychology*, *23*(1), 39–45.

Tamborini, R. (2013). Model of intuitive morality and exemplars. In R. Tamborini (Ed.), *Media and the moral mind* (pp. 43–74). London: Routledge.

Tamborini, R., Eden, A., Bowman, N. D., Grizzard, M., Weber, R., & Lewis, R. J. (2013). Predicting media appeal from instinctive moral values. *Mass Communication and Society*, *16*(3), 325–346. https://doi.org/10.1080/15205436.2012.703285

Tamborini, R., & Weber, R. (2020). Advancing the model of intuitive morality and exemplars. In K. Floyd & R. Weber (Eds.), *The handbook of communication science and biology* (pp. 456–469). New York, NY: Routledge.

Tooby, J., & Cosmides, L. (2001). Does beauty build adapted minds? Toward an evolutionary theory of aesthetics, fiction, and the arts. *SubStance*, *30*(1), 6–27. https://doi.org/10.1353/sub.2001.0017

Tran, Q. D., & Jung, J. E. (2015). CoCharNet: Extracting social networks using character co-occurrence in movies. *Journal of Universal Computer Science*, *21*(6), 796–815. https://doi.org/10.3217/jucs-021-06-0796

Truby, J.(2007). *Anatomy of story*. New York, NY: Faber and Faber.

Vorderer, P., Klimmt, C., & Ritterfeld, U. (2004). Enjoyment: At the heart of media entertainment. *Communication Theory*, *14*(4), 388–408. https://doi.org/10.1111/j.1468-2885.2004.tb00321.x

Wachowski, L., & Wachowski, L. (Directors). (1999). *The Matrix* [Screenplay]. USA: Warner Bros.

Weber, R., & Hopp, F. R. (in press). Moral emotions and conflict motivate actions. *Insights: Consumer Neuroscience in Business*.

Weber, R., Hopp, F. R., & Fisher, J. T. (2020). *Predicting movie performance from latent moral values in movie scripts*. Santa Barbara, CA: University of California.

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., . . . Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, *12*(2/3), 119–139. https://doi.org/10.1080/19312458.2018.1447656

Weber, R., Popova, L., & Mangus, J. M. (2013). Universal morality, mediated narratives, and neural synchrony. In R. Tamborini (Ed.), *Media and the moral mind* (pp. 50–66). New York, NY: Routledge.

Weber, R., Tamborini, R., Lee, H. E., & Stipp, H. (2008). Soap opera exposure and enjoyment: A longitudinal test of disposition theory. *Media Psychology*, *11*(4), 462–487. https://doi.org/10.1080/15213260802509993

Weng, C. Y., Chu, W. T., & Wu, J. L. (2007). Movie analysis based on roles' social network. In X. Zhuang & W. Gao

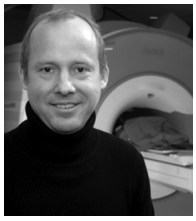
(Eds.), *2007 IEEE International Conference on Multimedia and Expo* (pp. 1403–1406). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICME.2007.4284922

Zillman, D., & Cantor, J. R. (1977). Affective responses to the emotions of a protagonist. *Journal of Experimental Social Psychology*, *13*(2), 155–165. https://doi.org/10.1016/S0022-*1031*(77)80008-5

About the Authors

Frederic René Hopp is a Doctoral Candidate in the Department of Communication at the University of California, Santa Barbara. His research advances theory and techniques for the algorithmic extraction of latent moral content and moral conflict from narratives, establishes computational models for predicting large-scale sociopolitical events and narrative performance, and illuminates the neural representations that undergird moral message processing. Frederic holds a BA in Media Psychology from the University of Mannheim and a MA in Communication from UCSB.

Jacob Taylor Fisher is a Doctoral Candidate in the Department of Communication at the University of California, Santa Barbara, and a Trainee of the National Science Foundation IGERT in Network Science and Big Data. He researches multimedia processing and media multitasking from a network neuroscience perspective. His current work investigates how certain digital environments can modulate attentional networks in the brain, and how these modulatory effects can be harnessed to develop novel treatments for cognitive processing disorders like ADHD.



René Weber received his PhD in Psychology from the University of Technology in Berlin, Germany, and his MD in Psychiatry and Cognitive Neuroscience from the RWTH University in Aachen, Germany. He is a Professor at UCSB's Department of Communication and director of UCSB's Media Neuroscience Lab (https://medianeuroscience.org). He was the first media psychology scholar to regularly use neuroimaging technology to investigate various media effects, from the impact of violence in video games to flow experiences, attention disorders, and the effectiveness of anti-drug PSAs. He has published four books and more than 140 journal articles and book chapters. His research has been supported by grants from national scientific foundations in the US and Germany, as well as through private philanthropies and industry contracts. He is a Fellow of the International Communication Association.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 180–190
DOI: 10.17645/mac.v8i3.3046

Article

(A)synchronous Communication about TV Series on Social Media: A Multi-Method Investigation of Reddit Discussions

Julian Unkel * and Anna Sophie Kümpel

Department of Media and Communication, LMU Munich, 80538 Munich, Germany; E-Mails: unkel@ifkw.lmu.de (J.U.), kuempel@ifkw.lmu.de (A.S.K.)

* Corresponding author

Submitted: 20 March 2020 | Accepted: 10 June 2020 | Published: 13 August 2020

Abstract
Audiences' TV series entertainment experiences are increasingly shaped not only by the events on the 'first screen' but also by discussions on social media. While an extensive body of research has examined practices of 'second screening,' especially on Twitter, online discussions before and after the live broadcast and on other platforms have received less attention. On Reddit—one of the most important platforms for Social TV—discussions often take place in temporally structured threads that allow users to discuss an episode before (pre-premiere thread), during (live premiere thread), and after (post-premiere thread) it airs. In this project, we examine whether these spaces mainly indicate temporal preferences among users or are associated with different usage practices and motives. To do so, we conducted two case studies of the Reddit community r/gameofthrones: a survey about usage motives (n = 417) and an automated content analysis of approximately 1.2 million comments left on the episode discussion threads in which we examined thread use over time, interactions between users, and discussion content. The results revealed differing usage motives and practices for the three thread types, illustrating the distinct function that these communication spaces fulfil for users.

Keywords
entertainment; multi-method; Reddit; second screen; social media; social TV; TV series; usage motives

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

In the midst of the third 'golden age of television,' we are currently experiencing a peak of quality serial television, which is characterized by complex stories and narrative structures, but also by a high degree of intertextuality and new usage and communication practices (Schlütz, 2016). In a convergent media environment, the entertainment experiences of TV series such as *The Walking Dead*, *Doctor Who*, and *Game of Thrones* (*GoT*) not only result from what happens on the 'first screen' (i.e., in the series itself), but are increasingly shaped by all the content and discussions with which users can interact on social media (Raney & Ji, 2017; Sutter, 2017). Drawing on Raney and Ji (2017, p. 428), we have reason to believe that communication about TV series on social media not only complements the TV entertainment experience, but represents "an entertainment experience, in and of itself." While funny memes or cynical comments cannot improve a bad episode, they do entertain recipients on a whole other level.

According to our review of the literature, Social TV research has mostly focused on second screening, which encompasses practices that occur "while watching television" (Gil de Zúñiga, Garcia-Perdomo, & McGregor, 2015, p. 793), thus ignoring mediated communication that hap-

pens prior to or after an episode airs. Considering the success of video-on-demand services (e.g., Netflix, Amazon Prime Video) and their associated viewing practices, this focus on parallel communication seems to reflect social reality less and less. Moreover, extant empirical studies seem to be centered around parallel communication on Twitter (e.g., Buschow, Schneider, & Ueberheide, 2014; Ji & Raney, 2015; Schirra, Sun, & Bentley, 2014), leaving us with research gaps regarding both the temporal and platform structures of Social TV activities. In order to address these gaps, this multi-method research project: a) focuses on parallel communication (i.e., second screening) and pre- and follow-up communication; and b) investigates the usage practices and motives of these three forms of communication on Reddit, which is one of the most important platforms for Social TV (Bentley, 2017). Empirically, this research builds on two case studies of the subreddit r/gameofthrones—a discussion board for talking about *GoT*—and combines a user survey (n = 417) with an automated content analysis of about 1.2 million user comments left in the temporally structured Reddit discussion threads (pre-premiere, live premiere, and post-premiere) about the show. This approach allows us to acquire differentiated insights into the interplay between Social TV users' motives and their actual communication practices.

2. Social TV: Communicating about TV Series on Social Media

In the literature, definitions of Social TV often focus on "real-time backchannel communication…during a live television broadcast" (Lim, Hwang, Kim, & Biocca, 2015, p. 158) or "technologies that permit synchronous social interactions" (Raney & Ji, 2017, p. 425), thus explicitly excluding asynchronous forms of communication. Drawing on Buschow et al. (2014, p. 131), we adopted a broader definition of Social TV that also includes communication practices that take place before and after a TV show airs (see Figure 1). For example, TV users might turn to Facebook groups to exchange predictions about the upcoming episodes of their favorite show (pre-communication), use hashtags to follow and engage in live discussions on Twitter (parallel communication), or log in to Instagram to share memes related to key events of the most recent episode (follow-up communication).

To our knowledge, most prior studies have investigated parallel communication on Twitter, which does not adequately reflect the plethora of Social TV activities. In fact, survey studies have indicated that users seem to prefer asynchronous forms of communication, particularly follow-up communication (Bentley, 2017, p. 125): Asked about their latest Social TV activity, 60% of respondents indicated that they have engaged in follow-up communication. Considering the rise of streaming services, such as Netflix and Amazon Prime Video, and the practices associated with them, such as 'binge-watching' (Jenner, 2017), asynchronous Social TV activities are likely to gain even more importance. The more TV series can be watched anytime and anywhere, the more crucial understanding the uses and motives of non-parallel Social TV experiences becomes.

2.1. Uses and Motives of Social TV

To understand how and why users engage in (a)synchronous Social TV activities, we built on empirical research conducted in the tradition of the uses and gratifications approach (Katz, Blumler, & Gurevitch, 1973; Rubin, 2002). Following this perspective, Social TV users can be characterized as "active, discerning, and motivated in their media use" (Quan-Haase & Young, 2010, p. 351), resorting to specific communicative means and platforms to address diverse psychological and social needs. Social media play an increasingly important role in fulfilling these needs, as the "water cooler moments" (Lochrie & Coulton, 2012, p. 199) associated with linear television—getting together with colleagues or acquaintances to discuss the previous night's episode—might become less frequent due to both the abundance of available shows and the mentioned changes in viewing practices (e.g., binge-watching). To put it more generally, if people want to communicate about a specific episode of a specific TV series at a specific time, social media offers targeted opportunities (Shim, Shin, & Lim, 2017, p. 340).

Previous research on people's motivations to engage in Social TV activities has exclusively focused on parallel communication (e.g., live tweeting). Researchers have conducted both qualitative (Han & Lee, 2014; Schirra et al., 2014) and quantitative (Gil de Zúñiga et al., 2015; Krämer, Winter, Benninghoff, & Gallus, 2015; Shim et al., 2017) studies, thereby revealing a diverse set of affective, cognitive, personal-integrative, and social motives. For example, people engage in parallel Social TV activities to experience companionship, share their experience with other viewers, receive and/or share background information, and display their own knowledge or wit. However, we do not yet know whether these motives are more or less important in the context of asynchronous pre- and follow-up communication.

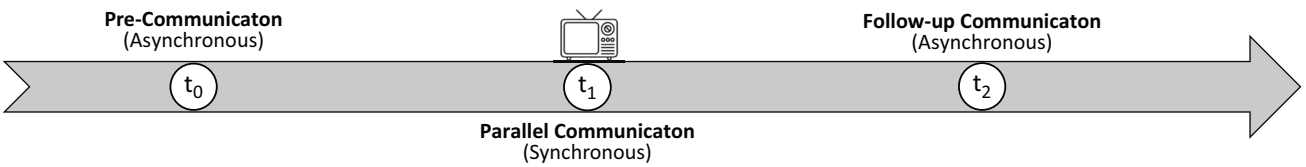


Figure 1. The Social TV experience: Forms of (a)synchronous communication about TV series on social media.

Thus, whereas potential differences between different (a)synchronous Social TV activities have—to our knowledge—not been addressed, there are indications that channel properties and related affordances affect the uses and motives of parallel Social TV activities (Han & Lee, 2014; Krämer et al., 2015). These studies show that using messaging apps, such as WhatsApp, is associated more with personal social needs, while using more public platforms, such as Twitter, is more strongly related to informational needs and a curiosity about other people's opinions. More generally, messaging apps seem to be used for private, in-depth conversations about TV series, whereas public social media platforms seem to be better suited for information seeking and sharing and getting a sense of a wider range of opinions. In line with these findings and the above-introduced temporal systematization of Social TV experiences, we assumed that people do not only choose specific channels/platforms in accordance with their needs but also prefer synchronous or asynchronous forms of communication for specific motives. For example, although there is mixed evidence on whether spoilers increase or decrease the enjoyment of narratives (Johnson & Rosenbaum, 2015; Leavitt & Christenfeld, 2013), it can be expected that some users specifically use forms of pre-communication to discuss expected developments in a show or, conversely, deliberately avoid forms of parallel communication on social media (e.g., by muting hashtags on Twitter) when they are unable to watch the live broadcast. However, it is still an open question whether meaningful differences exist in the uses and motives of different (a)synchronous Social TV activities. Thus, building on the discussed survey and interview studies (Gil de Zúñiga et al., 2015; Han & Lee, 2014; Krämer et al., 2015; Schirra et al., 2014; Shim et al., 2017), we aim to explore diverse motivations, including the informational, social, and entertainment-related needs for synchronous Social TV activities established in previous research while also considering needs that might be more germane to asynchronous pre- and follow-up communication (e.g., creating memes, discussing spoilers).

2.2. Reddit as a Social TV Platform

As stated in the introduction, our research is focused on the social media platform Reddit. As the self-proclaimed "front page of the Internet," Reddit is essentially an extensive collection of thematically structured forums (known as subreddits), in which users can submit own content (e.g., texts, links, pictures, videos) in the form of posts and/or comment either directly on these original posts (top-level comments) or reply to the comments of other users, thus creating a discussion thread (Widman, 2020; also see Figure A.1 in the Supplementary File). According to recent online traffic data, Reddit is the 18th most popular website in the world, with most of its desktop traffic originating from the US (49.9%), the UK (7.9%), and Canada (7.5%; Alexa, 2020; SimilarWeb, 2020). In addition to its popularity among (English-speaking) Internet users, Reddit has become an attractive data source (Amaya, Bach, Keusch, & Kreuter, 2019) because it is: a) mostly public and can be browsed by unregistered users; b) allows the identification and study of special or otherwise hard-to-recruit and -observe populations; and c) is perceived as suitable in investigations of users' "true beliefs" (Amaya et al., 2019, p. 2) due to its largely anonymous environment.

Even more pertinent to this research is the fact that Reddit is highly relevant to Social TV (Bentley, 2017; Cassis, 2018). In Bentley's survey (2017, p. 125), 47 % of respondents (19- to 69-year-old US Americans recruited via Amazon Mechanical Turk) said that they had turned to Reddit in the past month to find additional information about a TV show, actor, or character, making it the most important social media platform for such activities. Moreover, Reddit is ideal for simultaneously investigating asynchronous (pre-communication, follow-up communication) and synchronous (parallel communication) forms of communication. To facilitate separated discussions before (pre-premiere thread), during (live premiere thread), and after (post-premiere thread) an episode airs, many TV subreddits (e.g., those dedicated to *GoT*, *The Walking Dead*, and *Doctor Who*) offer temporally structured threads that mirror the conceptual tripartition of Social TV activities (see Figure 1). This tripartition of threads allows us to get a sense of usage motives and practices while keeping possible influences of platform affordances or user characteristics more or less constant.

Building on the prior discussion of the uses and motives of Social TV and given the scarcity of research on asynchronous Social TV activities on platforms other than Twitter, our project was guided by two overarching research questions:

> RQ1: Usage motives: What motivates users to engage in (a)synchronous communication about TV series on Reddit?
>
> RQ2: Usage practices: What characterizes (a)synchronous communication about TV series on Reddit?

3. Method

To address these research questions, we conducted a multi-method case study of r/gameofthrones, which is the largest *GoT* subreddit and has been among the most-frequented TV-centric subreddits, with about 11.5 million unique users per month in 2018 (Cassis, 2018). Episode discussion threads on r/gameofthrones have followed the aforementioned tripartition into pre-premiere, live premiere, and post-premiere discussion threads since the middle of Season 4. All three thread types are 'official' discussion threads created by the moderators or select subreddit users, meaning that every episode has a dedicated pre-premiere, live premiere, and post-premiere thread. For this case study, we drew from

two data sources that will be described in more detail below: the content and meta-textual features of all comments posted in the episode discussion threads for Seasons 5–8, and a survey of subreddit users conducted during the airing of the eighth and final season. The data and reproducible *R* analysis scripts can be obtained from this study's Open Science Framework (OSF) repository: https://osf.io/7v49t/.

3.1. Comments on Subreddit Discussion Threads

There are two main ways to access organic (i.e., user-generated) Reddit data: through the official Reddit application programming interface (API) and the Pushshift Reddit dataset (Amaya et al., 2019). While both generally provide the same data, we opted for the latter, as it provides a much easier and more flexible way of querying Reddit data (e.g., higher rate limits). The Pushshift Reddit dataset (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020) is a privately maintained dataset that ingests Reddit data in real time and contains all posts and comments made on the platform since June 2005. Data are split into two files (one for original posts and the other for comments) and may be accessed in a variety of ways, including monthly dumps downloadable directly from the Pushshift website and an open API. For this project, we wrote a simple *R* function that, for a given list of Reddit posts (identified via IDs that are also part of each post's URL), queries the Pushshift API and returns all comments posted in those threads, including associated meta-textual information (e.g., creation time, parent comment ID, comment score). Data are returned in JavaScript Object Notation format, which can be easily converted to various tabular data frame formats for further analysis. All API querying and data handling steps were conducted using various functions from the "Tidyverse" meta package (Wickham et al., 2019).

We queried the Pushshift API for all comments posted in the episode discussion threads for seasons 5–8, resulting in 1,252,971 individual comments. All queries were performed in September 2019, about three months after the broadcast of the final *GoT* episode. All comments either deleted by the original author or removed by a moderator between its original post date and the query date as well as all comments by discussion moderation bots were removed from the data, leading to a final sample of 1,203,666 comments from 243,997 unique users.

3.2. Survey of Subreddit Users

A survey of 417 users of r/gameofthrones was conducted from April 11, 2019—three days before the US premiere of the first episode of Season 8—to April 22, 2019. Prospective participants were recruited by posts and sponsored links on Reddit. Participation was completely voluntary, anonymous, and unincentivized. The majority of participants (58 %) self-identified as male (41% female; 1% non-binary) and were rather young ($M = 28.30$ years, $SD = 8.75$). Further sample demographics included education (61 % university degree) and nationality (58 % US citizens).

A list of 25 subreddit usage motives (see Table 1) was compiled based on previous Social TV studies (Gil de Zúñiga et al., 2015; Han & Lee, 2014; Krämer et al., 2015; Puschmann, 2017; Shim et al., 2017). Participants were first asked to rate the importance of each of these motives to their personal subreddit use on a Likert-type scale from 1 (strongly disagree) to 7 (strongly agree). Afterwards, participants were presented the same list again and asked to select which, if any, of the three thread types (pre-premiere, live premiere, post-premiere) was best suited to fulfilling each motive. The full survey questionnaire and an overview of all motive ratings are available on the study's OSF repository.

4. Results

4.1. Usage Motives of Discussion Threads

In order to identify distinct motive factors, we conducted an exploratory maximum likelihood factor analysis with oblique rotation on the 25 surveyed usage motives using the *R* package "psych" (Revelle, 2019). We retained five factors based on a preceding parallel analysis (see Table 1). The most important motive factor for r/gameofthrones subreddit users was 'Knowledge' ($M = 5.60$, $SD = 1.19$, 7-point agreement scale), which spoke to users' need to obtain background lore and expand their knowledge. This factor was followed by 'Humor' ($M = 5.09$, $SD = 1.79$), which represented the need for entertainment through memes and humorous comments. Virtual 'Company' ($M = 3.14$, $SD = 1.72$)—the feeling of a shared viewing experience and the social gratifications connected with it—played a rather minor role, as did the needs to actively contribute one's own content, theories, or comments ('Share,' $M = 2.34$, $SD = 1.42$) and to share/receive spoilers ('Spoilers,' $M = 2.02$, $SD = 1.35$).

However, thread type suitability ratings indicated that the temporally structured discussion threads did indeed appear to be differently able to fulfil usage motives (see Figure 2). Although the post-premiere thread seemed to be the 'all-rounder' and suited to fulfilling all five overarching motive factors, those motives related to the need for company could be best satisfied by parallel communication in the live-premiere thread. The pre-premiere thread, on the other hand, was mostly relevant to refreshing one's knowledge or receiving background information but seemed to be of limited relevance to the other motive factors. Crucially, for each motive, a substantial share of participants did not specify a most-adequate thread type (ranging from 14.9% for Motive 14 ["know whether I have missed something important"] to 46.3% for Motive 08 ["earn some Reddit karma/silver/gold"]). This may be attributable to partic-

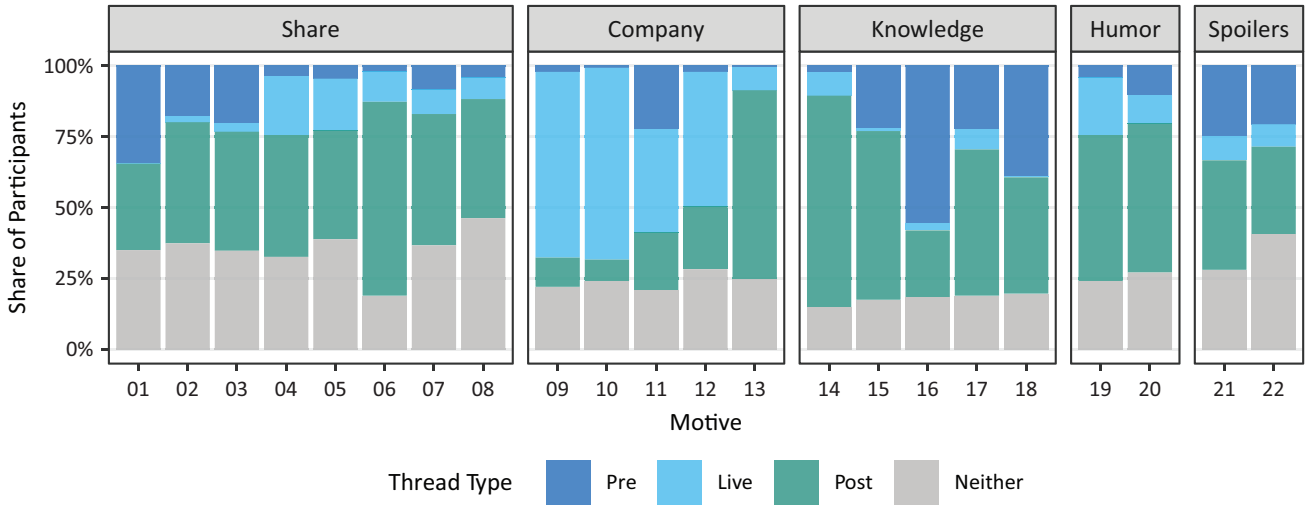


Figure 2. Thread most suitable to fulfilling usage motives. Notes: *n* = 417. For motive item wordings see Table 1.

ipants perceiving the respective motives as less important and/or to users' perception of multiple thread types as equally suitable to fulfilling these motives.

4.2. Usage Practices of Discussion Threads

To investigate whether the temporally structured discussion spaces were also characterized by differing usage practices, we turned to the content and meta-textual information of the comments left in those threads. In purely quantitative terms, the three thread types were used very differently (see Figure 3). While the popularity of all thread types increased over time, by far the most comments were found in the post-premiere threads, with all Season 8 episodes generating more than 40,000 comments (and peaking at around 100,000 comments for the controversial episode "The Long Night"). This was followed by synchronous live commenting on the episodes. The pre-premiere threads rarely exceeded 5,000 comments. Of course, the temporal structuring of the threads predefined when comments were made in relation to the episode's airing, with the pre-premiere thread starting 48 hours before the US premiere of an episode and the live and post-premiere threads appearing shortly before the start and the end of each episode. Thus, despite the possibility of asynchronous communication on the episodes, thread usage was strongly influenced by linear program planning. All three threads received the most attention in a window of only a few hours immediately before and after the US premiere of the respective episode (see Figure 4). Even in the post-premiere thread, the vast majority of comments were made within three to four hours after the episode aired, and almost no new comments appeared after about a day.

Looking closer at the content of the comments, we found that the use of the live premiere threads in particular differed from the asynchronous thread types. Not

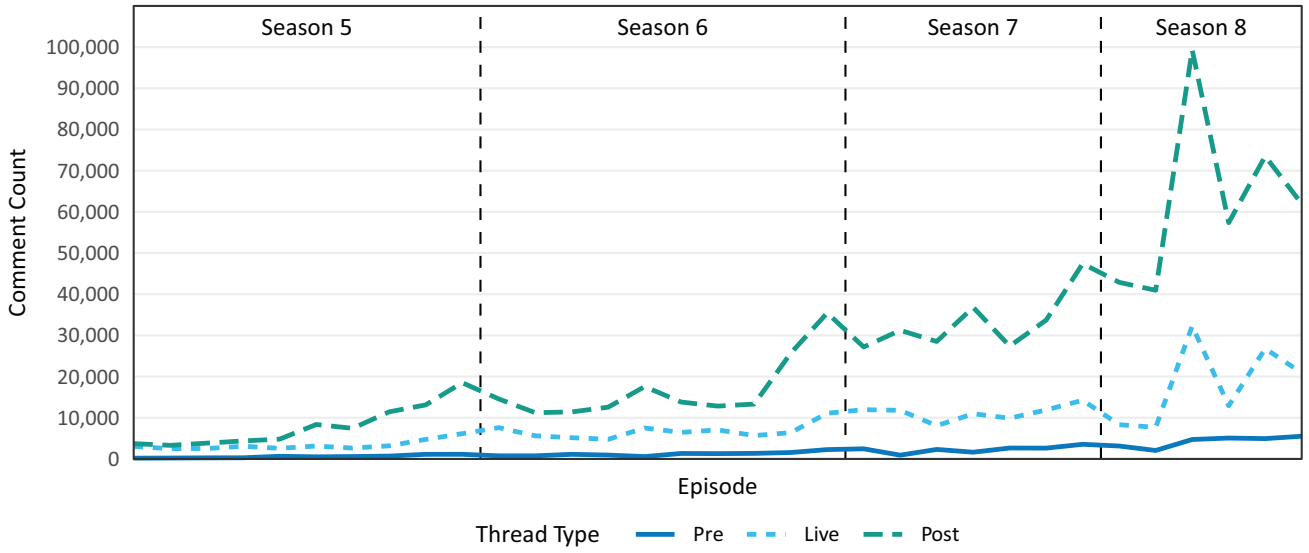


Figure 3. Comment counts per episode and thread type. Note: *n* = 1,203,666.

Table 1. Factor loadings of thread usage motives.

I use the discussion threads to…	I Share	II Company	III Knowledge	IV Humor	V Spoilers
01: share background lore	0.885	−0.155	0.103	−0.046	−0.069
02: share my own theories about developments in the series	0.853	−0.103	0.079	−0.101	−0.051
03: show my knowledge about the series	0.808	0.086	−0.011	−0.085	−0.034
04: post humorous comments	0.793	−0.051	0.004	0.144	−0.076
05: show my wit and humor	0.765	0.047	−0.050	0.062	0.044
06: discuss or argue about events in the episode	0.722	0.142	0.006	−0.131	−0.146
07: post or share memes	0.519	−0.073	−0.083	0.320	0.143
08: earn some Reddit karma/silver/gold	0.423	−0.028	−0.048	0.036	0.222
09: feel like watching together with other people	−0.126	0.957	−0.083	0.003	−0.033
10: not be alone while watching	−0.148	0.846	−0.153	−0.002	0.084
11: share excitement or suspense	0.271	0.626	0.036	−0.044	−0.112
12: interact with other people	0.395	0.586	−0.055	−0.115	0.031
13: deal or cope with tragic events in the series	0.011	0.572	0.082	0.025	−0.080
14: know whether I have missed something important	−0.020	0.092	0.510	0.052	0.028
15: read theories about developments in the series	0.011	−0.011	0.406	0.012	0.011
16: refresh my knowledge about past developments in the series	0.005	−0.026	0.742	−0.081	0.058
17: receive specific information about characters, locations, music, etc.	0.050	−0.069	0.736	0.025	0.064
18: read background lore	0.015	−0.080	0.831	−0.038	−0.035
19: read humorous comments	0.024	0.038	0.030	0.755	−0.120
20: read or view memes	−0.082	−0.005	−0.027	0.872	−0.030
21: read spoilers	−0.172	−0.044	0.145	−0.060	0.695
22: share spoilers	0.263	−0.094	−0.022	−0.048	0.552
23: bypass boring moments	0.000	0.157	−0.053	−0.011	0.250
24: get to know about other people's opinions	−0.029	0.297	0.193	0.091	0.003
25: get myself in the mood for watching	0.029	0.311	0.149	0.039	0.091
% Variance	0.19	0.12	0.09	0.06	0.04
M (*SD*)	2.34 (1.42)	3.14 (1.72)	5.60 (1.19)	5.09 (1.79)	2.02 (1.35)
ω_h	0.90	0.84	0.79	0.78	0.58

Notes: Maximum likelihood factor analysis with oblique rotation (Promax) for all 25 motive items. Scale: 1 (strongly disagree) to 7 (strongly agree). $n = 417$, Kaiser-Meyer-Olkin measure (KMO) = 0.87 (all KMO values for individual items ≥ 0.68), Bartlett's $K^2(24) = 808.34$, $p < 0.001$; parallel analysis suggested five factors. All factor loadings < |0.4| grayed out.

only were the comments in the live premiere threads on average much shorter ($M = 76$ characters, $SD = 118$) than in the pre- ($M = 142$, $SD = 231$) and post-premiere threads ($M = 126$, $SD = 189$), there was less interaction between users, as indicated by the fact that more than half (52%) of all comments in the live premiere threads were top-level comments—comments that replied to the original post and thus did not directly engage with comments from other users. In contrast, top-level comments only accounted for about one-fifth to a quarter of all comments in the pre- (24 %) and post-premiere threads (21 %).

To further investigate the content of the comments, we focused on keywords, that is, words that were distinctive in each thread type. Using the *R* package "quanteda" (Benoit et al., 2018), we first preprocessed all comments by removing URLs, symbols (including emojis), stopwords based on the Snowball stopword list, and further uninformative terms that were identified during preliminary data analysis (e.g., all-text emojis such as "xD" and remaining parts of HTML entities like '&' and '>'). We then calculated log-likelihood keyness scores for each thread type, using the other two thread types as reference groups. The identified keywords (see Figure 5) offered insights into the differences in discussion content. Comments in the pre-premiere thread seem to focus on predictions and discussions of (leaked) episode plot points and developments (e.g., "predict(ion/s)," "theory," "kills," "dies," "leaked," and "spoilers") and on the framing of the individual reception situation (e.g., "watch," "see," "premiere," and "hbo;" see also Puschmann, 2017). Live premiere threads were characterized by expletives and exclamations of surprise (e.g., "omg," "oh," and various four-letter words). Finally,

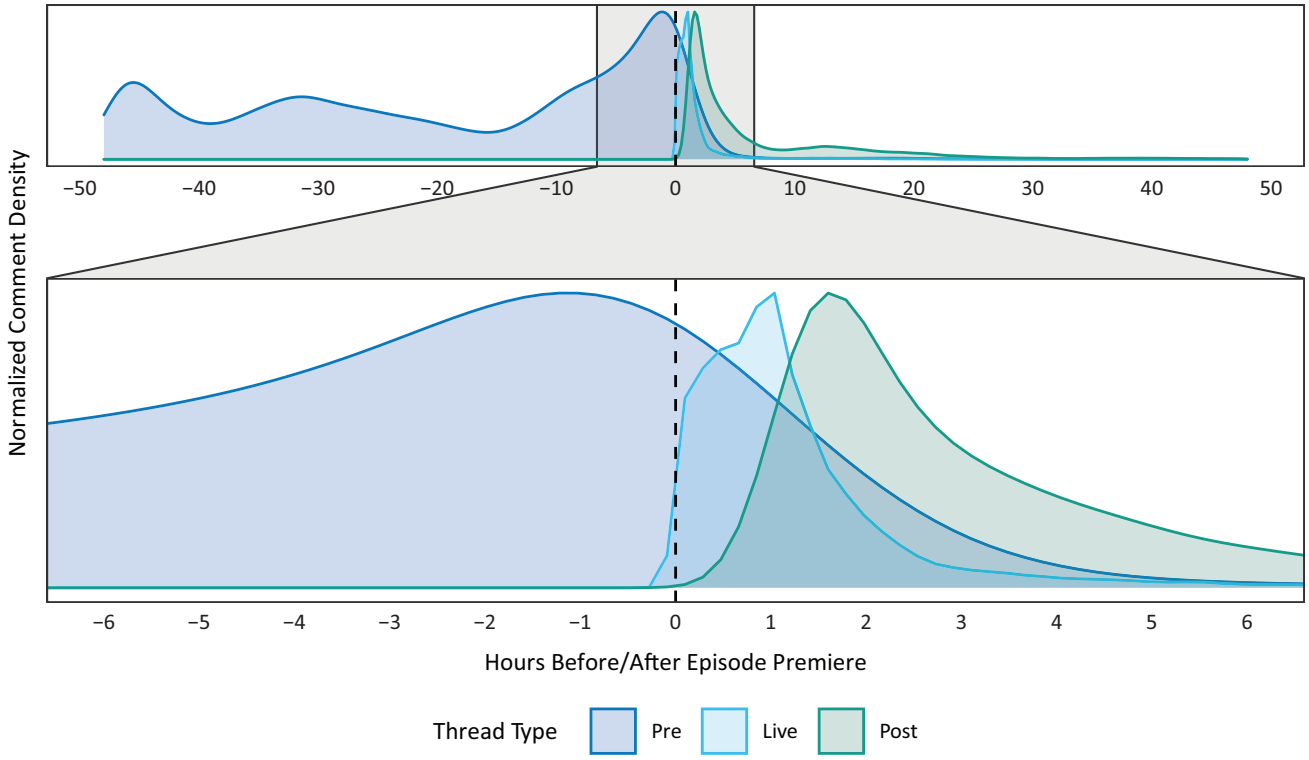


Figure 4. Comment density in relation to US episode premiere. Note: $n = 1{,}203{,}666$.

the post-premiere threads contained more in-depth discussions and evaluations of recent series events (e.g., "loved," "felt," "thought," and "scene"); crucially, other Reddit users and their opinions were explicitly addressed and consulted (e.g., "anyone" and "else").

5. Discussion and Conclusion

The present study attempted to explore which usage motives and practices are characteristic in different forms of asynchronous and synchronous communication about

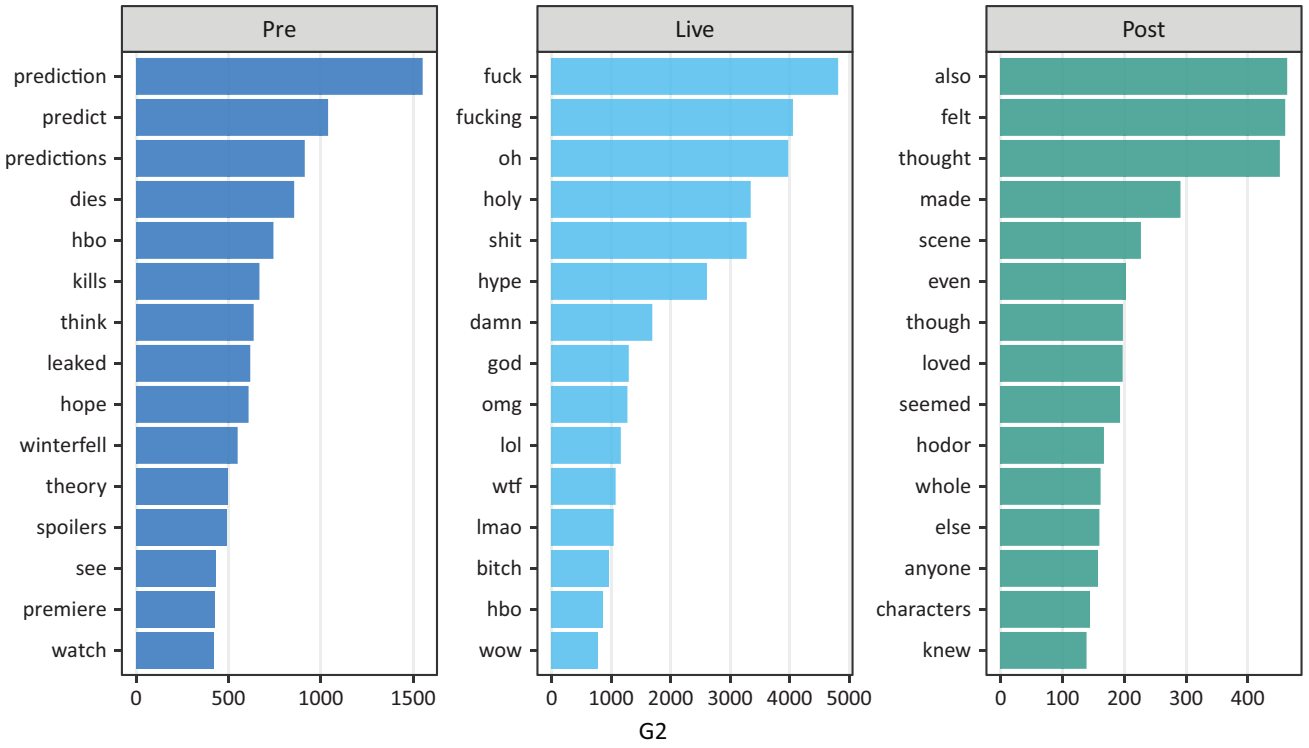


Figure 5. Keywords with highest likelihood-ratio G2 score per thread type.

TV series on social media. Building on prior research, we simultaneously examined the uses and motives of pre-communication, live communication, and follow-up communication on Reddit, a highly important yet understudied platform in the context of Social TV activities. Our multi-method research project combined an automated content analysis of more than 1.2 million user comments from the subreddit r/gameofthrones—an online community centered on discussions of the HBO series *GoT*—with a survey of users active in this community ($n = 417$).

While prior research already found that messaging apps and more public social media platforms are used differently and for different reasons in the context of Social TV (Han & Lee, 2014; Krämer et al., 2015), our project has shown that discussion spaces dedicated to asynchronous and synchronous communication about TV series are also frequented and valued for different reasons. Focusing on our case study of discussions about *GoT*, we found that refreshing and extending one's knowledge of the show (factor 'Knowledge') and viewing memes and humorous comments (factor 'Humor') were overall the most important for users of the r/gameofthrones subreddit. However, the temporally structured threads (pre-, live-, and post-premiere thread) were perceived to be more or less suited to fulfilling specific motives. According to our participants, the need for feeling like they were watching with others and not being alone (factor 'Company') was best satisfied by parallel communication in the live premiere thread, while almost all motives related to sharing one's own information or theories (factor 'Share') were perceived as best fulfilled in the post-premiere thread. Thus, not all Social TV activities are created equal, and users seem to engage in these activities with the goal of addressing specific needs and desires. This also became apparent when we considered the broader information environment on Reddit. While r/gameofthrones is by far the most frequented and broadest *GoT*-related subreddit, there are also specialized subreddits that are, for example, exclusively focused on humorous content and memes (e.g., r/aSongOfMemesAndRage).

To get a more detailed picture, we complemented the survey with a large-scale automated content analysis of actual usage practices (i.e., commenting behavior) in the three thread types. This analysis has shown that: (1) follow-up communication in the post-premiere threads was most popular among users of r/gameofthrones; (2) parallel communication in the live premiere threads was less extensive and less interactive than pre-communication and follow-up communication; and (3) the focus and content of discussions varied in the three threads as well. Looking at the most important terms per thread type, we found that the post-premiere threads seemed to be characterized by users sharing background information and wanting to discuss theories, as other users were addressed explicitly and specific developments were debated; this mirrored our survey results. The live premiere threads, on the other hand, were used to express one's surprise and emotions, while discussions in the pre-premiere threads were indicative of talking about one's ability to watch the current episode and for exchanging predictions.

Taken together, our multi-method approach allowed us to gain differentiated insights into (a)synchronous Social TV activities and has shown the potential of combining survey and content data. This potential is best illustrated by an example from our data: if we had relied solely on the survey data, we might have suspected that the importance of the live premiere thread to the motive factor 'Company' (i.e., feeling like watching together, sharing one's feelings with others) would have led users to engage in lively exchanges and more interactions with others. Instead, the thread analysis indicated that comments in the live premiere threads were the shortest, and the level of direct interactions was, by far, the lowest. This finding implies that the need for company might already be stimulated by the mere feeling of being in the same (online) space as other viewers—comparable to the collective cinema experience of sitting and watching in silence, united simply by joint attention to the screen (Hanich, 2018). Direct interactions with other users happened more often—and were actively sought out—in the asynchronous discussion threads. On the other hand, the survey responses provided us with deeper insights into the users' motives in addition to the discussion content, especially as the automated content analysis faced some challenges and limitations. Specifically, at least with data-driven, unsupervised text analysis methods, we were not able to extract discussion patterns in the actual Reddit comments that showed substantial overlap with the surveyed usage motives. In future studies, more theory-based analysis procedures such as dictionaries or adaptations of topic modeling procedures that allow for the prior specification of keywords (e.g., Eshima, Imai, & Sasaki, 2020) might provide viable alternatives in the more meaningful combination of user-centric survey results and content-centric automated text analysis.

Although the foci on a single platform, a single community, and a single TV series reduces the generalizability of our findings, this case study approach allowed us to gain insights into distinct Social TV activities while reducing the influence of confounding variables introduced by platform affordances (e.g., number of characters allowed in comments), discussion dynamics (e.g., users' mode of interacting on Instagram versus on Reddit), or audience characteristics. Future research should build on our findings by investigating pre-communication, parallel communication, and follow-up communication on different social media and with a focus on different TV series. For example, it might be expected that viewers of TV series that are less 'story-heavy' than *GoT* have a diminished need for background knowledge and theories, but are even more interested in humorous interactions or memes. However, due to the fact that the show's content "serves as the source material for content shared on the second screen" (Raney & Ji, 2017, p. 429), it will prob-

ably never be possible to completely separate Social TV activities from the TV series they relate to.

With the success of streaming services and viewing behaviors that are less and less tied to single episodes, the necessity to investigate asynchronous Social TV activities will certainly increase. For example, as the episodes of most shows of subscription-based video-on-demand services, such as Netflix and Amazon Prime Video, are released en bloc, parallel communication becomes less important because there is no actual 'live' moment anymore, not even for users in the same time zone. Popular discussions indicate that tweeting behavior seems to be particularly challenged by this all-at-once release strategy (see Weller, 2016): indeed, when, if at all, is it acceptable to post about a new Netflix show on Twitter—and what is one allowed to give away? Considering these debates in the context of our findings, we would particularly encourage a closer look at forms of follow-up communication, as the post-premiere threads consistently emerged as the most frequented ones. Moreover, it is likely that the engagement in Social TV discussions also alters the experience of first screen content (Krämer et al., 2015; Raney & Ji, 2017). As such, differences in Social TV usage motives and practices may lead to differential effects on users' entertainment experiences. From a methodological standpoint, it is interesting to consider that these entertainment experiences are likely to be reflected in the content of Social TV discussions, as indicated by expressions of suspense, surprise, or shock in the live episode discussion threads and in-depth discussions of affective responses in the post-premiere threads. Being able to identify the different entertainment experiences of TV series/episodes (e.g., hedonic and eudaimonic, see Oliver & Raney, 2011; Vorderer, 2011) by observing discussions on social media would benefit media entertainment research by reducing the need to rely on retrospective self-reports. As a first step in this direction, dictionary-based methods may provide a scalable procedure to investigate hedonic and eudaimonic entertainment experiences based on user comments.

Despite the inherent limitations of our project, this research presents exploratory first steps in understanding the different temporal, motivational, and behavioral facets of Social TV experiences. By recognizing that communication about TV series in social media goes beyond synchronous interactions, media entertainment research will be in a better position to explore the gratifications of reading, learning, and talking about one's favorite TV show. After all, TV has always been a social experience and will continue to be so in online environments.

Acknowledgments

The authors wish to thank the four anonymous reviewers and the thematic issue editors for their valuable comments.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the authors (unedited). The replication code and data are available at https://osf.io/7v49t/.

References

Alexa. (2020). The top 500 sites on the web. *Alexa*. Retrieved from https://www.alexa.com/topsites

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2019). New data sources in social science research: Things to know before working with Reddit data. *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/0894439319893305

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 830–839. Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/7347

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Bentley, F. R. (2017). Understanding secondary content practices for television viewing. In *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video* (pp. 123–128). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3077548.3077554

Buschow, C., Schneider, B., & Ueberheide, S. (2014). Tweeting television: Exploring communication activities on Twitter while watching TV. *Communications*, *39*(2), 129–149. https://doi.org/10.1515/commun-2014-0009

Cassis, C. (2018). Discuss and discover TV with Reddit. *Upvoted*. Retrieved from https://redditblog.com/2018/08/10/discuss-and-discover-tv-with-reddit

Eshima, S., Imai, K., & Sasaki, T. (2020). *Keyword assisted topic models*. Unpublished manuscript. Retrieved from http://arxiv.org/abs/2004.05964

Gil de Zúñiga, H., Garcia-Perdomo, V., & McGregor, S. C. (2015). What is second screening? Exploring motivations of second screen use and its effect on online political participation. *Journal of Communication*, *65*(5), 793–815. https://doi.org/10.1111/jcom.12174

Han, E., & Lee, S.-W. (2014). Motivations for the complementary use of text-based media during linear TV viewing: An exploratory study. *Computers in Human Behavior*, *32*, 235–243. https://doi.org/10.1016/j.chb.2013.12.015

Hanich, J. (2018). *Audience effect: On the collective cinema experience*. Edinburgh: Edinburgh University Press.

Jenner, M. (2017). Binge-watching: Video-on-demand, quality TV and mainstreaming fandom. *International Journal of Cultural Studies*, *20*(3), 304–320. https://doi.org/10.1177/1367877915606485

Ji, Q., & Raney, A. A. (2015). Morally judging entertainment: A case study of live tweeting during Downton Abbey. *Media Psychology*, *18*(2), 221–242. https://doi.org/10.1080/15213269.2014.956939

Johnson, B. K., & Rosenbaum, J. E. (2015). Spoiler alert: Consequences of narrative spoilers for dimensions of enjoyment, appreciation, and transportation. *Communication Research*, *42*(8), 1068–1088. https://doi.org/10.1177/0093650214564051

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *The Public Opinion Quarterly*, *37*(4), 509–523.

Krämer, N. C., Winter, S., Benninghoff, B., & Gallus, C. (2015). How "social" is Social TV? The influence of social motives and expected outcomes on the usage of Social TV applications. *Computers in Human Behavior*, *51*, 255–262. https://doi.org/10.1016/j.chb.2015.05.005

Leavitt, J. D., & Christenfeld, N. J. S. (2013). The fluency of spoilers: Why giving away endings improves stories. *Scientific Study of Literature*, *3*(1), 93–104. https://doi.org/10.1075/ssol.3.1.09lea

Lim, J. S., Hwang, Y., Kim, S., & Biocca, F. A. (2015). How social media engagement leads to sports channel loyalty: Mediating roles of social presence and channel commitment. *Computers in Human Behavior*, *46*, 158–167. https://doi.org/10.1016/j.chb.2015.01.013

Lochrie, M., & Coulton, P. (2012). Sharing the viewing experience through second screens. In *Proceedings of the 10th European Conference on Interactive TV and Video* (pp. 199-202). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2325616.2325655

Oliver, M. B., & Raney, A. A. (2011). Entertainment as pleasurable and meaningful: Identifying hedonic and eudaimonic motivations for entertainment consumption. *Journal of Communication*, *61*(5), 984–1004. https://doi.org/10.1111/j.1460-2466.2011.01585.x

Puschmann, C. (2017). Beitragstypen der öffentlichen rezeptionsbegleitenden Kommunikation auf Twitter bei fiktionalen TV-Inhalten [Types of tweets in the public, synchronous communication about fictional TV content on Twitter]. In U. Göttlich, L. Heinz, & M. R. Herbers (Eds.), *Ko-Orientierung in der Medienrezeption: Praktiken der Second-Screen-Nutzung* [Co-orientation in media use: Practices of second screen use] (pp. 195–218). Cham: Springer.

Quan-Haase, A., & Young, A. L. (2010). Uses and gratifications of social media: A comparison of Facebook and instant messaging. *Bulletin of Science, Technology & Society*, *30*(5), 350–361. https://doi.org/10.1177/0270467610380009

Raney, A. A., & Ji, Q. (2017). Entertaining each other? Modeling the socially shared television viewing experience. *Human Communication Research*, *43*(4), 424–435. https://doi.org/10.1111/hcre.12121

Revelle, W. psych: Procedures for psychological, psychometric, and personality research (Version 1.9.12) [Computer software]. (2019). Evanston, IL: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Rubin, A. M. (2002). The uses-and-gratifications perspective of media effects. In J. Bryant & D. Zillmann (Eds.), *Media effects: Advances in theory and research* (2nd ed.) (pp. 525–548). Mahwah, NJ: Lawrence Erlbaum Associates.

Schirra, S., Sun, H., & Bentley, F. (2014). Together alone: Motivations for live-tweeting a television series. In *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2441–2450). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2556288.2557070

Schlütz, D. (2016). Contemporary quality TV: The entertainment experience of complex serial narratives. *Annals of the International Communication Association*, *40*(1), 95–124. https://doi.org/10.1080/23808985.2015.11735257

Shim, H., Shin, E., & Lim, S. (2017). What makes us two-screen users? The effects of two-screen viewing motivation and psychological traits on social interactions. *Computers in Human Behavior*, *75*, 339–346. https://doi.org/10.1016/j.chb.2017.05.019

SimilarWeb. (2020). Reddit.com. *SimilarWeb*. Retrieved from http://similarweb.com/website/reddit.com

Sutter, T. (2017). Kommunikation über Fernsehen im Internet. Social TV als Anschlusskommunikation [Communication about TV on the Internet. Social TV as follow-up communication]. In U. Göttlich, L. Heinz, & M. R. Herbers (Eds.), *Ko-Orientierung in der Medienrezeption: Praktiken der Second-Screen-Nutzung* [Co-orientation in media use: Practices of second screen use] (pp. 29–46). Cham: Springer.

Vorderer, P. (2011). What's next? Remarks on the current vitalization of entertainment theory. *Journal of Media Psychology*, *23*(1), 60–63. https://doi.org/10.1027/1864-1105/a000034

Weller, C. (2016, July 5). The complicated ethics of tweeting about your favorite Netflix show. *Business Insider*. Retrieved from https://www.businessinsider.com/ethics-of-tweeting-about-netflix-show-2016-7

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Widman, J. (2020, July 1). What is Reddit? *Digital Trends*. Retrieved from https://www.digitaltrends.com/web/what-is-reddit

About the Authors

Julian Unkel (Dr. rer. soc., LMU Munich) is a Postdoctoral Researcher at the Department of Media and Communication at LMU Munich. His research interests focus on online media selection and effects, media entertainment, and communication research methods. As a member of the LMU Open Science Center, he seeks to promote and to foster open science practices in communication research. More information: http://julianunkel.com

Anna Sophie Kümpel (Dr. rer. soc., LMU Munich) is a Postdoctoral Researcher at the Department of Media and Communication at LMU Munich. Her research interests are focused on media effects, particularly in the context of social media, (incidental exposure to) online news, and media entertainment. Her research has been published in *Journal of Communication*, *Journal of Media Psychology*, and *Social Media + Society*, among others. More information: http://anna-kuempel.de

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 191–204
DOI: 10.17645/mac.v8i3.3153

Article

Popular Music as Entertainment Communication: How Perceived Semantic Expression Explains Liking of Previously Unknown Music

Steffen Lepa [1,*], Jochen Steffens [2], Martin Herzog [1] and Hauke Egermann [3]

[1] Audio Communication Group, TU Berlin, 10587 Berlin, Germany; E-Mail: steffen.lepa@tu-berlin.de (S.L.), herzog@tu-berlin.de (M.H.)
[2] Media Department, University of Applied Sciences Düsseldorf, 40476 Düsseldorf, Germany; E-Mail: jochen.steffens@hs-duesseldorf.de
[3] York Music Psychology Group, University of York, York, YO10 5DD, UK; E-Mail: hauke.egermann@york.ac.uk

* Corresponding author

Submitted: 14 April 2020 | Accepted: 26 June 2020 | Published: 13 August 2020

Abstract
Our contribution addresses popular music as essential part of media entertainment offerings. Prior works explained liking for specific music titles in 'push scenarios' (radio programs, music recommendation, curated playlists) by either drawing on personal genre preferences, or on findings about 'cognitive side effects' leading to a preference drift towards familiar and society-wide popular tracks. However, both approaches do not satisfactorily explain why previously unknown music is liked. To address this, we hypothesise that unknown music is liked the more it is perceived as emotionally and semantically expressive, a notion based on concepts from media entertainment research and popular music studies. By a secondary analysis of existing data from an EU-funded R&D project, we demonstrate that this approach is more successful in predicting 10000 listeners' liking ratings regarding 549 tracks from different genres than all hitherto theories combined. We further show that major expression dimensions are perceived relatively homogeneous across different sociodemographic groups and countries. Finally, we exhibit that music is such a stable, non-verbal sign-carrier that a machine learning model drawing on automatic audio signal analysis is successfully able to predict significant proportions of variance in musical meaning decoding.

Keywords
entertainment; genre preferences; musical expression; music preferences; musical taste; popular music; push scenarios; semantics

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

Popular music (in the broadest sense, also encompassing 'oldies,' jazz and hits from the classical repertoire, based on the definition of Tagg, 2000), is one of the most prevalent types of entertainment content in everyday media use, especially in social media. It is nowadays predominantly consumed in *push scenarios*—socio-musical contexts, in which music is selected and played back for us by someone else (e.g., when listening to radio programs, curated playlists, DJ sets, in-store music, YouTube videos, shuffle-mode, and music in virtual worlds) or by recommendation algorithms. The current abundance of listening situations where people are confronted with previously unknown popular music is part of the ongoing "musicalization" of society (Pontara & Volgsten, 2017).

In the past age of audio storage media, the breadth of existing music tracks available for airplay was generally limited by material physical and economic restrictions, similar to the number and stylistic variety of albums and singles published and sold. Consequently, the question of which records to buy and which musical genre and artists to adhere to has led to a great deal of socio-cultural distinction practices (Bourdieu, 1984). Out of limited and heterogenous economic and cultural capital, people typically stuck to their cultural habitus acquired during their formative years which then formed an essential part of their identity during their later lives (Frith, 1996). Therefore, personal taste and choices in music selection have always tended to separate people of different generations, milieus, and cultures from one another, which remains partly the case today (Mellander, Florida, Rentfrow, & Potter, 2018; Vlegels & Lievens, 2017). However, we claim that due to the inflation of musico-technological repertoires in the age of digital media (Lepa & Hoklas, 2015) as well as the global introduction of 'flatrate' streaming offerings (Drott, 2018) and music recommendation algorithms (Krämer, 2018), there is a growing tendency that the logics governing people's socio-musical practices are converging to new patterns (see a review in Section 1.1). In consequence, theoretical models from music psychology and cultural sociology that successfully described musical preference dynamics in the past require re-examination (Brisson & Bianchi, 2019; Prior, 2013).

Accordingly, in the present article, we propose and empirically compare alternative explanations for music liking in existing popular music push scenarios by drawing on concepts from empirical aesthetics, media entertainment research, and popular music studies.

1.1. Challenging the Perceived View on Music Liking: Personal Genre Preferences

Research on music liking has often employed questionnaires asking participants for the degree of liking regarding several given musical genre labels. This practice has revealed sociodemographic differences in obtained preference patterns (typically, a high-brow vs low-brow cultural gap; Roose & Stichele, 2010), as well as correlations with personality traits (Fricke & Herzberg, 2017) and political attitudes (Feezell, 2017). However, observed effect sizes are comparatively small (Schäfer & Mehlhorn, 2017), and it remains unclear whether obtained answer patterns relate to music listening practices in the push scenarios under discussion here. This is due to the ambiguous intensional content of musical genres. In catalogues of music stores or streaming providers, different taxonomies exist (e.g., Spotify vs Apple Music), and listeners tend to have heterogeneous and historically changing ideas about the musical attributes and values defining specific genres (Lahire, 2008). Moreover, non-musicians in particular associate genres rather with social stereotypes and identity concepts related to artists, epochs, subcultures, and fandom the music stems from (Shevy, 2008). As a result, genre-based expressed musical taste has become an elemental part of postmodern identity and distinction practices (Lonsdale & North, 2009).

However, while being an interesting phenomenon in itself, taste performances (Hennion, 2001) are not necessarily informative for the actual patterns of music liking and listening practice (Lonsdale & North, 2012). Also, recent empirical studies suggest a growing tendency towards genre "omnivorousness" (Peterson, 1992) spreading across social classes (Vlegels & Lievens, 2017) and a continuous development towards new genre taxonomy logics based on social context or lifeworld functions (Airoldi, Beraldo, & Gandini, 2016). It is therefore unsurprising that the actual power of traditional genre labels for predicting musical liking is rather low (Brisson & Bianchi, 2019). Hence, we infer that explicitly stated genre affinities might only explain a minor portion of music liking in music push scenarios.

1.2. Familiarity, Prominence and Popularity as 'Cognitive Side Effects' in Music Liking

In contrast, theories from empirical aesthetics and social psychology appear better suited to explaining music liking in times of musicalization. For instance, repeated exposure to a stimulus leads to cognitive fluency effects (Jacoby & Dallas, 1981) and a more positive evaluation. However, familiarity and pleasantness of artworks only covary up to a certain ideal point, from where pleasantness decreases again in terms of a saturation effect, often idealised graphically by an inverted U-curve (Chmiel & Schubert, 2017). In the original theory of Berlyne (1971), this effect was said to interact with stimulus complexity. However, this has rarely been successfully demonstrated in experimental works with musical stimuli (Madison & Schiölde, 2017). An alternative explanation of the advantage of widely-known music in push scenarios is the elaboration likelihood model (Petty & Cacioppo, 1986). According to the model, the fact that a piece of music is well-known and appreciated by others (e.g., in our culture or peer-group) constitutes a peripheral persuasive cue that can positively influence aesthetical experiences, in particular in the low-involvement scenarios that we discuss here (Egermann, Grewe, Kopiez, & Altenmüller, 2009). Overall, we assume personal *familiarity* and social *popularity* effects in combination with incidental *saturation* resulting from over-prominence can explain experienced liking in situations where we are confronted with familiar-sounding music. We will denote them throughout this paper as "cognitive side effects" because they affect music liking independently of actual musical content or perceived expression.

1.3. Musical Expression Strength and Breadth as New Explanation for Music Liking

Rentfrow, Goldberg, and Levitin (2011) and Rentfrow et al. (2012), criticizing the genre label-based approach,

suggested working with sounding questionnaires to operationalize musical preferences. Following Hevner (1936), they also introduced adjective inventories allowing listeners to describe perceived 'attributes of music.' Based on aesthetical judgements gathered in this way, Greenberg et al. (2016) identified three major dimensions of perceived musical expression, two of them representing affect ("Arousal" & "Valence") and one representing the felt degree of aesthetic-cognitive stimulation ("Depth"). While this has generated significant progress for the field of music liking research, choosing a rather small convenience sample for 'judging' the semantics of popular music may lead to a narrowing of possible meanings as to what is deemed valuable from a high-brow perspective. This might lead to the perspective of the 'people,' the actual producers and addressees of popular music (Frith, 1996), becoming neglected.

Furthermore, analogous to discussions in media entertainment research (Klimmt, 2011), it appears crucial to acknowledge that beyond affect-guided hedonism and intellectual appreciation, people might also enjoy specific types of music because they fulfil their eudaimonic needs and help them to find identity, truth, and transformational experiences, overall rendering their everyday existence meaningful (Vorderer & Reinecke, 2015). To explain how meaning is imparted, Tagg (2013) argues that a majority of meanings conveyed by popular music, including substantial parts of affect expression, are due to so-called "para-musical fields of connotation." This term describes extra-musical meanings that are bestowed upon musical sign-carriers by human appropriation practices during the music's semiotic carrier as part of the circulation of culture (Herzog, Lepa, Egermann, Schönrock, & Steffens, 2020).

Based on the theoretical perspective of non-verbal communication theory adopted for music (Brunswik, 1952; Juslin, 2000) we further assume that perceived musical expression is only partly idiosyncratic, and rather by and large pragmatically 'understood' homogenously by other recipients, because most of our conspecifics take part in the same cultural game of musical semiosis as we do. In empirical studies on music expression drawing on Brunswik's (1952) lens model (Eerola, Friberg, & Bresin, 2013; Juslin, 2000), it has been found that the communicative cues employed in music work in a linear-additive fashion, sometimes redundantly, but often imparting several dimensions of meaning at the same time, which renders some pieces so expressive and popular. Hence, we postulate that liking for previously unknown popular music is dependent on the *breadth and strength* of perceived musical expression, which, according to cultural theorist Alison Stone (2016) should encompass the dimensions of affect, values, aesthetic commitments, identity, location and time. We furthermore suspect that these decoded connotations do exhibit a certain degree of idiosyncratic and cultural heterogeneity (Kristen & Shevy, 2013) in terms of content and their specific weight in personal preference judgments.

1.4. The Constructivist Challenge: To Which Degree Are Perceived Musical Expressions Socially Uniform and Predictable?

The current lack of prior systematic research on semantic musical expression might stem from the 'constructivist challenge' imposed by the concept of para-musical fields of connotations. If musical meaning lies to a greater extent in the 'ear of the beholder' and is not immediately inherent to the acoustic stimulus, how can we systematically measure it? On the other hand, it is known that film scores and advertisement music work well in communicating certain connotations successfully to recipients (Bouvier & Machin, 2013). Furthermore, it should be considered that musicalization has probably already led to a perceptible degree of musical sign-disambiguation across the globe and emotional music expressions might be based to some degree on cross-cultural universals (Sievers, Polansky, Casey, & Wheatley, 2012). Finally, similar to story or movie interpretations, the empirically found extent of non-uniformity in meaning decoding might also be due to the specificity of meanings searched for (Lepa, 2010). Following Tagg (2013), the actual degree of non-uniformity could be analysed either by measuring the variance of a small audience's actual meaning productions regarding a smaller pool of music or by formalising human meaning attribution regarding a larger pool of music with machine learning (ML) methods and then checking the resulting explanatory model power when applied to new music. Both approaches are pursued in the current contribution.

1.5. Resulting Hypotheses and Research Questions

Due to participation in a multi-national European research and development project on music branding funded by the EU (www.abc_dj.eu), we had the opportunity to test some of the assumptions mentioned above with an existing dataset. Even if the actual expression potential of popular music most probably reaches beyond the commercially exploitable domain, this nevertheless provided an excellent opportunity to test our following theoretical hypotheses based on the possibilities of this specific dataset:

H1: Liking of presented music is (positively) dependent on personal *genre affinity* strength.

H2: Liking of presented music is (positively) dependent on personal *familiarity* with a track.

H3: Liking of presented music is dependent on society-wide *popularity* and dependent on society-wide *prominence* of a track.

H4: Liking of presented music is dependent on *strength* and *breadth* of perceived musical expression regarding affect and values.

Additionally, four overarching open-ended research questions were explicated that address the constructivist challenge of music semantics:

RQ1: What is the relative importance of hypothesized predictors (and their sub-dimensions) regarding liking of presented music?

RQ2: Are there socio-cultural differences in perceived musical expression or relative preferences for different dimensions of musical expression?

RQ3: To what extent is it possible to predict perceived musical expression based on algorithmic audio signal analysis?

RQ4: Which acoustical attributes of popular music are best suited to predict perceived musical expression dimensions?

Two empirical studies were conducted. Based on a secondary analysis of existing data, Study 1 addresses the four main hypotheses, as well as RQ1 and RQ2. In order to answer RQ3 and RQ4, Study 2 then employs numerical results from Study 1 and combines them with ML and music information retrieval (MIR) techniques.

2. Study 1: Explaining Music Liking in Push Scenarios

To perform systematic inquiry on hypotheses H1–H4, as well as RQ1 and RQ2, we drew on available data from a cross-national online survey experiment which was part of an EU Horizon 2020 research & development project (Herzog, Lepa, & Egermann, 2016; Herzog, Lepa, Steffens, Schönrock, & Egermann, 2017a).

2.1. Methods

Due to space limitations, details on participants, sampling, stimulus material and data pre-processing are documented in the Supplementary File (A1.1.–A1.3). The resulting net sample comprises n = 9,197 subjects from three generations (gen Y: age 18–34, gen X: age 35–51, gen B: age 52–68) and three countries, with gender being approximately equal-distributed.

2.1.1. Procedure

Based on initial sociodemographic screening procedures organised by panel providers, subjects received an online questionnaire formulated in their country's primary language (English, German, Spanish). They conducted a short sound test and were then presented with either four (wave 1) or six (wave 2) randomly selected 30s popular music excerpts from a larger pool (see Section 2.4 for details). Afterwards, they rated the subjectively perceived fit between the music and 15 adjective attributes (GMBI_15 inventory, see Figure 3) employing a 6-point scale, as well as the degree of familiarity with and liking for the excerpt. Finally, subjects stated the extent of their general personal affinity to each of 10 different musical genres in the pool, which were presented to them as linguistic labels (see Figure 1).

2.2. Results

2.2.1. Personal Genre Affinities versus Actual Liking

Figure 1 provides an overview of obtained genre preference ratings. Highest personal affinities were found for *Pop & Charts*, and *Rock & Punk*, while *Hip Hop & Trap*, *Country & Folk* and *World Music* received the lowest sympathies. Also, we computed a multivariate generalized linear model (cumulative logit link, n = 9,197) to test for socio-cultural differences in genre preference patterns revealing various highly significant differences in line with the literature (not documented here), altogether explaining 15% R^2 (Nagelkerke).

We then calculated ordinal Kendall-Tau correlations between stated affinity for a genre and the actual liking of excerpts from that genre in the prior listening experiment, resulting in an average correlation of $\tau = 0.22$. Hence, stated affinities appear to be rather weak predictors for actual liking. Furthermore, as depicted by Figure 6 in the Supplementary File, we observed substantial differences in correlation size across genres and country of residence, hinting at cultural heterogeneities in genre label understanding.

2.2.2. Track Familiarity, Prominence, and Popularity

To test whether there was a sufficient amount of 'novel' music in the pool presented to participants, we inspected histograms of excerpt familiarity ratings by musical genre. Results demonstrated an expected long-tail distribution with the far more frequent *House & Techno* and *World Music* excerpts being rather unfamiliar to respondents, while *Rock & Punk* and *Pop & Charts* were most familiar to them (see Figure 7 in the Supplementary File).

We then calculated a society-wide prominence score for each excerpt, based on the mean track familiarity rating per country. Similarly, we computed a society-wide popularity score for each excerpt, based on the mean track liking rating per country. Afterwards, we estimated the ordinal Kendall-Tau correlation between both measures, resulting in $\tau = 0.34$. To check for a possible non-linear dependency, we plotted both aggregated index variables against each other (Figure 2), obtaining a clear 'hinge' effect with weaker dependencies for prominence values above a scale value of 2, but no substantial relationship between both indices and specific musical genres.

2.2.3. Perceived Musical Expression

To measure perceived musical expression, our survey utilised a multi-lingual questionnaire inventory

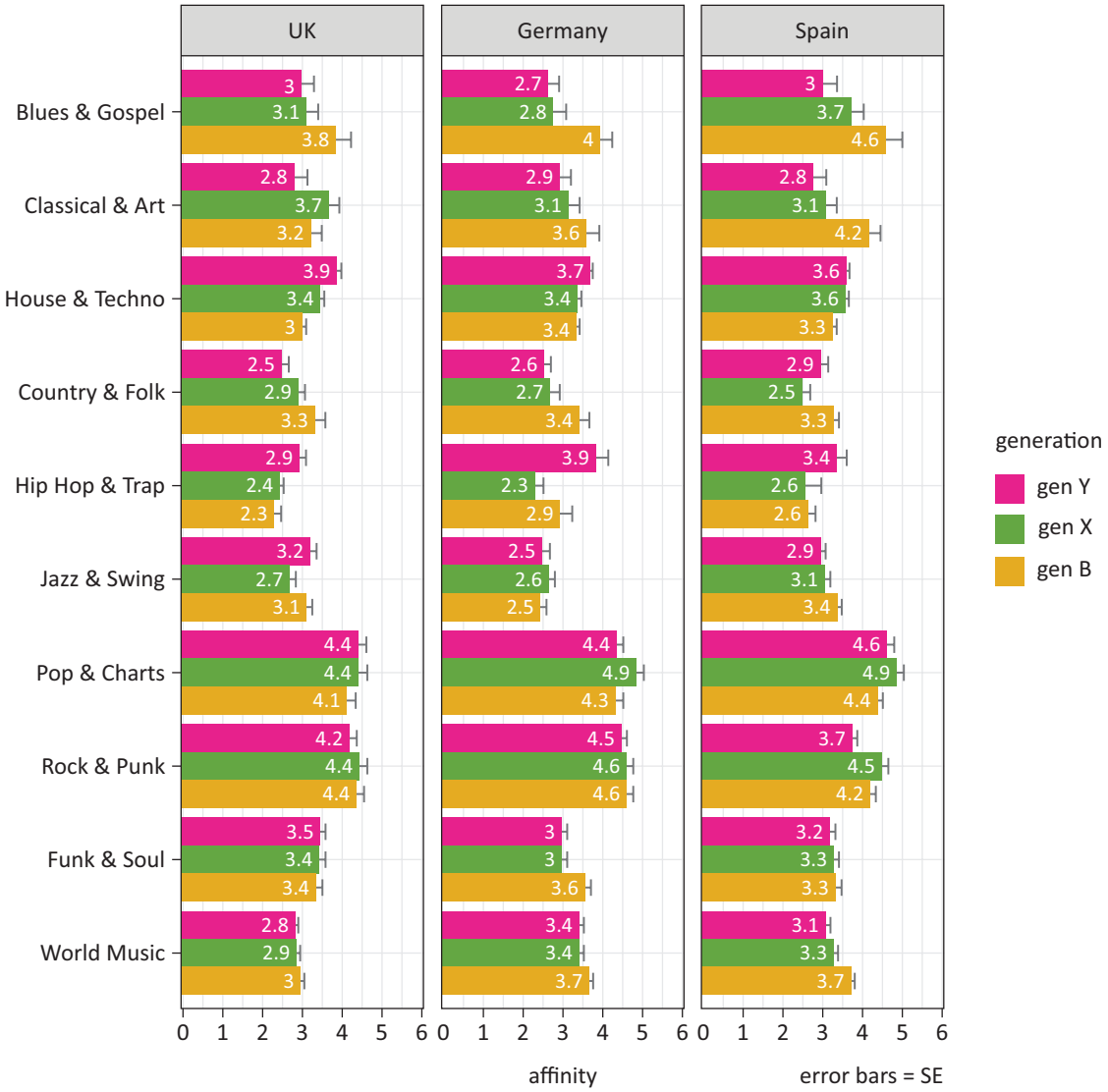


Figure 1. Mean genre affinity by country and generation (scaled from 1 to 6).

(GMBI_15) that had been developed based on results of an expert focus group and a marketing expert survey (Herzog et al., 2020; Herzog, Lepa, Steffens, Schönrock, & Egermann, 2017b). The GMBI_15 measures five musical expression dimensions relevant for branding campaigns, each operationalized by three manifest item indicators. Two dimensions represent musical affect expression (*Arousal, Valence*), while three others represent musical value expression (*Authenticity, Timeliness, Eroticity*). For interpretation of resulting factor scores, it is worth noting that, while items are formulated unipolar, the dimensions of the latent variables are interpreted bipolar (*Arousal*: relaxing–stimulating, *Valence*: dark–bright, *Authenticity*: conventional–authentic, *Timeliness*: traditional–futuristic, *Eroticity*: mental–sensual).

The empirical fit of the employed GMBI_15 measurement model (Herzog et al., 2017a) was estimated using MLR estimation and a sandwich-estimator to compensate for unbalanced measurement repetition within subjects (see Figure 3). This procedure resulted in a very good fit with X^2 = 515.239; df = 80; $p < 0.01$; RMSEA = 0.040 [0.039—0.041], CFI = 0.968; SRMR = 0.028 (note that significant p-values for model fit are expected for this sample size). Since items had been presented in three different languages, we tested measurement invariance across language versions following Cheung and Rensvold (2002), resulting in a fair degree of scalar invariance (see Table 8 in the Supplementary File). After inverting the polarity of *Arousal* to improve interpretability, we finally calculated z-standardised factor scores for each observation in the dataset. For each musical expression factor, we then performed an ANOVA-based variance component estimation (Searle, 1995), resulting in ~1% R^2 for socio-demographics, while track identity explained between 12–26% R^2 (see Table 9 in the Supplementary File).

2.2.4. Results of Hypotheses Tests

Hypotheses regarding music liking were tested by a blockwise ordinal logistic regression model (cumulative logit link, n = 9,197), calculating cluster-robust standard er-

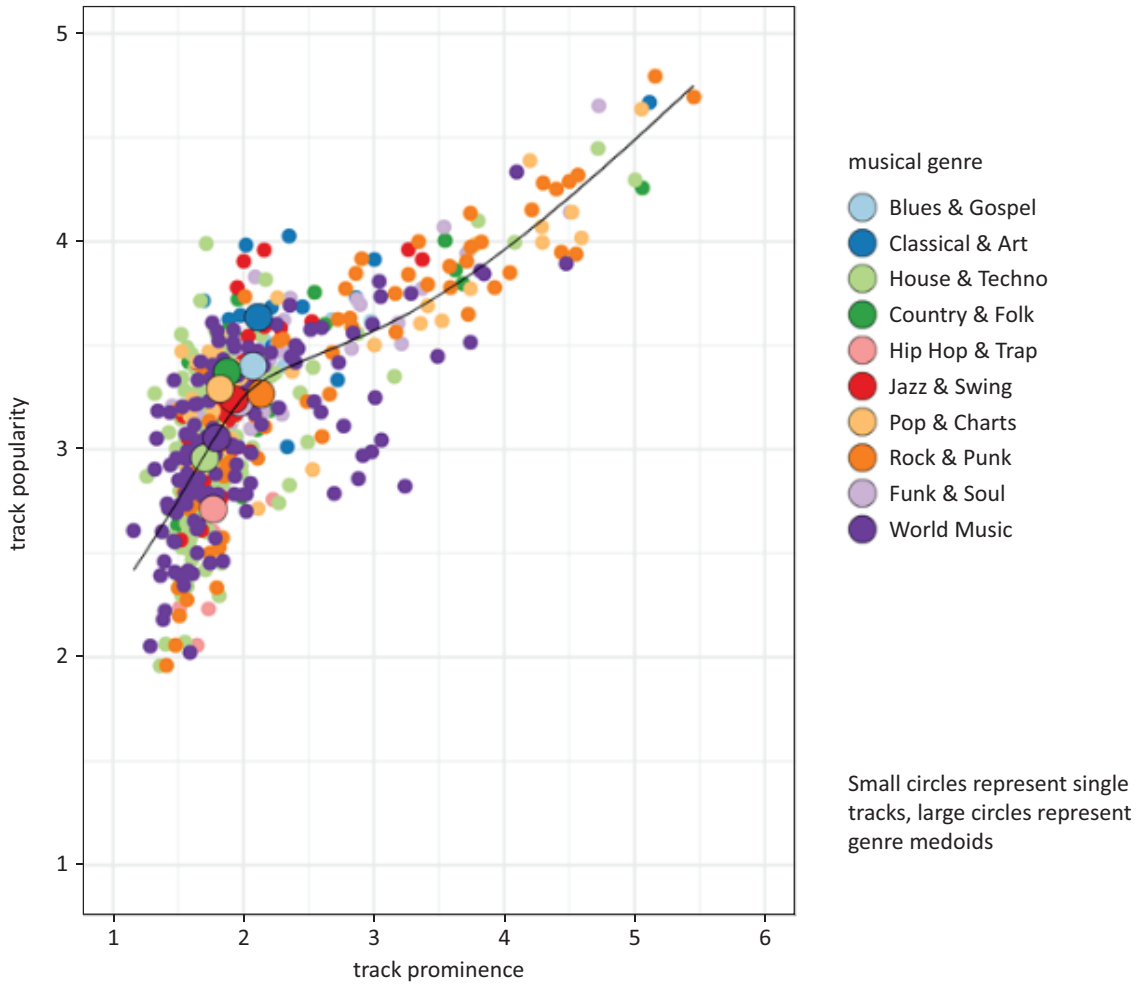


Figure 2. Track prominence vs track popularity by genre.

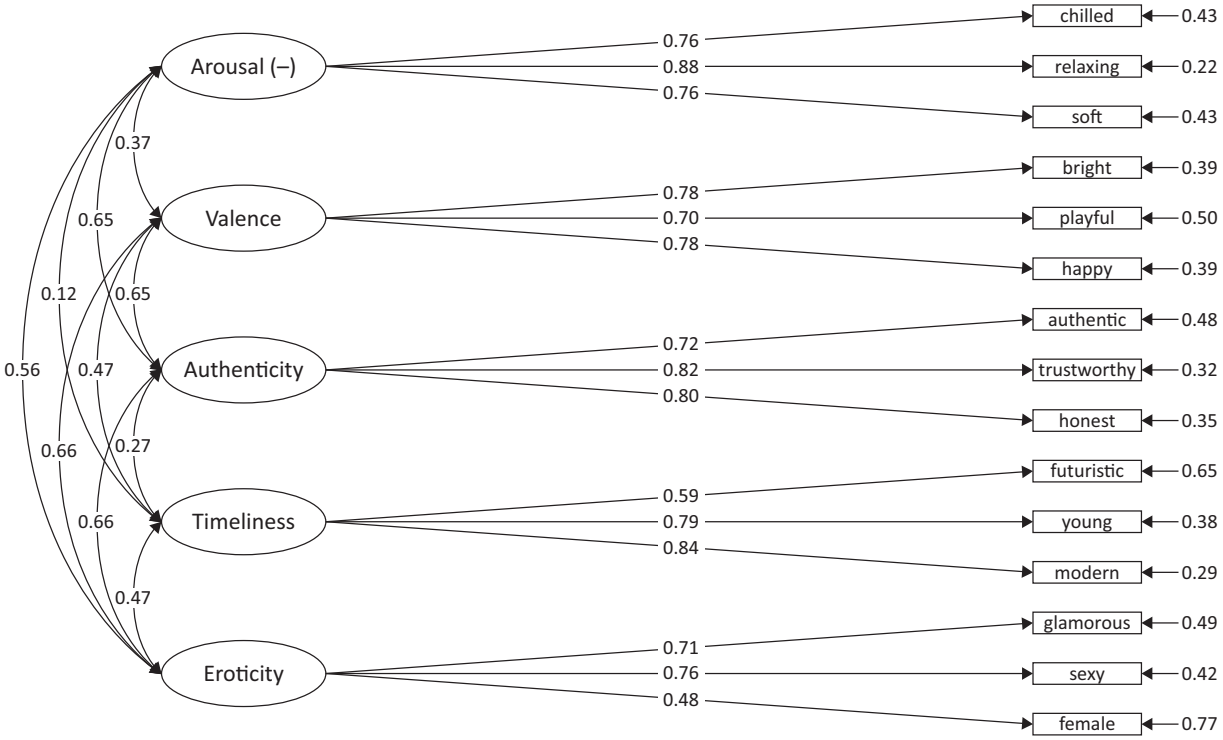


Figure 3. GMBI_15 measurement model for perceived musical expression (affect/values).

rors to compensate for unbalanced measurement repetitions within subjects. Hypothesis 1 regarding *genre affinity* as well as Hypothesis 3 assuming an influence of *popularity* and *prominence* were deliberately tested as last theoretical model blocks. This was done to allow for estimation of their 'true' R^2 contribution after having controlled for predictors sharing common variance. The incremental gain in Nagelkerke's R^2 was calculated for each block corresponding to one of the four major hypotheses (model 1–4), as well as for two additional models (model 5–6), which statistically corrected for the imbalance in track genres and socio-demographics. All hypotheses were confirmed as highly significant, and estimated beta-values for each predictor were only slightly altered when entering controls (Table 1). In an extended model version (not documented here due to space restrictions), we also tested all two-way-interactions between the six expression factor variables and socio-demographics which resulted in some significant, but minor effects contributing to an overall additional R^2 gain of only 1%.

2.3. Discussion

Results obtained from Study 1 confirmed our assumption concerning the obsolescence of genre labels for explaining musical liking. In contrast, assumed 'cognitive side effects' related to *familiarity, popularity*, and *prominence* play an essential role (H2 + H3). Notably, on a societal level, we identified a dampening effect of too much *prominence* on *popularity*. Furthermore, as postulated, it is predominantly music's perceived expression of affect and values (with nearly similar weights) that explain best why individual people enjoy previously unknown music played back to them (H4). When controlling for these effect clusters, preferences expressed by genre labels only explain a small residual portion of music liking (H1), feasibly representing associated non-musical stereotypes connected with genre labels. Finally, we observed only minor heterogeneities in perceived musical expression across socio-demographics and cultures (RQ1) and, similarly, we only found minor socio-cultural differences in weights for different musical expression dimensions predicting musical liking (RQ2).

Table 1. Results of ordinal block-wise regression (cumulative logit link) performed to explain liking of musical excerpts.

Predictor/Model	(1)	(2)	(3)	(4)	(5)	(6)
track familiarity (H2)	0.85***	0.57***	0.54***	0.55***	0.54***	0.54***
expression: arousal (H4)		−0.33***	−0.34***	−0.30***	−0.32***	−0.33***
expression: valence (H4)		0.44***	0.44***	0.47***	0.48***	0.49***
expression: authenticity (H4)		0.70***	0.66***	0.65***	0.62***	0.62***
expression: timeliness (H4)		0.17***	0.13***	0.19***	0.19***	0.18***
expression: eroticity (H4)		0.15***	0.16***	0.08***	0.09***	0.08***
genre affinity (H1)			0.30***	0.31***	0.33***	0.32***
track popularity (H3)				0.51***	0.51***	0.49***
track prominence (H3)				−0.28***	−0.28***	−0.26***
track genre: Blues and Gospel					0.24***	0.25***
track genre: Classical and Art					0.17***	0.17***
track genre: House and Techno					−0.05*	−0.05*
track genre: Country and Folk					0.01	−0.01
track genre: Hip Hop and Trap					0.21**	0.20**
track genre: Jazz and Swing					−0.05	−0.05
track genre: Pop and Charts					−0.40***	−0.40***
track genre: Rock and Punk					0.05	0.05
track genre: Funk and Soul					−0.07*	−0.06*
residency: Germany (def: UK)						0.03
residency: Spain (def: UK)						0.12***
age group: generation X (def: gen. Y)						−0.03
age group: generation B (def: gen. Y)						0.08*
education: ISCED 3–4 (def: ISCED 0–2)						0.02
education: ISCED 5–8 (def: ISCED 0–2)						0.11***
gender: female (def: male)						0.07**
Nagelkerke's R^2	18.57%	49.42%	50.71%	52.68%	52.92%	53.02%
Incremental R^2	18.57%	30.86%	1.29%	1.96%	0.24%	0.1%

Notes: all non-dummy predictors are standardized (beta-coefficients); standard errors are cluster-robust; track genres are effect-coded (redundant category: World Music); * $p < .05$, ** $p < .01$, *** $p < .001$.

3. Study 2: Predicting Perceived Musical Expression by Algorithmic Audio Signal Analysis

In the course of Study 2 addressing RQ3 and RQ4, we developed computational prediction models explaining the non-individual parts of variance contained in the scores of perceived musical expression factors Arousal, Valence, Authenticity, Timeliness, and Eroticity (see Study 1).

3.1. Methods

To this end, we utilised technical audio signal and music descriptors as predictors, which either stem from ML of branding experts' knowledge or algorithmic MIR toolboxes describing music and sound parameters. Details on development and selection of these predictor variables are provided in the Supplementary File (A2.1 and A2.2). Dependent variables were created by calculating the arithmetic mean of each perceived musical expression factor across all respondents of Study 1.

3.1.1. Statistical Aggregation of Descriptors and Feature Selection for Computational Prediction Models

Linear hierarchical stepwise regression procedures were employed to aggregate the descriptors' explanatory power. In detail, predictor variables were always entered in a block-wise fashion, based on toolbox origin or ML descriptor group (see Table 11 in the Supplementary File, for a list of all predictor blocks). Within each block, a stepwise variable selection procedure (forward/backward-method with p_in = .05/p_out = .10) was performed. We finally computed (incremental) adjusted R^2 for each predictor block to estimate the explanatory power of the different descriptors.

3.2. Results

3.2.1. Accuracy of ML Classifiers

ML of the various classifiers led to very robust results (see Table 10 in the Supplementary File). ML Classification of *musical style* and the presence of *vocals* was accomplished with over 90% accuracy. By contrast, recognition of *instrumentation* (81% accuracy), *production timbre* (82% accuracy), and *vocals gender* (76% accuracy) turned out to be more challenging.

3.2.2. Obtained Prediction Models

Across all computational models identified by hierarchical stepwise regression (see Table 11 in the Supplementary File), *musical style* and *instrumentation*, as learned by the ML algorithm play the most significant role in variance explanation of perceived musical expression ($\overline{R^2_{adj}}$ [style] = .191, $\overline{R^2_{adj}}$ [instrumentation] = .183). Also, rhythmic features extracted by the IRCAM beat toolbox explain a substantial amount of variance ($\overline{R^2_{adj}}$ [IRCAM beat] = .151), in particular related to perceived *Authenticity* and *Timeliness* of a musical excerpt. ($\overline{R^2_{adj}}$ [instrumentation] = .183). Finally, the remaining predictor blocks play a lesser important role in variance explanation, suggesting various interacting levels and facets of musical meaning. In the following, single prediction models obtained for the five musical expression factors will be described in detail.

3.2.3. Valence

As already suggested by the overall results in Table 11 in the Supplementary File, *musical style* adherence probabilities (R^2_{adj} = .177) and instrumentation (R^2_{adj} = .132) play a crucial role in variance explanation of perceived Valence. Table 2 presents results of the hierarchical stepwise regression model, revealing *Hip Hop*, *Blues*, and *Oriental* as slightly associated with negative *Valence*, whereas *Samba, Rock and Roll*, and *Latin* are related to more positive *Valence*. Additionally, the probability of a track containing an *electric guitar* implies more negative *Valence*, possibly because electric guitars are often connotated as 'aggressive.' Finally, two production sound descriptors emerged in the list of the ten most potent predictors, namely the proportion of *noise energy* in the audio signal and its *periodicity* both being associated with more positive *Valence*.

Table 2. Hierarchical stepwise regression model predicting *Valence*, ten best predictors with largest β values.

Predictor	β	SE	t	p
Style (ML): HipHop	−0.178	0.035	−5.064	< .001
IRCAM descriptor: total noise energy	0.166	0.061	2.725	.007
IRCAM descriptor: periodicity	0.139	0.043	3.216	.001
Style (ML): Samba	0.139	0.035	3.967	< .001
Style (ML): Rock and Roll	0.136	0.033	4.058	< .001
Style (ML): Latin	0.129	0.035	3.712	< .001
IRCAM descriptor: sharpness SD	−0.127	0.044	−2.892	.004
Style (ML): Blues	−0.121	0.033	−3.709	< .001
Style (ML): Oriental	−0.119	0.032	−3.740	< .001
Intrumentation (ML): Electric Guitar	−0.118	0.035	−3.354	.001

3.2.4. Arousal

Adherence of an audio track to a *style* (R^2_{adj} = .239) and its *instrumentation* (R^2_{adj} = .219) also play a dominant role in variance explanation of *Arousal* (see Table 3). Musical *styles* such as *Downbeat*, *Balearic, Reggae, Boogie*, and *Soul* commonly associated with lower tempi and relaxation and calmness are major predictors of lowered arousal. The three best predictors, however, are directly related to production sound: Firstly, the more *harmonic energy* an audio track contains, the less it is perceived as arousing. This corroborates common knowledge in music psychology and psychoacoustics stating that the noisier (i.e., less harmonic) an audio track is, the more it is perceived as arousing (Juslin & Laukka, 2004). Secondly, the mean and standard deviation of the *first MFCC band* highlight the arousing role of the amount and fluctuation of low-frequency content (i.e., pumping beats) in a musical track. Finally, the model supports everyday experience that the more *percussive* and the less *warm* the sound of a musical track is, the more it will be perceived as arousing.

3.2.5. Authenticity

Regarding the attribution of *Authenticity, rhythmic* features as measured by the *IRCAM beat* toolbox (R^2_{adj} = .214), as well as the adherence to a musical *style* and associated image (R^2_{adj} = .169), are crucial for variance explanation. Amongst the most critical features (Table 4), eight *styles* are negatively related to authenticity, four of them from the electronic dance music genre. This resonates with findings that the use of synthesised instruments, studio production, and more contemporary styles are often associated with lesser authenticity (Wu, Spieß, & Lehmann, 2017). The fact that instrumental (i.e., non-vocal) music, in general, predicts lesser *Authenticity* can be explained by assuming that it is foremost the vocal intonation of a singer that helps to represent human values such as being 'honest.' Finally, two production sound descriptors, namely *total harmonic energy* and fluctuations in the higher mid-frequency range (MFCC Band 05 [SD]) appear in the list. The former is associated with 'non-distorted' acoustical sounds in a track contributing to perceived *Authenticity*; the latter might be related to pulsating synthetic sounds occurring in electronic music and thus leading to less perceived authenticity.

3.2.6. Timeliness

Analogously to previous musical expression factors, musical *style* is also crucial for the variance explanation of *Timeliness* (R^2_{adj} = .213), together with *instrumentation* (R^2_{adj} = .216) and features related to *rhythm* (R^2_{adj} [IRCAM beat] = .297). Nine of the ten most potent single vari-

Table 3. Hierarchical stepwise regression model predicting *Arousal*, ten best predictors with largest β values.

Predictor	β	SE	t	p
IRCAM descriptor: total harmonic energy	−0.210	0.046	−4.531	< .001
MFCC Band 01 SD	0.193	0.037	5.205	< .001
MFCC Band 01 Mean	0.146	0.056	2.592	.010
Style (ML): Downbeat	−0.141	0.024	−5.856	< .001
IRCAM descriptor: Percussivity	0.124	0.031	4.029	< .001
Style (ML): Balearic	−0.114	0.023	−4.954	< .001
Production Timbre (ML): warm	−0.109	0.028	−3.831	< .001
Style (ML): Reggae	−0.100	0.023	−4.386	< .001
Style (ML): Boogie	−0.100	0.022	−4.476	< .001
Style (ML): Soul	−0.100	0.023	−4.416	< .001

Table 4. Hierarchical stepwise regression model predicting *Authenticity*, ten best predictors with largest β values.

Predictor	β	SE	t	p
Style (ML): UK Funky	−0.205	0.029	−7.125	< .001
Style (ML): Hip Hop	−0.203	0.035	−5.745	< .001
IRCAM descriptor: total harmonic energy	0.188	0.053	3.535	< .001
Vocals present (ML): no	−0.175	0.042	−4.148	< .001
Style (ML): Dubstep	−0.173	0.029	−6.000	< .001
Style (ML): Electro (ML)	−0.165	0.029	−5.635	< .001
MFCC Band 05 (SD)	−0.151	0.036	−4.136	< .001
Style (ML): Drum and Bass (ML)	−0.145	0.031	−4.740	< .001
Style (ML): Krautrock (ML)	−0.135	0.029	−4.714	< .001
Style (ML): Tech House (ML)	−0.134	0.031	−4.287	< .001

ables constitute *musical styles* that can be regarded as rather traditional (e.g., *German Schlager, Chanson, Classical Jazz, Country*) and were thus negatively associated with perceived Timeliness (Table 5). Also, the *proportion of noise* (i.e., non-harmonic) energy in an audio signal was a positive predictor of timeliness. High total noise energy often results from using (non-harmonic) synthetic sounds and effects, as typically found in rather modern industrial-sounding music styles (e.g., *Dubstep*).

3.2.7. Eroticity

Finally, concerning perceived *Eroticity* of a musical track, *instrumentation* (R^2_{adj} = .205) and *style* (R^2_{adj} = .155) explained the most substantial amount of variance (Table 6). A musical track is more likely to be perceived as erotic if containing female vocals, in particular as opposed to an instrumental track. In contrast, the presence of an *electric guitar* (presumably associated with rather 'manly' musical genres such as *Rock* and *Heavy Metal*) contributed negatively to perceived *Eroticity*. Moreover, *Soul* music is a positive predictor of *Eroticity*, whereas tracks from the styles *HipHop, Oriental*, and *UK Funky* styles are perceived as less erotic. Finally, a *warm* timbre as well as high mean values in the 11th MFCC band are related to stronger perceived *Eroticity* of a musical piece. The latter might be related to aspirated female vocals which are perceived as erotic.

3.3. Discussion

Findings from Study 2 demonstrate that it is possible to predict major portions of the non-individual parts of perceived expression in popular music with the aid of audio signal analysis, MIR, and ML techniques (RQ3). Inspecting obtained computational prediction models and addressing RQ4, it turned out that questions of musical *style, instrumentation,* and *rhythm* dominate perceived affective and semantic expressivity of popular music. Meanwhile, production *sound, keys, chords* and *lyrics* play only a minor role. Finally, we also found differences regarding the importance of specific musical elements when it came to different dimensions of musical expression. However, these were largely in line with existing research literature and too complex to be discussed here in further detail due to space limitations.

4. General Discussion

With the present contribution, we empirically compared different ways of explaining music liking in the 'push scenarios' which are becoming more prevalent in the age of digital media. Aiming at demonstrating the importance of the hitherto underestimated role of perceived musical expression, we compared its explanatory power with that of the received genre preference approach while controlling for well-known 'cognitive side-effects' in mu-

Table 5. Hierarchical stepwise regression model predicting *Timeliness*, ten best predictors with largest β values.

Predictor	β	SE	t	p
Style (ML): Schlager	−0.204	0.022	−9.270	< .001
Style (ML): Balkan	−0.202	0.022	−9.058	< .001
Style (ML): Oriental	−0.198	0.022	−9.139	< .001
Style (ML): Chanson	−0.195	0.023	−8.459	< .001
Style (ML): Asia	−0.176	0.022	−7.852	< .001
Style (ML): Calypso	−0.168	0.024	−6.941	< .001
Style (ML): Latin Style	−0.166	0.024	−6.984	< .001
Style (ML): Classical Jazz	−0.165	0.025	−6.606	< .001
IRCAM descriptor: Total Noise Energy	0.158	0.039	4.042	< .001
Style (ML): Country	−0.155	0.023	−6.618	< .001

Table 6. Hierarchical stepwise regression model predicting *Eroticity*, ten best predictors with largest β values.

Predictor	β	SE	t	p
Vocals (ML): no	−0.221	0.039	−5.617	< .001
Female vocals (ML): yes	0.221	0.044	4.987	< .001
Style (ML): HipHop	−0.156	0.038	−4.082	< .001
Instrumentation (ML): Electric Guitar	−0.149	0.034	−4.371	< .001
Style (ML): Oriental	−0.137	0.031	−4.409	< .001
Production timbre (ML): Dark	−0.135	0.034	−3.958	< .001
MFCC Band 11 MEAN	0.130	0.034	3.892	< .001
Style (ML): Soul	0.127	0.032	4.015	< .001
IRCAM key: Db (effect-coded)	0.121	0.165	3.368	0.001
Style (ML): UK Funky	−0.116	0.031	−3.686	< .001

sic liking (Study 1). While the latter (foremost familiarity, but also prominence and popularity) were shown to explain a fair amount of variance, the explanatory potential of genre affinities expectedly turned out to be minor compared to the influence of perceived musical expression. Notably, advancing the state of research in the field, we demonstrated that perceived semantic meaning is as important as perceived strength of expressed emotions when it comes to explaining liking for previously unknown music. In summary, all our hypotheses were confirmed. Additionally, attribution of meaning towards presented and largely previously unknown music was found to be particularly homogenous across sociodemographic groups and countries. Similarly, sociodemographic differences regarding the weighting of different musical expression dimensions for music liking turned out to be small. Nevertheless, a significant degree of individual (presumably also encompassing situational) heterogeneity in musical meaning attribution still exists.

Based on these findings, we used MIR and ML in Study 2 to test the algorithmic predictability of the perceived musical expression. As expected, it turned out that meaning attribution concerning popular music appears to a substantial degree to be uniform and rule-based. The explanatory power of musical style in Study 2, when compared to our findings regarding the related, but coarser concept of musical genre in Study 1 hints at the possibility that fine-grained, highly standardised algorithmic style descriptors instead of subjective ratings might form a solution for the 'genre dilemma' discussed in the introduction and the research literature (Brisson & Bianchi, 2019). Taken together, the findings point out the importance of *communicative aspects of popular music* when it comes to empirically explaining and predicting music liking in basic musicological research on music preferences as well as in applied scenarios such as music recommendation algorithms.

Overall, the findings of our two studies stress the importance of a hitherto underdeveloped area in quantitative music reception research: music semantics. Previously, music psychology tended to analyse popular music predominantly as an art form or as a sensual media offering that may emotionally move us and entrain our bodies into dancing. However, with this contribution, we suggest conceiving of popular music also as a semiotic device, a carrier of complex meanings, similar to oral language or any other communicative sign system. This can be interpreted in terms of music's anthropological main functions of self-awareness and social relatedness (Schäfer, Sedlmeier, Städtler, & Huron, 2013). Popular music once more presents itself as something that brings people together, not only in terms of affect, but also in terms of identity and values (Frith, 1996). Until now, however, expression of these aspects in pop music have been researched predominantly by cultural studies scholars, either by employing discourse analysis (Machin & Richardson, 2012) or interpretive interview studies (Hesmondhalgh, 2007). Here, our paper demonstrates that meaning structures in music excert strong measurable quantitative effects, and that these are relatively homogenous across social groups and cultures, making them well-suited for statistical analyses with larger samples and also largely predictable by ML.

Several limitations regarding the generalisation of our findings have to be addressed. Popular music is a complex global cultural phenomenon, and the existing repertoire of genres, styles, artists and scenes is vast. Our study was only able to analyse music listeners from three European countries and only employed a very limited, though comparatively heterogeneous selection of popular music. In general, it appears hard to claim with any sample of any size to have a proper representation of popular music as such, due to its breadth, complexity, and everchanging nature. Furthermore, we conducted a secondary analysis of popular music titles all deemed suitable for branding purposes, which necessarily leads to the exclusion of more extreme, fringe styles of pop music. The finding that the ML features operationalising the content of song lyrics did not play a substantial role in the final models of Study 2 could be related to this fact. Further, it is crucial to acknowledge that—by design—the musical expression space operationalised by the GMBI_15 questionnaire instrument does not exhaust the full breadth of musical expression. Hence, further research should expand from our findings, especially with a sharper focus on the expression of identity and humanistic, political and religious values.

Summarising implications, we propose that musicology should consider taking a shift in research focus 'from mood to meaning' (Vorderer & Reinecke, 2015) that has already taken place in media research. The observed importance of the authenticity dimension further parallels the claim of a 'truth-seeking media recipient' that has recently gained prominence in media entertainment research (Oliver & Raney, 2011). Also, our findings suggest that popular music's meaning expression is a legitimate field of research for applying communication theory, because it appears to act like a rule-based language, as demonstrated by Study 2. It can thus be analysed similarly to linguistic or pictorial content and may also form an independent variable in media reception and effects research (Shevy, 2013). In conclusion, we argue that the results of our studies are of importance not only for the music industry and musicology but also for media studies and communication science.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 688122. We express our gratitude to Geoffroy Peeters and his IRCAM team for their contributions to the ML parts of Study 2. Similarly, we thank our project partner HearDis for contributing the stimulus material and tag knowledge, as well as Dr Andreas Schönrock for stimulus pool preparation.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Airoldi, M., Beraldo, D., & Gandini, A. (2016). Follow the algorithm: An exploratory investigation of music on YouTube. *Poetics*, *57*, 1–13. https://doi.org/10.1016/j.poetic.2016.05.001

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York, NY: Appleton Century-Crofts.

Bourdieu, P. (1984). *Distinction*. Polity Press. Bouvier, G., & Machin, D. (2013). How advertisers use sound and music to communicate ideas, attitudes and identities: A multimodal critical discourse approach. In B. Pennock-Spek & D. S. Rubio (Eds.), *The multimodal analysis of television commercials*. Valencia: University of Valencia Press.

Brisson, R., & Bianchi, R. (2019). On the relevance of music genre-based analysis in research on musical tastes. *Psychology of Music*. Advance online publication. https://doi.org/10.1177/0305735619828810

Brunswik, E. (1952). *The conceptual framework of psychology*. Illinois, IL: University of Chicago Press.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Chmiel, A., & Schubert, E. (2017). Back to the inverted-U for music preference: A review of the literature. *Psychology of Music*, *45*(6), 886–909. https://doi.org/10.1177/0305735617697507

Drott, E. (2018). Why the next song matters: Streaming, recommendation, scarcity. *Twentieth-Century Music*, *15*(3), 325–357. https://doi.org/10.1017/S1478572218000245

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00487

Egermann, H., Grewe, O., Kopiez, R., & Altenmüller, E. (2009). Social feedback influences musically induced emotions. *Annals of the New York Academy of Sciences*, *1169*(1), 346–350. https://doi.org/10.1111/j.1749-6632.2009.04789.x

Feezell, J. T. (2017). It's not only rock and roll: The influence of music preferences on political attitudes. In *Music as a platform for political communication* (pp. 167–186). Hershey, PA: IGI Global.

Fricke, K. R., & Herzberg, P. Y. (2017). Personality and self-reported preference for music genres and attributes in a German-speaking sample. *Journal of Research in Personality*, *68*, 114–123. https://doi.org/10.1016/j.jrp.2017.01.001

Frith, S. (1996). *Performing rites: On the value of popular music*. Oxford: Oxford University Press.

Greenberg, D. M., Kosinski, M., Stillwell, D. J., Monteiro, B. L., Levitin, D. J., & Rentfrow, P. J. (2016). The song is you: Preferences for musical attribute dimensions reflect personality. *Social Psychological and Personality Science*, *7*(6), 597–605. https://doi.org/10.1177/1948550616641473

Hennion, A. (2001). Music lovers: Taste as performance. *Theory Culture Society*, *18*(5), 1–22. https://doi.org/10.1177/02632760122051940

Herzog, M., Lepa, S., & Egermann, H. (2016). *Towards automatic music recommendation for audio branding scenarios*. Paper presented at 17th International Society for Music Information Retrieval Conference (ISMIR), New York, NY.

Herzog, M., Lepa, S., Egermann, H., Schönrock, A., & Steffens, J. (2020). Towards a common terminology for music branding campaigns. *Journal of Marketing Management*, *36*(1/2), 176–209. https://doi.org/10.1080/0267257X.2020.1713856

Herzog, M., Lepa, S., Steffens, J., Schönrock, A., & Egermann, H. (2017a). Predicting musical meaning in audio branding scenarios. In E. Van Dyck (Ed.), *Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*. Ghent: ESCOM.

Herzog, M., Lepa, S., Steffens, J., Schönrock, A., & Egermann, H. (2017b). *Predicting musical meaning in audio branding scenarios*. Paper presented at Conference of the European Society for the Cognitive Sciences of Music (ESCOM), 2017, Ghent.

Hesmondhalgh, D. (2007). Audiences and everyday aesthetics. Talking about good and bad music. *European Journal of Cultural Studies*, *10*(4), 507–527. https://doi.org/10.1177/1367549407081959

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, *48*(2), 246–268.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306–340.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(6), 1797–1813.

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217. https://doi.org/10.1080/0929821042000317813

Klimmt, C. (2011). Media psychology and complex modes of entertainment experiences. *Journal of*

Media Psychology: Theories, Methods, and Applications, *23*(1), 34–38. https://doi.org/10.1027/1864-1105/a000030

Krämer, B. (2018). Online music recommendation platforms as representations of ontologies of musical taste. *Communications*, *43*(2), 259–281. https://doi.org/10.1515/commun-2017-0056

Kristen, S., & Shevy, M. (2013). A comparison of German and American listeners' extra musical associations with popular music genres. *Psychology of Music*, *41*(6), 764–778. https://doi.org/10.1177/0305735612451785

Lahire, B. (2008). The individual and the mixing of genres: Cultural dissonance and self-distinction. *Poetics*, *36*(2/3), 166–188. https://doi.org/10.1016/j.poetic.2008.02.001

Lepa, S. (2010). Media appropriation from a critical realist point of view? Theoretical premises and empirical research. In A. S. da Silva, J. C. Martins, L. Magalháes, & M. Goncalves (Eds.), *Comunicação, cognição e media. Volume 1* (pp. 379–390). Lisbon: Aletheia.

Lepa, S., & Hoklas, A.-K. (2015). How do people really listen to music today? Conventionalities and major turnovers in German audio repertoires. *Information, Communication & Society*, *18*(10), 1253–1268. https://doi.org/10.1080/1369118X.2015.1037327

Lonsdale, A. J., & North, A. C. (2009). Musical taste and ingroup favouritism. *Group Processes & Intergroup Relations*, *12*(3), 319–327. https://doi.org/10.1177/1368430209102842

Lonsdale, A. J., & North, A. C. (2012). Musical taste and the representativeness heuristic. *Psychology of Music*, *40*(2), 131–142. https://doi.org/10.1177/0305735611425901

Machin, D., & Richardson, J. E. (2012). Discourses of unity and purpose in the sounds of fascist music: A multimodal approach. *Critical Discourse Studies*, *9*(4), 329–345. https://doi.org/10.1080/17405904.2012.713203

Madison, G., & Schiölde, G. (2017). Repeated listening increases the liking for music regardless of its complexity: Implications for the appreciation and aesthetics of music. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00147

Mellander, C., Florida, R., Rentfrow, P. J., & Potter, J. (2018). The geography of music preferences. *Journal of Cultural Economics*, *42*(4), 593–618. https://doi.org/10.1007/s10824-018-9320-x

Oliver, M. B., & Raney, A. A. (2011). Entertainment as pleasurable and meaningful: Identifying hedonic and eudaimonic Motivations for entertainment consumption. *Journal of Communication*, *61*(5), 984–1004. https://doi.org/10.1111/j.1460-2466.2011.01585.x

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics*, *21*(4), 243–258. https://doi.org/10.1016/0304-422X(92)90008-Q

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 123–205). Cambridge, MA: Academic Press.

Pontara, T., & Volgsten, U. (2017). Musicalization and mediatization. In O. Driessens, G. Bolin, A. Hepp, & S. Hjarvard (Eds.), *Dynamics of mediatization* (pp. 247–269). New York, NY: Springer International Publishing. https://doi.org/10.1007/978-3-319-62983-4_12

Prior, N. (2013). Bourdieu and the sociology of music consumption: A critical assessment of recent developments. *Sociology Compass*, *7*(3), 181–193. https://doi.org/10.1111/soc4.12020

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology*, *100*(6), 1139–1157. https://doi.org/10.1037/a0022406

Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D., & Levitin, D. J. (2012). The song remains the same: A replication and extension of the music model. *Music Perception: An Interdisciplinary Journal*, *30*(2), 161–185. https://doi.org/10.1525/mp. 2012.30.2.161

Roose, H., & Stichele, A. V. (2010). Living room vs. concert hall: Patterns of music consumption in flanders. *Social Forces*, *89*(1), 185–207. https://doi.org/10.1353/sof.2010.0077

Schäfer, T., & Mehlhorn, C. (2017). Can personality traits predict musical style preferences? A meta-analysis. *Personality and Individual Differences*, *116*, 265–273. https://doi.org/10.1016/j.paid.2017.04.061

Schäfer, T., Sedlmeier, P., Städtler, C., & Huron, D. (2013). The psychological functions of music listening. *Frontiers in Psychology*, *4*(511). https://doi.org/10.3389/fpsyg.2013.00511

Searle, S. R. (1995). An overview of variance component estimation. *Metrika*, *42*(1), 215–230. https://doi.org/10.1007/BF01894301

Shevy, M. (2008). Music genre as cognitive schema: Extramusical associations with country and hip-hop music. *Psychology of Music*, *36*(4), 477–498. https://doi.org/10.1177/0305735608089384

Shevy, M. (2013). Integrating media effects research and music psychology. In S.-L. Tan, A. J. Cohen, S. D. Lipscomb, & R. A. Kendall (Eds.), *The psychology of music in multimedia* (pp. 66–88). Oxford: Oxford University Press.

Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2012). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, *110*(1), 70–75. https://doi.org/10.1073/pnas.1209023110

Stone, A. (2016). Meaning and affect in popular music. In A. Stone (Ed.), *The value of popular music: An approach from post-kantian aesthetics* (pp. 173–211). Nova York, NY: Springer International Publishing. https://doi.org/10.1007/978-3-319-46544-9_6

Tagg, P. (2000). Folk music, art music, popular music: An

axiomatic triangle. In *Kojak. Fifty seconds of television music. Towards the analysis of affect in popular music* (2nd ed., p. 35). Larchmont, NY: Mass Media Music Scholars' Press.

Tagg, P. (2013). *Music's meanings: A modern musicology for non-musos*. Larchmont, NY: Mass Media Music Scholar's Press.

Vlegels, J., & Lievens, J. (2017). Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences. *Poetics*, *60*, 76–89. https://doi.org/10.1016/j.poetic.2016.08.004

Vorderer, P., & Reinecke, L. (2015). From mood to meaning: The changing model of the user in entertainment research. *Communication Theory*, *25*(4), 447–453. https://doi.org/10.1111/comt.12082

Wu, L., Spieß, M., & Lehmann, M. (2017). The effect of authenticity in music on the subjective theories and aesthetical evaluation of listeners: A randomized experiment. *Musicae Scientiae*, *21*(4), 1–23. https://doi.org/10.1177/1029864916676301

About the Authors

Steffen Lepa (PhD), born 1978, is a Postdoc Researcher and Lecturer at Audio Communication Group, TU Berlin. He holds MA studies in media, psychology, computer science and communication science and a PhD in educational and social sciences. He teaches on digital media change, research methodology, and sound design. In 2018–2019 he was a Visiting Professor for Media and Music at Hanover University of Music, Drama and Media. His current key research areas are mediatization, audio reception, audience studies (focus on music streaming and analogue media) and computational research methods.

Jochen Steffens (PhD) born 1981, is a music and sound Researcher and a Professor of Musical Acoustics at Düsseldorf University of Applied Sciences. After completing his PhD, he held postdoctoral positions at McGill University, the Max Planck Institute for Empirical Aesthetics and the TU Berlin, where he was awarded his Habilitation in systematic musicology and psychoacoustics. His research areas include the perception of music and sound, their effects on emotional, cognitive and behavioural processes, and associated practical applications.

Martin Herzog graduated in Computer Science at Humboldt University Berlin in 2011. Since 2016, he is a Research Associate at the Audio Communication Group at TU Berlin. His research areas include music perception, music information retrieval, and audio branding. In his PhD research he focusses on predicting musical meaning from high-level music features. Apart from science, Martin Herzog works as a consultant for digital communications in Berlin since 2011.

Hauke Egermann (PhD) is an Associate Professor in the Department of Music, University of York. He graduated in Systematic Musicology, Media Studies, and Communication Research (MA 2006, Hanover University for Music, Drama and Media). Subsequently, he studied at the Centre for Systems Neurosciences Hanover (PhD, 2009). He was Postdoctoral Research Fellow at the Centre for Interdisciplinary Research in Music Media and Technology (2009-2011, McGill University, Montreal, Canada). In 2016, he was awarded his Habilitation in Musicology at the Technische Universität Berlin. He directs the York Music Psychology Group since 2016.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 205–218
DOI: 10.17645/mac.v8i3.3128

Article

A Computational Approach to Analyzing the Twitter Debate on Gaming Disorder

Tim Schatto-Eckrodt *, Robin Janzik, Felix Reer, Svenja Boberg and Thorsten Quandt

University of Münster, Department of Communication, 48143 Münster, Germany;
E-Mails: tim.schatto-eckrodt@uni-muenster.de (T.S.-E.), robin.janzik@uni-muenster.de (R.J.),
felix.reer@uni-muenster.de (F.R.), svenja.boberg@uni-muenster.de (S.B.), thorsten.quandt@uni-muenster.de (T.Q.)

* Corresponding author

Submitted: 13 April 2020 | Accepted: 7 July 2020 | Published: 13 August 2020

Abstract
The recognition of excessive forms of media entertainment use (such as uncontrolled video gaming or the use of social networking sites) as a disorder is a topic widely discussed among scientists and therapists, but also among politicians, journalists, users, and the industry. In 2018, when the World Health Organization (WHO) decided to include the addictive use of digital games (gaming disorder) as a diagnosis in the International Classification of Diseases, the debate reached a new peak. In the current article, we aim to provide insights into the public debate on gaming disorder by examining data from Twitter for 11 months prior to and 8 months after the WHO decision, analyzing the (change in) topics, actors, and sentiment over time. Automated content analysis revealed that the debate is organic and not driven by spam accounts or other overly active 'power users.' The WHO announcement had a major impact on the debate, moving it away from the topics of parenting and child welfare, largely by activating actors from gaming culture. The WHO decision also resulted in a major backlash, increasing negative sentiments within the debate.

Keywords
addiction; content analysis; entertainment research; games; gaming disorder; social media

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

When television viewing became a mass phenomenon in the 1950s, it only took a few years until the first scientific works on television addiction were published (e.g., Meerloo, 1954). Today, discussions about excessive, pathological behavior primarily concern digital forms of media entertainment, such as social networking sites and/or video games. The debate on the latter reached a new peak in 2018 when the World Health Organization (WHO) decided to include the addictive use of digital games as a diagnosis in the 11th Revision of the International Classification of Diseases (ICD-11). Gaming disorder is defined as:

> A pattern of persistent or recurrent gaming behaviour…manifested by: 1) impaired control over gaming (e.g., onset, frequency, intensity, duration, termination, context); 2) increasing priority given to gaming to the extent that gaming takes precedence over other life interests and daily activities; and 3) continuation or escalation of gaming despite the occurrence of negative consequences. The behaviour pattern is of sufficient severity to result in significant impairment in personal, family, social, educational, occupational or other important areas of functioning. (WHO, 2019)

Some scholars support the idea of gaming disorder being recognized in official manuals. They argue that this

is a prerequisite to establish adequate treatment for the improvement of public health. From a societal perspective, a possible pathologization of players is of lower priority to this goal (e.g., Rumpf et al., 2018). Others, especially from the social sciences and communication science, question the scientific basis of this decision and warn against moral panics (e.g., van Rooij et al., 2018). Through a certain media presentation, gaming might be characterized as a threat, thus pathologizing normal behavior and putting scientific results in a different light (e.g., Bowman, 2016; Markey & Ferguson, 2017).

This is reflected in traditional mass media such as newspapers and television. Reports contain numerous portrayals of extreme cases: A mother starved her daughter to death because of her gaming habits (Thompson, 2011), a gamer died from thrombosis because of playing a game for 22 uninterrupted days (McCrum, 2015), and a desperate father tried to deter his son from playing by hiring other gamers to target his avatar repeatedly (Kleinman, 2013). In addition to traditional media and scientific outlets, the debate is increasingly taking place on social networking sites. A distinctive feature of these sites is that a diverse group of stakeholders are active participants in the debate; gamers, in particular, take part in it.

While the discussion of gaming disorder within academia can be understood on the basis of the many debate articles published in scientific journals, such as the *Journal of Behavioral Addictions* (e.g., Aarseth et al., 2017; Billieux et al., 2017; Griffiths, Kuss, Lopez-Fernandez, & Pontes, 2017; Rumpf et al., 2018; van den Brink, 2017; van Rooij et al., 2018), the public debate is more fragmented and less tangible. Thousands of social media posts, blog articles, and videos on the topic have been published by a very diverse set of actors. This large amount of data makes the use of traditional methods of content analysis almost impossible and requires innovative methods suitable for the analysis of large-scale datasets. In recent years, a new discipline known as 'computational social science' emerged, developing and refining the tools necessary for these kinds of large-scale analyses (Conte et al., 2012). Fields like political science (e.g., Hopkins & King, 2010), communication science (e.g., van Atteveldt & Peng, 2018), and subfields like journalism studies (e.g., Boumans & Trilling, 2016) have started to use computational methods in their research, but they have scarcely been utilized in the field of media entertainment research.

By applying automated content analysis to a Twitter dataset, the current study, on the one hand, provides insights into the public debate on gaming disorder, and on the other hand, showcases the usefulness of computational approaches in media entertainment research.

2. The Gaming Disorder Debate in Traditional Media, Science, and Beyond

Systematic analyses of the debate on gaming disorder in media coverage, scientific journals, or on other platforms, such as social networking sites, are rare, and their findings are fragmented. Prior studies predominantly looked at print media and identified addiction as one aspect in the general debate on gaming. Kirkpatrick (2016) examined the gaming discourse in American magazines in the 1980s and found that most articles considered games to be unsuitable for children, but did not differentiate between the addictive potential of games specifically and technology in general. In a review of Chinese historical and media sources, Szablewicz (2010) found that Internet addiction and Internet gaming are often portrayed in a sensationalistic way, suggesting the framing of the debate as a moral panic. Whitton and Maclure (2017) showed that the video game discourse in British print media is dominated by the narrative of naïve video game players becoming addicted because they cannot control the technology. In one of the few quantitative empirical studies, Jung (2019) investigated the Korean media landscape with regard to its stance on gaming regulations. By analyzing daily newspapers, digital news sites, and digital gaming magazines, he identified several frames in the debate, such as child protection, the preservation of the social order, freedom of choice, cultural consequences, and the effectiveness of therapies. Furthermore, Jung (2019) found that conservative, moderate, and specific IT news outlets differed in the extent to which they addressed these topics. While most conservative media emphasized the negative effects of gaming, IT news outlets only exhibited a positive or neutral stance; moderate media were more balanced, yet trending toward a negative opinion.

Taken together, previous works merely focused on traditional media in a national context, ignoring changes over time. However, with rise of social media platforms in the early 21st century, the media landscape as a whole has drastically changed—and with it, the dynamics of public discourse. Social media platforms have enabled almost everyone to not only observe but actively participate in ongoing debates, reaching large audiences that were previously only accessible via traditional mass media. With this change in the dynamics of the public sphere, many hopes were raised about the democratizing potential of these new platforms (Halpern & Gibbs, 2013; Levina & Arriaga, 2014). The ideal of a discursive public sphere—for instance, in the sense of the philosopher Jürgen Habermas (1991)—in which citizens and the political elite find the best solution to social problems together and at eye level, suddenly seemed to be within reach. The low entry barriers also enable actors from civil society to gain access to public discourse and reach a mass audience.

This ideal of openness and inclusivity is especially salient on Twitter, which attracts not only a sizable share of regular media users (22% of American adults use Twitter; Wojcik & Hughes, 2019), but also policy makers, celebrities, activists, and journalists of traditional and new media organizations (Groshek & Tandoc, 2017; Paulussen & Harder, 2014). The dynamics of public dis-

course on Twitter are shaped not only by the aforementioned openness of the platform and the diversity of its users' backgrounds, but also by two characteristic, technical affordances of the platform.

First, new accounts on Twitter are set to be public by default, that is, other users can subscribe or follow the account without asking permission. This follow relationship is asymmetrical, as the number of accounts a user follows can and often does differ from the number of accounts the user is followed by. This technical characteristic enables two types of network structures: the 'one-to-many' structure, where, similar to traditional mass media, single actors reach a large audience and the 'many-to-many' structure, where groups of users communicate among themselves. However, the potential reach of a user is not just defined by their follow network. The mechanic of the retweet lets users share a tweet of someone they follow with their own followers, thus making the barriers of the follow networks even more pervious. As a result of these network structures, even accounts with a low number of followers can potentially reach a large audience.

Second, a feature of almost all social media platforms—the hashtag—was first used on Twitter and is a central affordance of the platform. The hashtag makes it easier for Twitter users to find relevant tweets and to make their own posts easier for other users to discover. This feature also made it possible to quickly find and participate in ongoing debates, connecting different users with each other and potentially raising awareness about trending topics. In recent years, hashtags have also gained relevance in political activism. Campaigns and movements using a hashtag both as branding and a communication tool played a significant role in the context of many topics, such as feminism (e.g., #meToo, #whyIStayed), anti-racism (e.g., #blackLivesMatter, #takeAKnee), and other political movements (e.g., #ArabSpring, #UmbrellaRevolution). Hashtags have also been used by malicious actors to insert themselves into a discourse with the intent to disrupt the ongoing debate or to push their own political views, a practice called 'hashtag hijacking' (Hadgu, Garimella, & Weber, 2013; VanDam & Tan, 2016). A notable example of both hashtag hijacking and a movement using a hashtag to organize was #gamerGate, which was "spawned by individuals who purported to be frustrated by a perceived lack of ethics within gaming journalism" (Massanari, 2017, p. 330). Partly by outside agitation and hashtag hijacking by right-wing groups on 4chan, #gamerGate "became a campaign of systematic harassment of female and minority game developers, journalists, and critics and their allies" (p. 330). As evidenced by #gamerGate, Twitter, as a platform, has a history of video game-related activism.

An analysis of the debate on Twitter is particularly interesting, as the technical affordances of the platform enable dynamics of public discourse vastly different from those of the traditional media landscape. Following the theoretical considerations on Twitter as a platform for public discourse, our research questions focus on the actors, topics, and tone present in the debate, as well as potential changes in these categories arising from the decision by the WHO to include the addictive use of digital games as a diagnosis in the ICD-11. To our knowledge, prior research has not examined the debate on this level.

In order to investigate the claim of a more diverse debate based on the general heterogeneity of users on social networking sites, our first interest centered on the participating actors. We were not only interested in the opinions and background of the actors themselves, but also in determining whether their motivations were genuine. We wanted to know if they were actually interested in the debate or whether they were trying to disrupt the discourse in an orchestrated fashion (i.e., trolling). In traditional media, it is mainly scientists, politicians, and experts—or people who are presented as such—who have a say. By allowing anyone to post for the general public, social media involves a more heterogeneous group in the debate: gamers, gaming communities, and those affected by negative consequences may also contribute. Therefore, our first research question was:

RQ1: Which actors participate in the debate?

Given the range of issues discovered in prior research on traditional media, our second question asks if this also holds true for social media. One could assume that topics are being discussed that are not included in the scholarly debate and the reporting of traditional media. Therefore, we asked:

RQ2: What topics are being discussed?

Considering the two factions on gaming disorders in academia and differences in traditional media reports based on their background, we were interested in the sentiments expressed by the actors. Thus, we asked:

RQ3: How is the tone of the overall debate?

The dataset available to us also offered the interesting option of looking at the development of the debate over time. We wanted to know whether the WHO decision had an impact on the debate, and how, if at all, the answers to the previous research questions differed before and after the WHO decision. Our last research question was:

RQ4: How did the WHO decision influence the debate?

3. Data and Methods

3.1. Data

To answer these questions, a large-scale, automated content analysis of N = 16,831 tweets, of which 55.11%

were retweets, posted between March 16th, 2017 and November 15th, 2018 was conducted. The dataset was extracted from Twitter's Decahose stream, which represents a 10% sample of all public tweets and filtered for tweets mentioning the discourse on gaming disorder (for an extended description of the filtering process, see Supplementary File C).

At its peak on June 18th, 2018, a total of 3,308 tweets and retweets were posted, representing roughly 0.01% of the overall volume of all tweets posted during that day. The debate was clearly stimulated by both the release of the beta draft and the official version of the ICD-11, with the corresponding peaks labeled in Figure 1. Within those two peaks, 63.99% of the overall tweet volume was posted.

As we used archival data for our analysis, we were able to analyze tweets that were deleted between their original publishing and the time of analysis. Table 1 shows the proportion of tweets that are still online and the share of tweets that were no longer available online. There are three reasons why a tweet might be offline: 1) the tweet or the account was deleted by the user (deleted), 2) the user set their account to private (protected), or 3) the user was suspended by Twitter (suspended). Compared to a random sample of tweets with a similar age, this share of 73.1% online tweets is relatively

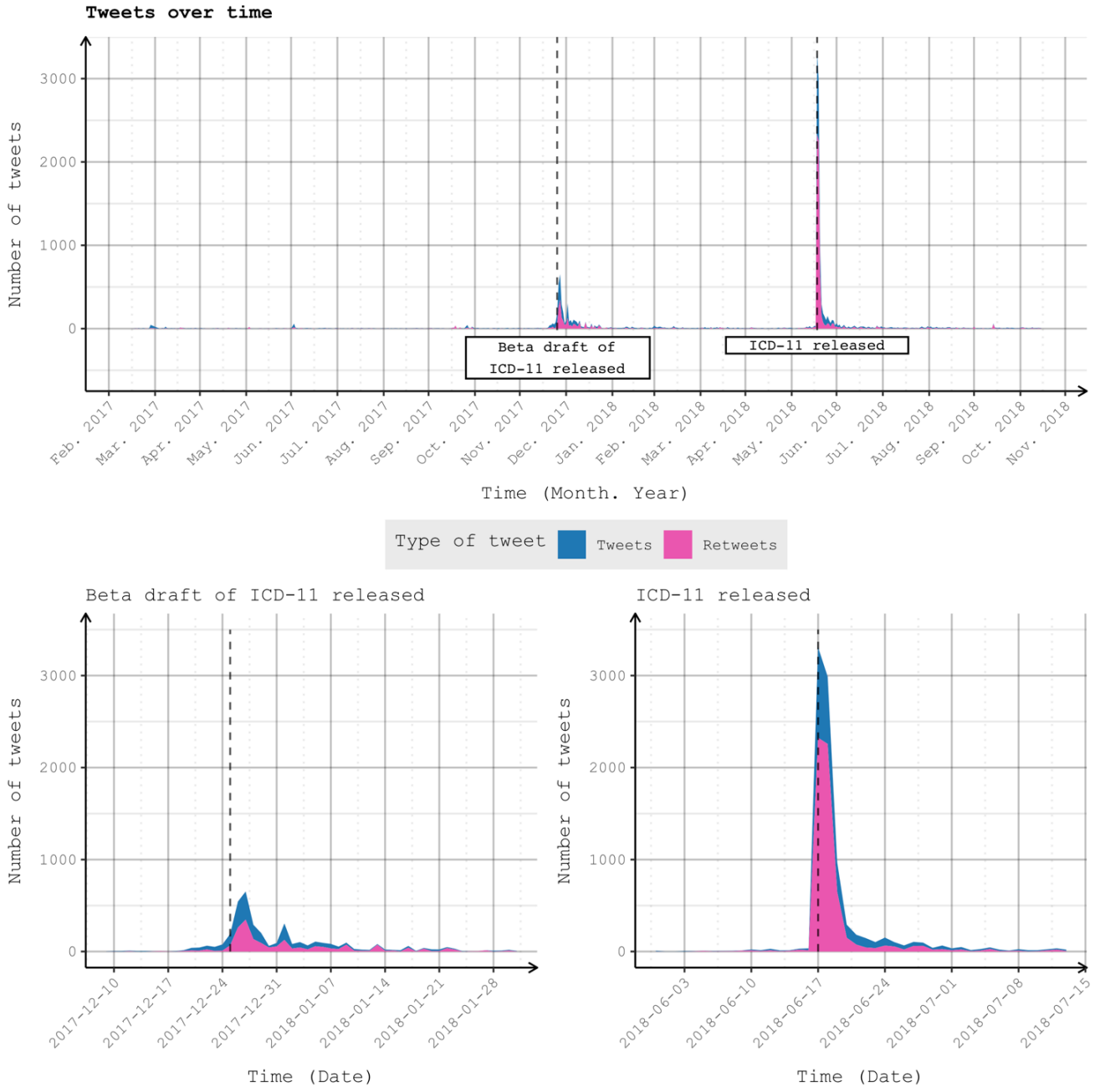


Figure 1. Number of tweets over time. Note: The vertical lines represent the release date of the ICD-11 beta draft (December 26th, 2017) and the release of ICD-11 (June 18th, 2020), respectively.

Table 1. Share of offline tweets.

Online status	%
Online	73.12
Deleted	12.24
Protected	1.30
Suspended	13.34

Notes: $N = 16{,}831$, including retweets. Checked in December 2019, an average 626 days after initial publishing.

high. This can be seen as evidence for an organic, not bot- and/or spam-driven discourse, since Twitter usually deletes spam accounts soon after their participation in a trending topic (Thomas, Grier, Song, & Paxson, 2011).

As we were interested in the effect the WHO decision had on the discourse, we split the dataset into three segments: Segment 1, comprised of tweets posted before the release of the ICD-11 beta draft, Segment 2, comprised of tweets posted after the release of the ICD-11 beta draft and before the official release of the ICD-11, and Segment 3, comprised of tweets posted after the official release of the ICD-11 (see Table 2).

3.2. Methods

The tweets' contents were analyzed using a combination of structural topic modeling, sentiment analysis, and an analysis of used hashtags and present actors. The following chapter gives a detailed description of these methods in order to provide other researchers with the tools to conduct similar analyses and to build upon this framework. All analyses were conducted with R (for a full list of used packages and versions, see Supplementary File A).

3.2.1. Preprocessing

A necessary prerequisite for all kinds of automated and semi-automated content analysis is the procedure of preprocessing. This procedure includes important and impactful decisions by the researchers and is often not well documented or dealt with in a non-transparent way (Denny & Spirling, 2018; Maier et al., 2018). In the current study, we pre-processed the documents by removing non-word characters and tokenizing the documents, by removing stopwords, by stemming, and by pruning. The R code for the preprocessing pipeline used in this study can be found in the Open Science Framework (Schatto-Eckrodt, Janzik, Reer, Boberg, & Quandt, 2020) and a detailed description of the preprocessing steps can be found in the Supplementary File D.

3.2.2. Topic Modeling

Topic modeling "is a computational content-analysis technique that can be used to investigate the 'hidden' thematic structure of a given collection of texts" (Maier et al., 2018, p. 1). In the context of topic modeling, the collection of texts to be analyzed is called a 'corpus,' while each text within the corpus is called a 'document.' In this case, the corpus consisted of every tweet, excluding retweets within the dataset, while a single tweet was a document. Retweets were excluded because including them could potentially introduce a bias towards topics that were represented in often-retweeted tweets. The structural topic model (STM) introduced by Roberts, Stewart, and Tingley (2019) is an extension of other probabilistic topic models, such as the latent dirichlet allocation (Blei, Ng, & Jordan, 2003), which enables researchers to "incorporate arbitrary metadata, defined as information about each document, into the topic model" (Roberts et al., 2019, p. 2). The topics modelled by STM and other topic modeling techniques represent latent content variables that should form a comprehensive representation of the corpus. Like most topic modeling techniques, STM tries to infer these topics from recurring patterns of word occurrence in documents, while ignoring the order of words within each document (i.e., using the bag-of-words assumption; Maier et al., 2018).

STM requires the researcher to choose a number of topics before applying the model. As there is no correct answer to the question of what this number should be (Grimmer & Stewart, 2013), we applied the elbow method on the measures for semantic coherence and held-out likelihood and found a five-topics solution to be optimal. As mentioned before, STM (Roberts et al., 2019) enables researchers to add covariates for topical prevalence to the topic model, which allows the observed metadata to affect the frequency with which a topic is discussed. In this analysis, we modeled the topical prevalence as a function of the time segment matching the publishing time of each tweet, as described above.

Table 2. Time segments.

Segment	Time frame	Number of tweets	%
1	March 16th, 2017–November 30th, 2017	810	10.72
2	December 1st, 2017–June 14th, 2018	2,886	38.20
3	June 15th, 2018–November 15th, 2018	3,859	51.08

Notes: $n = 7{,}555$, excluding retweets.

Models using the raw publishing timestamp of the tweet as a covariate and no covariates at all performed slightly worse than the final model.

3.2.3. Sentiment Analysis

The sentiment analysis was conducted using the opinion lexicon by Hu and Liu (2004), which features a list of English positive (2,001 words) and negative (4,779 words) opinion or sentiment words. This opinion lexicon is widely used for the analysis of social media data (Si et al., 2013; Zhang, Ghosh, Dekhil, Hsu, & Liu, 2011) and is available online, thus enabling replication. In order to provide a robust measure of sentiment, we created a corpus of 88 documents, with each document representing a full week's worth of tweets, including retweets. Pooling the tweet texts and moving the unit of analysis from single tweets to weeks within the debate, enabled us to show the change in the overall sentiment of the debate. For each week, we calculated the share of negative, positive, and neutral sentiment words. The same preprocessing steps were taken for the sentiment analysis as were for the topic modeling, with the exception of the removal of emojis, which were replaced with their Unicode Common Locale Data Repository short names, via the sentimentr package, as including emoticons significantly improves the accuracy of sentiment classification (Hogenboom et al., 2013). As expected, most words (83.29%) fall neither into the positive nor the negative sentiment category. Overall, 14% of all words were identified as negative, and only 2.44% as positive.

3.2.4. Co-Occurrence Graphs

To analyze the hashtags in our dataset, in addition to simple frequency tables, we calculated co-occurrence graphs of the used hashtags. In short, co-occurrence analysis is the analysis of the pairwise connection between elements in a set, with the connection modelled as the occurrence of two elements in a subset of all elements. In content analyses, the elements are often words, the subsets are documents, and the set of all elements is the corpus. Co-occurrence analysis can be used to construct networks (i.e., graphs) representing the connection between words, revealing thematic clusters (Buzydlowski, 2015). As a general method in content analysis, co-occurrence was used even before the introduction of computational methods (Harris, 1957) and has been used specifically for the analysis of social media data in numerous studies (e.g., Aiello et al., 2013; Pervin, Phan, Datta, Takeda, & Toriumi, 2015; Wang, Wei, Liu, Zhou, & Zhang, 2011). In the current study, we used co-occurrence graphs to gain insights into the use of hashtags within the debate. We extracted the hashtags from all tweets, excluding retweets, and built graphs with hashtags as vertices and the co-occurrence of two hashtags in the same tweet as edges. Edges were weighted by the number of co-occurrences.

4. Results

4.1. Actors

Of the 15,498 unique users that are represented in our dataset, 2.39% were verified by Twitter. According to Twitter (2020), "an account may be verified if it is determined to be an account of public interest. This includes accounts maintained by users in music, acting, fashion, government, politics, religion, journalism, media, sports, [and] business." This relatively high number (compared to less than 1% verified users in a random sample) suggests a high involvement of journalistic actors and actors who are otherwise involved in public life (Paul, Khattar, Kumaraguru, Gupta, & Chopra, 2019).

When analyzing the most important actors in a social media discourse, one must consider two characteristics: the reach of a user and the volume of their participation in the discourse. The reach of a user is defined by both their follower count and the number of times they were retweeted in the context of the debate. The higher the reach of a user, the higher the number of users getting into contact with their posts. In contrast, users who participate to a higher extent than the average user might have a lower reach than other users; yet, they are still responsible for a large share of the posts in the debate. These two characteristics are a consequence of the technical affordances of Twitter as a platform. It is possible that a user only mentions the discourse's topic in a single tweet, but—taking retweets into account—is still seen by most people participating in the discourse, thus being overrepresented in most users' timelines.

Those users of the second category, that is, users who participate more frequently than the average user, are shown in Table 3. The top-five most active users are supportive of the WHO decision and try to warn others of the dangers of gaming disorder; they have a parenting or professional education background. Users who oppose the WHO decision are also present in this group, and most of them have a background in gaming culture or technology journalism.

All of the users with the highest reach, except the CNN account, oppose the WHO decision and have a background in gaming culture (see Table 4). These accounts only began participating after the official release of the ICD-11, while the users with the larger extent of participation were part of the discourse months before the release of the beta draft on ICD-11.

To investigate whether a group of users was overrepresented in the dataset, we calculated the distribution of the tweet volume per user share. An equal distribution, that is, a distribution with the share of tweets equal to the share of users, would mean that there are no overly active 'power users.' Looking at all participating users, this kind of equal distribution can be observed. Again, this can be seen as evidence of organic discourse. However, the distribution for retweeted accounts was more skewed. In total, 10% of all tweets that

Table 3. Top users according to their extent of participation.

Username	Tweets by the user	% of all tweets	Verified	View on the WHO decision	Number of followers	Notes
MommyNooz *	32	0.19	No	Supporting	576	Parenting blog
camerondare	31	0.18	Yes	Supporting	2,545	Activist
AdvocateforEd	25	0.15	No	Supporting	26,261	Blog
Lynch39083	22	0.13	No	Supporting	45,143	Scholar/activist
techedvocate	22	0.13	No	Supporting	20,398	Tech blog
eplayuk	16	0.10	No	Opposing	389	Gaming blog
gamescosplay	14	0.08	No	Opposing	664	Gaming blog
gamingthemind	14	0.08	No	Opposing	779	NGO/activists
Pairsonnalites	13	0.08	No	Opposing	4,647	NGO
HealthyWrld*	12	0.07	No	Neutral	21,936	Health blog

Notes: N = 16,831. * = account suspended.

were potentially seen by other users, including retweets, were authored by six users.

4.2. Content

4.2.1. Topics

Analyzing the hashtags revealed that a large share of the discourse was neither centered around #gamingdisorder nor #gamingaddiction. Only 9.84% of all tweets included at least one hashtag. The strategy of including the terms (not hashtags) 'gaming disorder' and 'gaming addiction' for sampling the data was thus an adequate choice. Table 5 shows the 15 most frequently used hashtags. Besides the two topical hashtags used as the sample query (#gamingdisorder and #gamingaddiction) and their variations (#gaming, #addiction, #videogames), we also found hashtags referencing the WHO (#icd11, #who) hashtags related to parenting (#children, #parenting) and hashtags most likely used in journalistic reporting (#bbcbreakfast, #tech, #news).

Comparing the co-occurrence graphs of the hashtags used in Segment 1 and Segment 3 illustrates how the debate changed after the release of the ICD-11 (see Figures 2 and 3). The debate in Segment 1 consisted of two topical groups (education and a broad discussion of gaming disorder). This distinction fades in Segment 3, as the focus shifts away from the educational debate towards a more general discussion.

This shift was also noticeable when applying topic modeling to the data (see Table 6). Topic 3, which represents the topical group of tweets discussing educational and parenting-related arguments, is overshadowed by Topics 1, 2, and 4, which arise in Segment 3. Topic 5, where the classification of gaming addiction is compared to other mental conditions as gender dysmorphia, is represented almost only in Segment 3.

4.2.2. Sentiments

A sentiment analysis revealed that the topic was generally discussed with a relatively negative sentiment (see Figure 4). The tweets in our dataset were, in comparison to a random sample of English language tweets from the same time span, significantly more negative ($t(90) = 8.00$, $p < 0.001$, $d = 1.01$). Both the release of the beta draft and the official release of the ICD-11 resulted in a slight peak of negative sentiment.

Table 4. Top users according to their reach.

Username	Times the user was retweeted	% of all tweets	Verified	View on the WHO decision	Number of followers	Notes
CNN	403	2.39	Yes	Neutral	39,159,370	Media
deadmau5	385	2.29	Yes	Opposing	3,939,972	Musician
GaijinGoombah	318	1.89	No	Opposing	65,049	YouTube CC
BrendoTGB	274	1.63	No	Opposing	349	Regular user
LEGIQN	245	1.46	Yes	Opposing	306,185	Twitch CC
Pamaj	236	1.40	Yes	Opposing	1,174,043	E-sports athlete
NoahJ456	233	1.38	Yes	Opposing	949,370	YouTube CC
TheSmithPlays	227	1.35	No	Opposing	209,071	YouTube CC
Boogie2988	182	1.08	Yes	Opposing	685,290	YouTube CC
CaptainSparklez	170	1.01	Yes	Opposing	4,934,618	YouTube CC

Notes: N = 16,831. CC = content creator.

Table 5. Top hashtags.

Hashtag	Number of occurrences	% of hashtag occurrences
gaming	269	8.92
gamingdisorder	123	4.08
datascience	121	4.01
addiction	113	3.75
health	102	3.38
icd11	97	3.22
mentalhealth	97	3.22
bbcbreakfast	70	2.32
videogames	61	2.02
parenting	59	1.96
who	47	1.56
gamingaddiction	46	1.52
children	40	1.33
news	36	1.19
tech	32	1.06

Note: n = 3,017 hashtag occurrences.

Comparing the sentiment of Segment 1 against the combined sentiment in Segments 2 and 3 showed a significantly more negative sentiment in the latter ($t(66) = 4.87$, $p < 0.001$, $d = 1.09$).

In addition to this quantitative difference in sentiment, we also conducted a term frequency-inverse document frequency analysis, which is a statistical measure to determine the relative importance of a word within a document in a larger corpus, that is, words that are not only frequently used but are also used more frequently in a specific set of documents, as compared to others. Calculating the term frequency-inverse document frequency values for the tweets in our dataset and the random sample of English language tweets reveals that the terms 'irresponsible,' 'ridiculous,' and 'condemn' were the most relevant negative sentiment words associated with the topic.

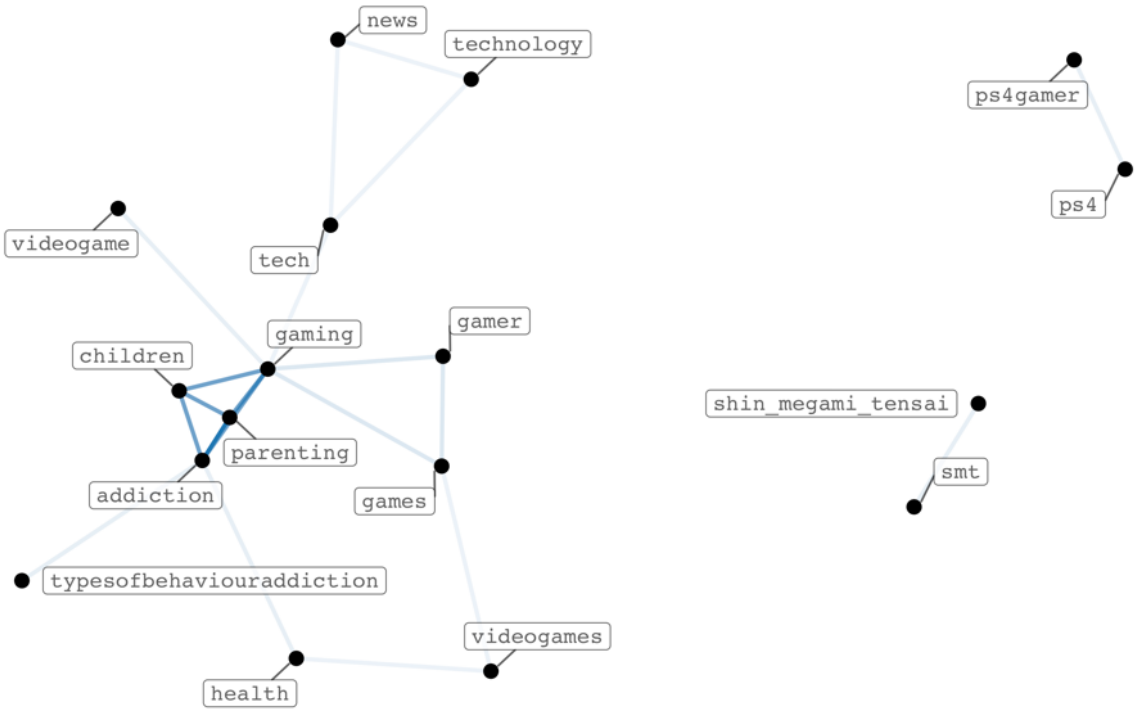


Figure 2. Co-occurrence graph for the top 30 hashtags in segment 1, force-directed layout algorithm by Fruchterman and Reingold (1991).

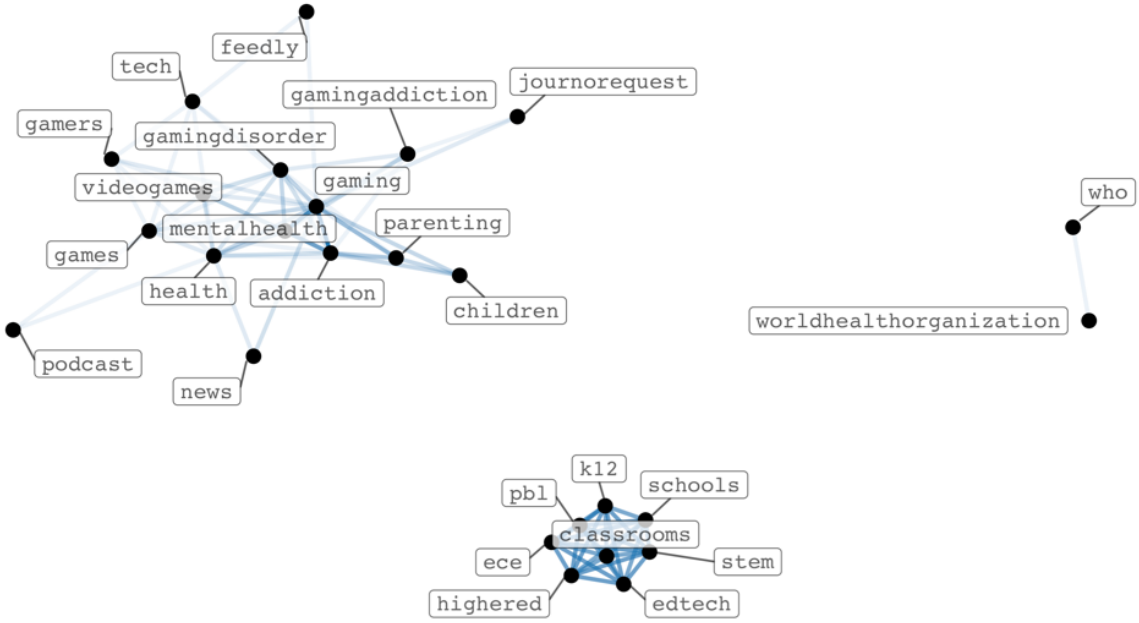


Figure 3. Co-occurrence graph for the top 30 hashtags in segment 3, force-directed layout algorithm by Fruchterman and Reingold (1991).

Table 6. Description, top terms, and representative quote of the topics.

Topic	Description	Terms	Quote
1	Breaking news: Sharing news articles on the WHO decision	Organization, recognizing, officially, mental, first time	"Gaming Disorders Officially Recognized by the World Health Organization: The World Health Organization (WHO) is recognizing 'gaming disorder' as a mental health issue in the beta draft of its upcoming 11th International Classification of Diseases"
2	Opposition: Voicing concerns on the validity of the WHO decision	Video, people, first time, mental, addiction	"'Gaming disorder' is total BS. People, especially adults, see gaming as such a taboo. It's ridiculous. Video games provide experiences that is as humans can't do in real life. Video games can connect people, video games provide a community and a sense of belonging."
3	Education: Information directed at educational professionals and parents regarding gaming addiction	Education, teachers, classroom, help, parents	"Learning Points: Parents need to wake up to gaming addiction"
4	Jokes: Mocking the WHO decision	Addict, play, video, time, mental	"PUBG is the cure to Gaming Disorder. 6 hours on that fucking thing will make you rage so hard you'll probably want to get outside for a bit."
5	Other conditions: Voicing concerns regarding the treatment of other mental afflictions in light of the WHO decision	Classification, time, video, recognizing, addiction	"So Gaming disorder is a mental health condition but gender dismorphia is not this is why government is garbage"

Note: STM of 5,378 documents, $K = 5$, time segment set as prevalence.

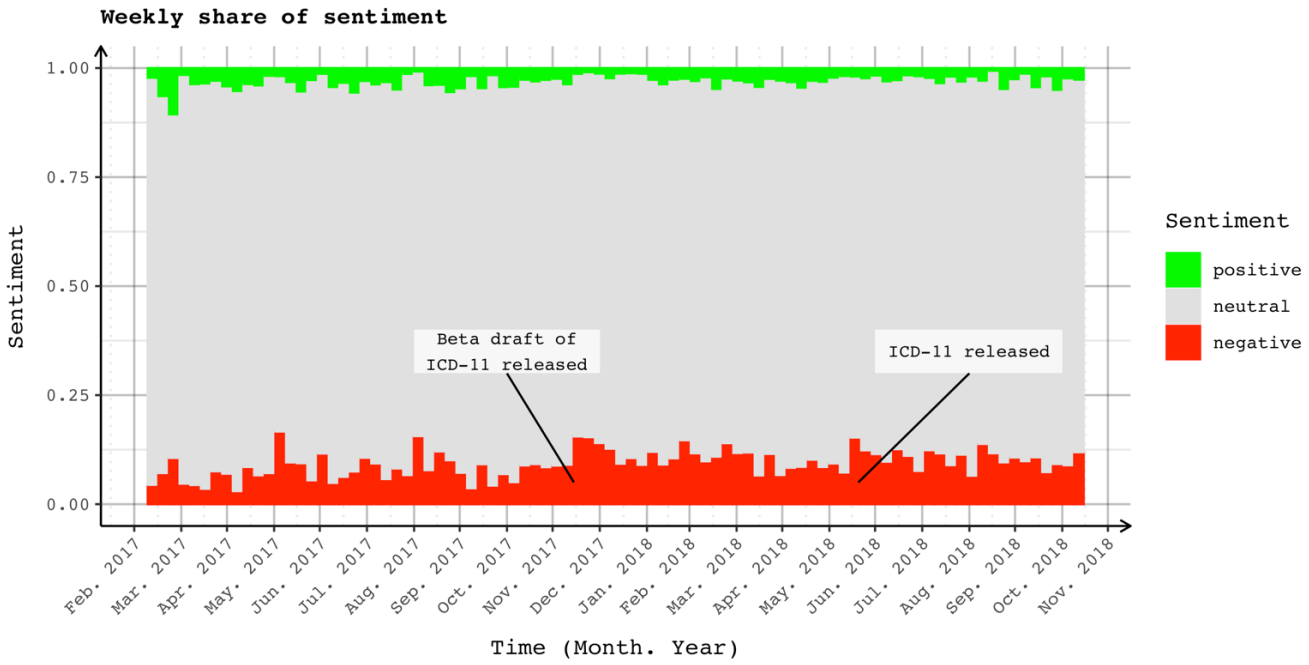


Figure 4. Share of positive and negative sentiment words per week over time.

4.2.3. Linked Websites

Users shared a total of 3,020 unique URLs with 33.92% of all (non-retweet) tweets containing at least one URL. Of these 3,020 URLs, 83.68% were shared only once. The five most often shared URLs link to Twitter's event page on the WHO decision to classify gaming disorder as mental health condition (294 shares), journalistic articles on the topic by the New York based media company Futurism (151 shares), and CNN Health (80 shares), a blog post critical of the WHO decision by the digital entertainment company Saljack Enterprises (66 shares), and a video by YouTube content creator Philip DeFranco, explaining why the WHO decision might villainize games (59 shares).

Most URLs shared by users belong to large online media outlets (e.g., CNN, ABC News) and contain factual reporting on the WHO decision. There are also multiple links to gaming-related blogs with arguments against the WHO decision. The most widely shared scientific content in the dataset was the open debate paper by Aarseth et al. (2017). Other than that, there seems to be little to no circulation of scientific studies within the debate.

5. Discussion and Conclusions

Gaming disorder is currently the most intensively discussed form of problematic entertainment media use. Our analysis shows that social media platforms, such as Twitter, are important forums where different actors (including gamers) discuss the topic. Following an explorative approach, our study was the first to examine the gaming disorder debate based on social media data. It can serve as a basis for more complex future analyses. Overall, our results showed that the debate is organic and not driven by spam accounts or other overly active 'power users.' There is no evidence of any orchestrated campaigns for or against the decision of the WHO.

Further, we see that analyzing the social media discussion has the potential to paint a more heterogenous and balanced picture of the public perception of gaming disorder than an analysis of classical media outlets where particular actors (like politicians, psychologists, and psychiatrists) are perhaps overrepresented. While it can be seen that CNN, as a traditional news medium, has a wide reach, it is largely followed by content creators; their level of participation is also higher. This suggests that traditional news media also play a role in social media for discussion, for instance, as a source of information, but the actual discussion is led by genuine stakeholders, such as the gamers themselves. A central distinction is that the accounts of news media represent organizations, while content creators are individuals who are given the opportunity to express their own opinions. While in traditional media, only public figures appear for their role as experts, on Twitter there is the chance to express one's thoughts through a medium without this prior decision.

With regard to topics and sentiment, our results showed that the social media discussion does more than cover the spectrum that previous studies have shown when examining traditional media (e.g., Kirkpatrick, 2016; Szablewicz, 2010; Whitton & Maclure, 2017). Although negative consequences of gaming are discussed, positive aspects are also emphasized. This suggests a diversification of the debate, which is also found in the academic discussion. Nevertheless, it can be seen that the discussion's sentiment is relatively negative. On the one hand, this is in line with the picture from tradi-

tional media; on the other hand, on closer examination, the terms used might rather indicate that users express their indignation about the WHO's decision. While previous research suggests that the discussion in traditional media focuses primarily on damage control, the topics we found indicate that there is a need to mention aspects that go beyond damage control, for example, the treatment, education, and significance of the decision. Thus, the discussion here is less to be compared with a moral panic, but rather attempts to differentiate.

The release of the ICD-11 draft and official version had a major impact on the debate. The decision moved the debate away from the topics of parenting and child welfare, largely by activating actors from gaming culture. The parenting and education-related discussion still took place, but it was overshadowed by a larger discussion. After the WHO decision, most tweets opposed the classification. Despite the research boom in recent years and the ongoing debate within academia, scientific studies barely played any role in the Twitter debate and research results were hardly considered. This can be interpreted as a hint that research results are perhaps not communicated effectively and are hardly known outside of an academic context.

From a more general perspective, the current study illustrates how computational methods, especially methods of automated content analysis, can be usefully utilized in the context of entertainment research. Using the example of the gaming disorder debate, we showed that these new tools can offer interesting insights into the public perception of risks that may be connected with the use of entertainment media. Future studies may follow this route and examine other diverse topics and their societal perception, such as the discussion about violent media content and aggressiveness or the question whether the use of digital entertainment media may negatively influence users' psychosocial well-being. Furthermore, our analytic framework showcases how social media (understood as a form of entertainment media itself) can be analyzed to identify subtopics and actors within large-scale debates. An interesting approach for future studies may be the combination of computational methods with qualitative methods to gain additional in-depth knowledge of selected networks and data patterns.

6. Limitations

The current study has some limitations. As we exclusively used Twitter data as the basis of our analysis, there might be a bias towards opinions shared by Twitter's relatively male and technophile userbase. The analysis is also limited to English language tweets, so there might be similar discourses in other languages, albeit using different hashtags.

The reduction in corpus size following our preprocessing procedure was relatively large, as a third of all Tweets were not considered in the topic modeling and a large share of tokens was removed due to frequency. This loss of information might mean that some nuanced distinction between topics was not detected. The findings should thus be considered as a broad overview of the debate.

The sentiment analysis conducted in this study used a dictionary-based approach and the dictionary used only includes the binary distinction between negative and positive sentiment words (Hu & Liu, 2004). More sophisticated methods of sentiment analysis enable researchers to investigate more complex emotions like disgust, anger, or surprise either by using a dictionary that includes those categories or by using a corpus-based approach where supervised machine learning techniques are utilized (Strapparava & Mihalcea, 2008). The simple method used in the current study reveals the shift in tone caused by the WHO decision and gives insight into the potential reasoning of users behind their emotional reactions but does not reveal any more detailed information on the sentiment of the debate. Future research might address this using the methods referenced above.

Another limitation of the methods of the current study is the use of topic modeling on a corpus of documents with a relatively short length. As most traditional topic modeling techniques like the latent dirichlet allocation (Blei et al., 2003) and other probabilistic topic models like the STM (Roberts et al., 2019) used in the current study, rely on document-level word co-occurrence patterns to reveal topics. In short texts, as commonly found in social media, this co-occurrence approach may not work very well, as there is only limited word co-occurrence information available in these texts (Jipeng, Zhenyu, Yun, Yunhao, & Xindong, 2019). Future research might mitigate this issue by using methods specifically developed for shorter texts, like the biterm topic model (BTM) by Yan, Guo, Lan, and Cheng (2013).

The exclusion of emojis from the topic modeling was, on the one hand, driven by the differentiation between the topic and the tone of the debate in our research questions and, on the other hand, motivated by the need to reduce the level of noise in the already noisy and sparse data. While being an interesting question, we did not feel confident enough to address the tonality of the discussed topics in a robust way.

In general, as the methods applied in the current study are meant for the analysis of large-scale datasets, the above findings should not be seen as a complete description of every facet of the debate, but as an overview of the discourse, revealing overarching structures and topics worth investigating in greater detail.

Acknowledgments

The research was funded in part by the German Federal Ministry of Education and Research. We acknowledge support by the open access publication fund at the University of Münster. We would also like to thank the reviewers and the Academic Editors for their constructive feedback.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Aarseth, E., Bean, A. M., Boonen, H., Colder Carras, M., Coulson, M., Das, D., . . . van Rooij, A. J. (2017). Scholars' open debate paper on the World Health Organization ICD-11 gaming disorder proposal. *Journal of Behavioral Addictions*, *6*(3), 267–270. https://doi.org/10/gb22vs

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., . . . Jaimes, A. (2013). Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, *15*(6), 1268–1282. https://doi.org/10/f5bzrz

Billieux, J., King, D. L., Higuchi, S., Achab, S., Bowden-Jones, H., Hao, W., . . . Poznyak, V. (2017). Functional impairment matters in the screening and diagnosis of gaming disorder: Commentary on: Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *Journal of Behavioral Addictions*, *6*(3), 285–289. https://doi.org/10/ggps6p

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(4/5), 993–1022.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23.

Bowman, N. D. (2016). The rise (and refinement) of moral panic. In R. Kowert & T. Quandt (Eds.), *The video game debate: Unravelling the physical, social, and psychological effects of digital games* (pp. 22–38). London: Routledge.

Buzydlowski, J. W. (2015). Co-occurrence analysis as a framework for data mining. *Journal of Technology Research*, *6*, 1–19.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., . . . Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, *214*(1), 325–346.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164.

Griffiths, M. D., Kuss, D. J., Lopez-Fernandez, O., & Pontes, H. M. (2017). Problematic gaming exists and is an example of disordered gaming: Commentary on: Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *Journal of Behavioral Addictions*, *6*(3), 296–301.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Groshek, J., & Tandoc, E. (2017). The affordance effect: Gatekeeping and (non)reciprocal journalism on Twitter. *Computers in Human Behavior*, *66*, 201–210.

Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Cambridge, MA: MIT Press.

Hadgu, A. T., Garimella, K., & Weber, I. (2013). Political hashtag hijacking in the U.S. In D. Schwabe, V. Almeida, H. Glaser, R. Baeza-Yates, & S. Moon (Eds.), *Proceedings of the 22nd international conference on world wide web* (pp. 55–56). New York, NY: Association for Computing Machinery.

Halpern, D., & Gibbs, J. (2013). Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior*, *29*(3), 1159–1168.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. *Language*, *33*(3), 283–340.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In S. Shin & J. Maldonado (Eds.), *Proceedings of the 28th annual ACM symposium on applied computing* (pp. 703–710). New York, NY: Association for Computing Machinery.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In K. Won, R. Kohavi, J. Gehrke, & W. DuMouchel (Eds.), *Proceedings of the 2004 ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177). New York, NY: Association for Computing Machinery.

Jipeng, Q., Zhenyu, Q., Yun, L., Yunhao, Y., & Xindong, W. (2019, April 13). Short text topic modeling techniques, applications, and performance: A survey. *Cornell University*. Retrieved from http://arxiv.org/abs/1904.07695

Jung, C. W. (2019). Media discourse and perception of game regulatory issues. *The Communication Review*, *22*(2), 139–161.

Kirkpatrick, G. (2016). Making games normal: Computer gaming discourse in the 1980s. *New Media & Society*, *18*(8), 1439–1454.

Kleinman, Z. (2013, January 7). Gamers hired by father to 'kill' son in online games. *BBC*. Retrieved from https://www.bbc.com/news/technology-20931304

Levina, N., & Arriaga, M. (2014). Distinction and status

production on user-generated content platforms: Using Bourdieu's theory of cultural production to understand social dynamics in online fields. *Information Systems Research*, *25*(3), 468–488.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2/3), 93–118.

Markey, P. M., & Ferguson, C. J. (2017). Internet gaming addiction: Disorder or moral panic? *American Journal of Psychiatry*, *174*(3), 195–196.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

McCrum, K. (2015, November 3). Tragic teen gamer dies after 'playing computer for 22 days in a row.' *Mirror*. Retrieved from https://www.mirror.co.uk/news/world-news/tragic-teen-gamer-dies-after-6373887

Meerloo, J. A. M. (1954). Television addiction and reactive apathy. *The Journal of Nervous and Mental Disease*, *120*(3), 290–291.

Paul, I., Khattar, A., Kumaraguru, P., Gupta, M., & Chopra, S. (2019). Elites tweet? Characterizing the Twitter verified user network. *Proceedings of the 2019 IEEE 35th international conference on data engineering workshops* (pp. 278–285). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Paulussen, S., & Harder, R. A. (2014). Social media references in newspapers: Facebook, Twitter and YouTube as sources in newspaper journalism. *Journalism Practice*, *8*(5), 542–551.

Pervin, N., Phan, T. Q., Datta, A., Takeda, H., & Toriumi, F. (2015). Hashtag popularity on twitter: Analyzing co-occurrence of multiple hashtags. In G. Meiselwitz (Ed.), *Social computing and social media* (pp. 169–182). Cham: Springer.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40.

Rumpf, H.-J., Achab, S., Billieux, J., Bowden-Jones, H., Carragher, N., Demetrovics, Z., . . . Poznyak, V. (2018). Including gaming disorder in the ICD-11: The need to do so from a clinical and public health perspective: Commentary on: A weak scientific basis for gaming disorder: Let us err on the side of caution (van Rooij et al.). *Journal of Behavioral Addictions*, *7*(3), 556–561.

Schatto-Eckrodt, T., Janzik, R., Reer, F., Boberg, S., & Quandt, T. (2020). Supplementary material to "A computational approach to analyzing the Twitter debate on gaming disorder". *OSF Home*. Retrieved from https://osf.io/vzymj

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st annual meeting of the Association for Computational Linguistics* (pp. 24–29). Stroudsburg, PA: Association for Computational Linguistics.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In R. Wainwright & H. Haddad (Eds.), *Proceedings of the 2008 ACM symposium on applied computing* (pp. 1556–1560). New York, NY: Association for Computing Machinery.

Szablewicz, M. (2010). The ill effects of "opium for the spirit": A critical cultural analysis of China's Internet addiction moral panic. *Chinese Journal of Communication*, *3*(4), 453–470.

Thomas, K., Grier, C., Song, D., & Paxson, V. (2011). Suspended accounts in retrospect: An analysis of Twitter spam. In P. Thiran & W. Willinger (Eds.), *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement* (pp. 243–258). New York, NY: Association for Computing Machinery.

Thompson, P. (2011, June 7). 'Sorry' mother jailed for 25 years for allowing her daughter to STARVE to death while she played an online video game. *Daily Mail*. Retrieved from https://www.dailymail.co.uk/news/article-1394903/Rebecca-Colleen-Christie-jailed-25-years-allowing-daughter-Brandi-Wulf-STARVE-death-played-World-Warcraft.html

Twitter. (2020, March 20). About verified accounts. *Twitter Help Center*. Retrieved from https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

VanDam, C., & Tan, P.-N. (2016). Detecting hashtag hijacking from Twitter. In W. Nejdl, W. Hall, P. Parigi, & S. Staab (Eds.), *Proceedings of the 8th ACM conference on web science* (pp. 370–371). New York, NY: Association for Computing Machinery.

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2/3), 81–92.

van den Brink, W. (2017). ICD-11 Gaming Disorder: Needed and just in time or dangerous and much too early? Commentary on: Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *Journal of Behavioral Addictions*, *6*(3), 290–292.

van Rooij, A. J., Ferguson, C. J., Colder Carras, M., Kardefelt-Winther, D., Shi, J., Aarseth, E., . . . Przybylski, A. K. (2018). A weak scientific basis for gaming disorder: Let us err on the side of caution. *Journal of Behavioral Addictions*, *7*(1), 1–9.

Wang, X., Wei, F., Liu, X., Zhou, M., & Zhang, M. (2011). Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In B. Berendt, A. de Vries, W. Fan, C. Macdonald, I. Ounis, & I. Ruthven (Eds.), *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 1031–1040). New York, NY: Association for Computing Machinery.

Whitton, N., & Maclure, M. (2017). Video game discourses and implications for game-based education. *Discourse: Studies in the Cultural Politics of Education*, *38*(4), 561–572.

Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. *Pew Research Center*. Retrieved from https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users

World Health Organization. (2019). 6C51 Gaming disorder. *International Classification of Diseases 11th Revision*. Retrieved from https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1448597234

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In D. Schwabe, V. Almeida, H. Glaser, R. Baeza-Yates, & S. Moon (Eds.), *Proceedings of the 22nd international conference on world wide web* (pp. 1445–1456). New York, NY: Association for Computing Machinery.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining lexicon-based and learning-based methods for twitter sentiment analysis* [Technical report HPL-2011-89]. Palo Alto, CA: HP Laboratories.

About the Authors

Tim Schatto-Eckrodt is a Research Associate at the Department of Communication at the University of Münster, Germany. He holds an MA degree in Communication Studies from the same department. He is currently working on his PhD project about online conspiracy theories and is part of the junior research group 'Democratic Resilience in Times of Online Propaganda, Fake News, Fear and Hate Speech (DemoRESILdigital).' His further research interests include computational methods and online propaganda.

Robin Janzik is a Research Associate in the Department of Communication at the University of Münster, Germany. He holds an MA degree in Communication Studies from the University of Münster, Germany, and is currently working on his PhD project about the acceptance of virtual reality technology for private use. His further research interests include trust in technology and problematic media use.

Felix Reer is a Postdoctoral Researcher in the Department of Communication at the University of Münster, Germany. His research interests include social media and online communication, effects of digital games, and the use of highly immersive media technologies, such as virtual reality and augmented reality devices. He is Vice-Chair of the Digital Games Research Section of the European Communication Research and Education Association (ECREA).

Svenja Boberg is a Communication Scientist at the University of Münster, who researches the dynamics of online debates and social media hypes using computational methods. In her dissertation she studies the spread of outrage in social media networks. In 2016 she joined the BMBF-funded project 'Detection, Proof and Combating of Covert Propaganda Attacks via Online Media.' Beforehand, she completed her MA in Communication Studies at the University of Münster addressing news consumption on Facebook.

Thorsten Quandt is a Professor of Online Communication at the University of Münster, Germany. His research fields include online communication, digital games, VR and journalism. Quandt is particularly interested in the societal changes connected to the Internet and new media, and the question how human beings have evolved in sync with these changes. His earlier works on participatory journalism and online newsroom production have been widely cited in the field of (digital) journalism research.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 219–230
DOI: 10.17645/mac.v8i3.3007

Article

Exploring the Effect of In-Game Purchases on Mobile Game Use with Smartphone Trace Data

Kristof Boghe [1,*], Laura Herrewijn [2], Frederik De Grove [1], Kyle Van Gaeveren [1] and Lieven De Marez [1]

[1] imec-mict-UGent, Ghent University, 9000 Ghent, Belgium; E-Mails: kboghe@gmail.com (K.B.), frederik.degrove@gmail.com (F.D.G), kyle.vangaeveren@ugent.be (K.V.G.), lieven.demarez@ugent.be (L.D.M.)
[2] Center for Persuasive Communication, Ghent University, 9000 Ghent, Belgium; E-Mail: laura.herrewijn@ugent.be

* Corresponding author

Submitted: 11 March 2020 | Accepted: 9 July 2020 | Published: 13 August 2020

Abstract
Microtransactions have become an integral part of the digital game industry. This has spurred researchers to explore the effects of this monetization strategy on players' game enjoyment and intention to continue using the game. Hitherto, these relationships were exclusively investigated using cross-sectional survey designs. However, self-report measures tend to be only mildly correlated with actual media consumption. Moreover, cross-sectional designs do not allow for a detailed investigation into the temporal dimension of these associations. To address these issues, the current study leverages smartphone trace data to explore the longitudinal effect of in-game purchase behavior on continual mobile game use. In total, approximately 100,000 hours of mobile game activity among 6,340 subjects were analyzed. A Cox regression with time-dependent covariates was performed to examine whether performing in-game purchases affects the risk of players removing the game app from their repertoire. Results show that making an in-game purchase decreases this risk initially, prolonging the survival time of the mobile gaming app. However, this effect significantly changes over time. After the first three weeks, a reversal effect is found where previous in-game purchase behavior negatively affects the further survival of the game. Thus, mobile games without previous monetary investment are more prone to long-term continual game use if they survive the first initial weeks. Methodological and theoretical implications are discussed. As such, the current study adds to those studies that use computational methods within a traditional inferential framework to aid theory-driven inquiries.

Keywords
computational methods; continual game use; in-game purchases; monetization; smartphone trace data; survival analysis

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

The mobile gaming market is expected to reach $77,2 billion in 2020, accounting for 48% of the global gaming revenue and seeing a year-on-year growth rate of 13.3% (Wijman, 2020). A contributing factor in this success is the increasing reliance on monetization through the use of microtransactions. Microtransactions refer to in-game purchases of additional downloadable content (DLC; e.g., adding new game modes) and/or virtual goods that can enhance the player's experience in the game (e.g., items that give players a winning edge in the game; Alha, Koskinen, Paavilainen, Hamari, & Kinnunen, 2014; Luton, 2013). The monetization strategy is being lucratively applied across game genres and platforms, and within both free-to-play games (e.g., Pokémon Go) and games that also involve an initial purchase price (e.g., Minecraft).

As a consequence, academic research has started to turn its attention to the study of in-game purchase be-

havior in recent years. In particular, research has started to investigate in-game purchase behavior in free-to-play games (both on mobile and on other platforms) by examining attitudes towards the microtransaction model (e.g., Alha et al., 2014; Hamari, 2015), which players decide to make in-game purchases and why (e.g., Balakrishnan & Griffiths, 2018; Hamari, 2015; Hamari, Alha, Järvelä, Kivikangas, & Koivisto, 2017; Hamari & Keronen, 2017; Hsiao & Chen, 2016), and what the relationship is between game enjoyment, the making of in-game purchases, and the intention to continue playing the game (e.g., Hamari, 2015; Hamari & Keronen, 2017; Hsiao & Chen, 2016).

Consistently, these studies have underlined the fundamental impact that the microtransaction model has on the game design philosophy. Specifically, when game developers decide to implement the use of microtransactions, they have to find ways to justify and create value for the virtual goods that they offer in order to motivate players to make in-game purchases as frequently as possible (Hamari, 2015; Hamari et al., 2017; Hamari & Keronen, 2017; Hamari & Lehdonvirta, 2010). Game developers themselves indicate that this balancing act has been a matter of great difficulty, since players have no reason to pay money for virtual goods if they are already having a great time with the core game (Alha et al., 2014). As such, artificial barriers are often integrated into the gameplay that make the game mechanics bothersome (e.g., by limiting the amount of lives, resources, or time that players have to play the game every day), after which virtual goods are offered that can break them down (Hamari, 2015; Hamari et al., 2017; Hamari & Keronen, 2017; Hamari & Lehdonvirta, 2010).

This suggests that in-game purchase decisions of players are no longer only influenced by their existing general attitudes and consumption motivations, but also by the developers' design decisions; and that in-game purchase behavior may in turn affect whether a person will continue playing a game (Hamari, 2015; Hamari et al., 2017; Hamari & Keronen, 2017; Hamari & Lehdonvirta, 2010; Hsiao & Chen, 2016). Indeed, the findings of prior research have shown that in-game purchase decisions are highly motivated by the appeal of 'unobstructed play' (i.e., circumventing the barriers that make the gameplay inconvenient), in addition to purchases being motivated by factors such as 'social interaction' (e.g., gift giving), 'economic rationale' (e.g., capitalizing on good deals) and 'unlocking content' (i.e., DLC; Hamari et al., 2017; Hsiao & Chen, 2016). Moreover, studies have found that there is a negative association between enjoyment from playing the game and in-game purchase intention (Hamari, 2015; Hamari & Keronen, 2017), and at the same time, a positive association between in-game purchase intention and continued playing intention (Hamari, 2015; Hamari & Keronen, 2017; Hsiao & Chen, 2016). Both results seem to imply that obstructing players' gameplay (generating frustration) may result in increased in-game purchasing (alleviating the barriers for enjoyment), which may in turn lead to continual game use, at least in the short term.

It is important to note, though, that prior research has measured the association between in-game purchases and players' intention for continual game use exclusively with cross-sectional survey data. Self-report measures of media use, however, are notoriously poor proxies for actual consumption, showing only moderate to low correlations with measures obtained from log data or experience sampling (Araujo, Wonneberger, Neijens, & de Vreese, 2017; Boase & Ling, 2013; Ellis, Davidson, Shaw, & Geyer, 2019; Scharkow, 2016). Biases in retrospective measurements might be especially prevalent for heavily fragmented and short media consumption patterns such as playing mobile games or making in-game purchases (Naab, Karnowski, & Schlütz, 2018). The current study aims to remedy this limitation by analyzing smartphone trace data to shed light on players' actual purchase and gaming behavior. Thus, the following research question is posed:

> RQ1: What is the relationship between performing in-game purchases and continual mobile game use?

In addition to increasing the validity of behavioral measures as such, leveraging log data also enables the modeler to discern granular and temporal patterns which are indistinguishable using cross-sectional survey data. More specifically, this approach makes it possible to establish whether the association between in-game purchases and continual game use might change over time. This seems especially relevant, as results from prior research making use of interview and survey research additionally suggest that implementing microtransactions in games might constitute a double-edged sword in the long term. Notwithstanding the initial proposed positive effects, players seem to argue that having to buy in-game goods with real money in order to be able to continue playing the game the way they want to weakens the game experience in the long run (Alha et al., 2014; Hamari, 2015; Hamari et al., 2017). Furthermore, being able to buy virtual goods that give the owner certain advantages in the game is also believed to skew the competition with other players, potentially resulting in unbalanced gameplay and some games getting called 'pay-to-win' (Alha et al., 2014; Hamari, 2015; Hamari et al., 2017; Hamari & Lehdonvirta, 2010). These previous studies therefore imply that these negative experiences might lead to the formation of negative player attitudes towards the microtransaction model over time (Alha et al., 2014; Hamari, 2015; Hamari et al., 2017). Since attitudes are positively associated with continued playing intention (Hamari, 2015; Hamari & Keronen, 2017), this may eventually result in an increased risk of abandoning the game. No research has actually investigated the existence of such a negative effect over time, however. Therefore, the following research question is posed:

RQ2: How does the relationship between performing mobile in-game purchases and continual game use change over time?

Notably, we frame this research as an exploratory inquiry, serving as a potential starting point for future researchers in sketching out the psychological mechanisms that could explain our findings. In this sense, our data-driven angle is an exemplar of what Margolin (2019) calls the symbiotic approach between computationally-intensive observational studies and more traditional methods within communication sciences. Within this framework, computational approaches can serve as a fertile ground for generating new hypotheses based on the observation of behavioral patterns. This is one of the key advantages of what Margolin (2019) calls "the computational niche." Thus far, methodological limitations have steered the field into inquiries that neglect the temporal nature of mobile gaming, i.e., the fact that purchase behavior should be situated within a game's total 'life span' and as part of an individual gamer's repertoire. It is exactly this temporal dynamic that could yield new valuable hypotheses. In this sense, we hope to stimulate other researchers to disentangle said causal mechanisms within carefully-controlled research settings (e.g., experiments).

2. Methods

We had access to a database containing log data from a free-to-install Android application (Mobile DNA) that gives consumers insight into their own smartphone behavior. The software and the data it generates is proprietary of our institution and is exclusively designed to serve academic research. Users agreed to potential use of their data for this purpose by the authors' research unit before opting in. No other organization or individual outside the research group has access to the database. A subset of our subjects was recruited by our organization, although most users installed the app on their own initiative thanks to a considerable amount of media exposure (e.g., the app has been featured on popular current affairs programs on national television). Subjects received no incentive for installing the application and were free to uninstall it at any time. For our purposes, we extracted all application-data from 01/01/2018 until 02/09/2019. These logs contain information on the specific app used by the subject, including start and end timestamps (precision: 1 millisecond). Moreover, all logs include an anonymized subject-key. Although the application collects geospatial data, these variables were not requested from the database. For this reason, the extracted data contain no identifiable information on our subjects whatsoever.

2.1. Sample

The database holds information on 14,426 subjects, totaling 202 million logs. Of special interest here is that the archive contains 287,789 hours (4,4 million logs) of mobile game use. 9,039 subjects opened at least one game during the data collection period and were included in our initial sample. To identify faulty data logs caused by an early software bug, we identified subjects that supposedly had spent more than 24 hours in a single day on their smartphone. As a result, six individuals were removed from the sample. Next, we accounted for non-human activity (e.g., test devices, bots) in our dataset by inspecting ids that appear in the 99th percentile on both of the following variables: median time spent on smartphone in a single day (median = 2.35 hours, 99th percentile = 7.38 hours) and median duration of a single smartphone session (i.e., opening and closing of smartphone; median = 53 seconds, 99th percentile = 10.21 minutes). Although a subject could legitimately obtain extreme scores on these indicators when the pool of observations is relatively small (e.g., the subject only logged his activity for two days), we would expect a reversion to the median when the number of days under study increases. For this reason, to be eligible for deletion subjects had to be included in the dataset for at least seven days. We deleted eight potential non-human subjects due to this procedure. The density functions of the abovementioned criteria can be found in the Supplementary File (see Figure A1 in the Supplementary File).

2.2. Defining Relevant Survival Periods

Crucially, from a conceptual viewpoint, we are not primarily interested in the total time spent on a mobile game. Instead, the so-called survival time is of key interest. The goal here is to make abstraction of the intensity of game use and rather capture consistent gaming behavior or how long an app remains in one's repertoire. However, unlike available approximations of app survival time in the literature (e.g., Jung, Baek, & Lee, 2012), we argue that the time interval between the first and last day of app usage is a poor approximation of how long an app actually remained in a subject's repertoire. In many cases, only a subset of the total time interval is relevant for measuring actual user interaction due to the many and long pauses in individual gaming behavior. To illustrate this, consider the fact that in our 2018 sample 71% of all days between the first and last active date of a specific game are dates without any gaming activity. At the same time though, it is unreasonable to define continued game usage as a completely non-interrupted streak of gaming activity. Thus, the challenge here is to define what could be considered a maximal allowed tolerance (in days) or time gap between play days.

For this purpose, we formulate an elementary gain metric. The gain metric aims to balance two co-occurring tendencies when one increases this so-called tolerance: Although the amount of captured play days invariably increases with more liberal windows, the interval will include—proportionally—more and more non-play days

as well. The metric aims to establish at which tolerance-level the increase in non-play days far outweighs the gain in captured play days. A specific, though fictitious, example of how different tolerant levels impact the survival time of an app can be found in the Supplementary File (see Table A1 in the Supplementary File).

The gain metric employs the survival time defined by $\text{tolerance}_{(1)}$ as the baseline, which is the most strict tolerance level (i.e., no pauses are allowed) and compares it with the play days/total days ratio captured by applying a more liberal tolerance level. It is thus defined as follows:

$$\text{Gain}_{tx} = \frac{\text{play days}_{tx} - \text{play days}_{t1}}{\text{total days}_{tx} - \text{total days}_{t1}}$$

Here, $t(x)$ represents the results obtained by applying tolerance level x, while $t(1)$ represents the number of (play) days captured by the baseline $\text{tolerance}_{t(1)}$. We calculated the gain for each game played by a specific individual when employing a tolerance level between two and 14 days. Only tolerance-levels which succeeded in increasing the total interval were retained (as indicated by the crossed-out tolerances in Table A1 in the Supplementary File). We subsequently calculated the mean gain (and standard error) for each tolerance-level.

Similar to how eigenvalues are used for determining the optimal number of components in principal component analysis (Schönrock-Adema, Heijne-Penninga, Van Hell, & Cohen-Schotanus, 2009), the preferred solution is determined by taking the tolerance-level just before the point of inflection, which seems to be seven days (see Figure 1). In other words, mobile gamers are allowed to take a break for six consecutive days from playing a specific mobile game. If the game remains untouched on day seven, the last play day before this seven (+) days gap is determined as the end date of the gamer's continual app usage. This procedure captured 76% of all play days while reducing the amount of dates without gaming activity with 90%. In total, 228,035 hours (around 80% of all gaming activity) is retained after applying the maximally allowed tolerance level.

2.3. Left Truncation

Apps were left truncated and thus removed for analysis if the mobile game appears during the first seven days of an individual's data collection. This minimizes the risk of including apps that were already part of one's mobile game repertoire before the subject actually started logging. After all, our gain metric reveals that gamers tend to remove a game from their repertoire if they pause their game activity for more than six days. This procedure diminished the amount of gaming hours in the final dataset to around 96,000 hours, a reduction of 58%.

2.4. Extracting In-Game Purchases

We defined in-game purchase processes by employing an algorithm that looks for specific sequences within an individual's logs set. Users who instantaneously (< one second) switched from playing a mobile game to visiting the Google Play Store (for at least four seconds) and subsequently switched back (< one second) to playing the very same mobile game were considered to have purchased an item in-game. This specific log-chain proved to be most predictive for actual Google Play vending processes after investigating log data of in-game purchases performed by the researchers. We aimed for a conservative estimate to limit the inclusion of accidental or other non-purchase related switches to the Play Store, such as clicking on an advertisement by accident. For this reason, we incorporated two additional decision rules within our purchase detection algorithm. First, all first switches to the Play Store (for each game) are ignored by default. Similarly, switches to the Play Store within the initial five

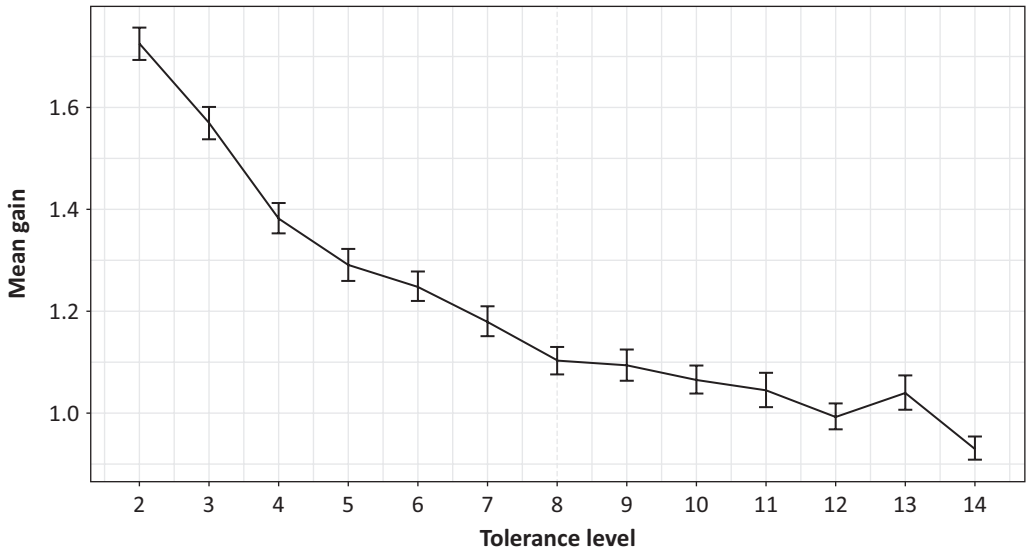


Figure 1. Relation tolerance level and mean gain. Note: Error bars represent standard errors (+/−).

minutes during the first gaming session are disregarded. This allows the user to fall prey to predatory advertising techniques once per game before a vending-chain is categorized as an in-game purchase.

2.5. Scraping Application Data

Since our database as such only includes the application name without any additional info, we developed our own web scraper to obtain relevant metadata. This allowed us to differentiate between mobile games and other apps. In total, we crawled five online app repositories using a sequential scraping method. In order, we scraped the following repositories: Google Play Store, APKMonk, APK Support, APKsHub and APK Pure. If an app was unavailable in the Google Play Store, we opted for the second most reliable online repository, and so on. Next to the general app category (i.e., 'mobile game,' 'social app') and the availability of in-game purchases, we collected other relevant variables which serve as covariates in our model. For more information on the web scraper we refer to Boghe, De Grove, Herrewijn, and De Marez (2020). Since all of these scraped variables are covariates within our model, we removed apps which were unavailable in the Google Play repository from our final dataset to avoid missing values. This led to an exclusion of 3,746 games (7% of all games) from our analysis. However, to check whether this exclusion had an impact on our findings, we ran the model again for all mobile games without the aforementioned covariates.

2.6. Measures

The following descriptive statistics count each and every unique id/application combination as a single mobile game unless mentioned otherwise. Control variables, however, are aggregated on game or id-level, since they do not vary within the same user or app level (e.g., game rating, median time spent gaming for each subject).

2.6.1. Purchases

In total, 46,184 mobile games were included in the sample among 6,340 mobile gamers. 1,082 mobile games included in-game purchases, with a total of 3,884 purchases. Evidently, some games were shared among multiple subjects. When aggregated on game-level, 7,901 unique mobile games were included in the dataset, of which 527 included (a) purchase(s) of at least one subject.

2.6.2. Control Variables

2.6.2.1. Rating and Number of Downloads

Both the average rating of the app on the Google Play Store (1–5) and the number of downloads (ordinal scale from 100–1 billion) were incorporated as proxies of mobile game quality. The assumption here is that highly entertaining mobile games tend to dominate the (free) market. Indeed, popular games tend to receive higher ratings ($r_s = 0.18$, $p < 0.001$). Unsurprisingly, the apps in our dataset tend to be relatively popular (median: one million downloads) and highly rated (median: 4.20, min: 1.30, max: 5.00). For modeling purposes, we aggregated the variable into three categories using the 1st (one million downloads) and 3rd (five million downloads) quartile as cut-offs.

2.6.2.2. Free-to-Play Versus Paid Apps

The differentiation between free-to-play and paid apps is of key interest given our focus on in-game purchases. If in-game purchases have a determining impact on the survival time of a mobile game, this covariate might serve as an important confounding factor. After all, some mobile gamers already made a monetary investment before installing the app in the first place. In our dataset, 6% of all mobile games were paid apps.

2.6.2.3. Availability of Multiplayer Component

Previous research has uncovered that social play is an important motivational factor for continual mobile game use (e.g., Hsiao & Chiou, 2012; Teng & Chen, 2014). In total, 14% of all mobile games in our dataset contain a multiplayer component.

2.6.2.4. Median Game Session Duration

We calculated the median duration of a single game session (in minutes) for each unique application in our dataset. Sessions were defined as the opening and closing of one's smartphone. This metric serves as a proxy of the time investment needed for continual game use. It is not unreasonable to assume that more time-intensive games might exhibit different survival patterns than apps which are more easily appropriated for short game sessions. The median time spent on a mobile game during a single session is 4.10 minutes (min: 0.01, max: 622.67). Three out of four games have a median game session duration of less than 7.60 minutes.

2.6.2.5. Median Time Spent Gaming

Finally, the median time (in minutes) spent gaming in a single day was added for each individual to the model. Avid gamers might exhibit different consumption patterns than sporadic gamers and therefore might show different survival curves. Only days with at least one game log for a specific individual were considered. The distribution is heavily right skewed, with a median time of 16 minutes spent on games in a single day (min: 0.01, max: 625.00). Around one out of four gamers tend to play more than 34 minutes in a single day.

2.6.2.6. Game Repertoire

The survival of an app might be dependent on the current game repertoire of an individual. Specifically, removing an app might be a reasonable strategy to reduce a cluttered game repertoire. We measured the amount of gaming apps opened by a specific individual in a specific day (min: 1, median: 2, max: 32). For days without any gaming activity (12% of all days), we imputed the most recent game repertoire for each individual. On most days (23%), gamers played a single mobile game.

The density functions (for interval variables) or bar plots (for categorical variables) of the abovementioned covariates can be found in the Supplementary File (see Figures A1 and A2 in the Supplementary File).

2.7. Analysis

2.7.1. Time-Dependent Cox Regression

To determine whether in-game purchases have an impact on continual gaming behavior, we performed a Cox regression with time-dependent covariates (Therneau, Crowson, & Atkinson, 2017). The Cox model, unlike traditional regression models, is a survival model which accommodates to censoring techniques necessary for most time-to-event data (Prinja, Gupta, & Verma, 2010; Therneau & Grambsch, 2000). In essence, a survival model estimates the association between covariate scores and the average hazard or failure rate, which constitutes the event rate (e.g., removing an app from the repertoire) at time *t* given that the subject (i.e., mobile game) survived up until time *t*. Censored apps only have an impact on the hazard function until their censored survival time. After that, they are removed from the risk set entirely and therefore have no impact on subsequent hazard estimates. For our purposes, apps were censored (i.e., coded as having survived the period under study) if the last activity log lies within less than seven days (the maximal tolerance-level) from the last log of the subject.

Moreover, the longitudinal variant of the Cox model employed here estimates the average hazard (also called risk score) while correcting for the fact that values of (some) predictors might be correlated with survival time (Suissa, 2008). In our case, the probability of having made an in-game purchase rises steadily as survival time increases. The probability of having made an in-game purchase after ten days of playing a mobile game is a mere 10%, which rises to 26% after 20 days. Since in a cross-sectional survival analysis these purchases are observed only at the end of said life span, any observed relationship between in-game purchases and survival outcomes might be an effect of the variable 'time under study' as such. To avoid this, a key advantage of the longitudinal model is that it compares covariate-scores of apps with a survival time of *t* with the covariate-scores of other apps up until time *t*. In other words, only current values have an impact on the estimated hazard and the model is—on purpose—ignorant of any future state of the subject.

To account for correlated events within-subject, we used the cluster variance by id to estimate a robust standard error (Therneau et al., 2017). Since more than half (51%) of all video games were not shared among more than three subjects, we did not account for the shared variance in survival times within the same game-cluster. However, we reran our analyses while excluding games shared among more than three subjects to check for the potential influence of specific survival times on game-level (see Section 3.3).

2.7.2. Proposed Model

We ran four blockwise Cox regressions with time-dependent covariates. The first model aims to establish the relationship between having performed at least one in-game purchase up until time *t*, before adding other relevant app characteristics (Model 2), proxies of app quality (Model 3) and player characteristics (Model 4). Importantly, the purchase variable could take on three different values depending on the specific game played and the actions undertaken by the user up until time *t*. More specifically, the user could play a game where: a) microtransactions were simply unavailable; b) microtransactions were available, but no in-game purchases were made up until time *t*; or c) microtransactions were available, and the user has made at least one in-game purchase in the past. The time-dependent covariates in our model allowed the user to switch from category b) (purchase available but not yet performed) to c) (purchase performed) on any specific day. We entered category b) as the reference category in the model. Moreover, we entered the median game session duration (Model 2), app rating (Model 3), and both covariates on id-level (Model 4) on a logarithmic scale due to its heavily skewed distribution.

Data cleaning and the calculation of general summary statistics for each user was done in Python using the Dask library as it supports parallel processing. The actual modelling was performed in R using the survival package (Therneau, 2020).

3. Results

3.1. Effect of Purchases on Survival Time

Performing an in-game purchase significantly decreases the risk estimate for app deletion (hazard ratio [HR]: 0.72; 95% confidence interval [CI]: 0.68–0.77). More specifically, apps where in-game purchases are available but none have been made up until time *t* experience a 39% increase in risk of app removal. Interestingly, games without any available in-game purchases experience a 1.22 fold increase in risk (HR: 1.22; 95% CI: 1.17–1.26) when compared to games where in-game purchases are available but not yet performed. Thus, the mere availability

of in-game purchases decreases the likelihood of app removal on time *t*, even without having actually performed said purchase(s). However, actually performing at least one in-game purchase further decreases the risk. Notably, these relationships remain relatively robust regardless of the control variables added to the model.

When it comes to app characteristics, there is no significant association between playing a pay-to-play game and the estimated hazard (HR: 1.00; 95% CI: 0.93–1.08). Contrary to this, playing a game with a multiplayer component does decrease the risk score slightly but significantly (HR: 0.96; 95% CI: 0.93–0.98). The relationship between the median duration of a single play session and app survival is less clear-cut. Although time-intensive games initially seem to have a lower risk for app removal, the final model suggests a small but significant increase in the hazard ratio as the median time spent per session increases (HR: 1.04, 95% CI: 1.02–1.05). Both proxies for app quality—rating and the amount of downloads—have a negative association with the risk estimate. First, for each log unit increase in game rating, the hazard decreases with 11% (HR: 0.89, 95% CI: 0.86–0.92). Second, games between 0.5 and ten million (HR: 0.93, 95% CI: 0.90–0.97) and more than ten million downloads (HR: 0.82, 95% CI: 0.79–0.86) experience a lower risk of app removal when compared with less popular (< 0.5 million downloads) games. Finally, our model suggests that gamers who tend to spend more time on mobile games in a single day retain apps for a longer time period (HR: 0.79, 95% CI: 0.78–0.81). At the same time, though, apps are more likely to be deleted as an individual's game repertoire on time *t* increases (HR: 1.70, 95% CI: 1.65–1.75). Detailed parameter estimates for all four models can be found in Table 1.

3.2. Time-Dependent Coefficients

Finally, the presence of time-dependent coefficients was explored by calculating the Spearman rank correlation between survival time and scaled Schoenfeld's residuals. This analysis shows that the parameter of 'in-game purchases performed' changes significantly over time ($\chi^2 = 27.30$, $p < 0.001$).

Therefore, we reran the fourth model containing a step function for $\beta(t)$ divided over multiple time periods (Zhang, Reinikainen, Adeleke, Pieterse, & Groothuis-Oudshoorn, 2018). The direction of the relationship seems to reverse after playing a game for three weeks (see Figure 2), which served as a cut-off point for our new model. The step function shows a clear reversal effect. While the risk for app removal decreases for games with an in-game purchase during the first 21 days (HR: 0.64; 95% CI: 0.59–0.70), in-game purchases are associated with a 38% increased risk after said time period (HR: 1.38; 95% CI: 1.19–1.59). Thus, while at first glance the model suggests that performing (an) in-game purchase(s) stimulates continual game use, it is exactly this type of previous monetary investment that increases the risk of app removal later in a game's life span.

3.3. Model Robustness

We validated our model by checking for influential observations, outliers, and non-normality. Based on a cross-sectional Cox model, we calculated DfBetas (DfBeta/SE) for each observation. Following recommendations by Belsley, Kuh, and Welsch (2004), values larger than two should raise our attention. There were no influential observations present in our data, with a maximal DfBeta

Table 1. Estimated parameters of Cox regression with time-dependent covariates.

Variable	Model 1 HR (95% CI)	Model 2 HR (95% CI)	Model 3 HR (95% CI)	Model 4 HR (95% CI)
purchases: [a]				
purchase unavailable	1.34 (1.30–1.38) ***	1.31 (1.27–1.36) ***	1.22 (1.18–1.26) ***	1.22 (1.17–1.26) ***
purchase performed	0.68 (0.64–0.73) ***	0.68 (0.63–0.73) ***	0.69 (0.64–0.73) ***	0.72 (0.68–0.77) ***
paid app		1.00 (0.93–1.08)	0.97 (0.90–1.05)	1.00 (0.93–1.08)
multiplayer		0.89 (0.87–0.92) ***	0.91 (0.89–0.94) ***	0.96 (0.93–0.98) **
median session game		0.98 (0.97–0.99) **	1.00 (0.98–1.01)	1.04 (1.02–1.05) ***
rating			0.83 (0.80–0.86) ***	0.89 (0.86–0.92) ***
downloads: [b]				
0.5–ten million			0.97 (0.94–1.01)	0.93 (0.90–0.97) **
+ ten million			0.87 (0.83–0.91) ***	0.82 (0.79–0.86) ***
time spent on games				0.79 (0.78–0.81) ***
game repertoire				1.70 (1.65–1.75) ***
Wald-score (*p*)	443 ($p < 0.001$)	518 ($p < 0.001$)	751 ($p < 0.001$)	1909 ($p < 0.001$)

Notes: [a] reference category ('purchases available, but not performed'); [b] reference category ('less than or equal to 0.5 million downloads'). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

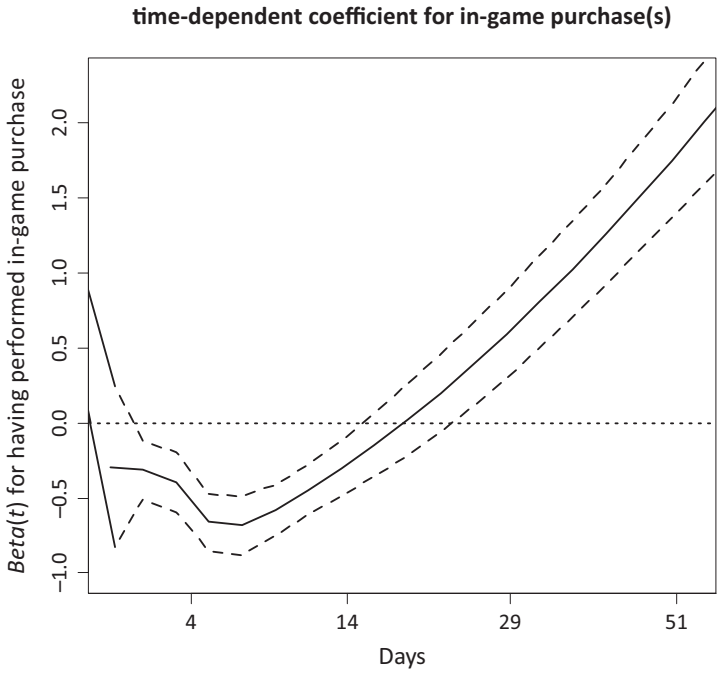


Figure 2. Estimated *Beta*(*t*) (*df* = 8) for having performed at least one in-game purchase up until time *t* over time (log[days]).

of 0.2 (see Figure A3 in the Supplementary File). Similarly, an inspection of the deviance residuals of the longitudinal Cox model show no clear deviations from normality nor any indication of outliers, with residuals centered around zero (see Figure A4 in the Supplementary File). To check whether specific mobile games shared by multiple users had an undue influence on our parameter estimates, we reran our analyses twice; once with a corrected standard error estimate by game-variance (instead of id-variance) and once by only including games that were shared by at most three subjects. Both alterations to our model had no considerable impact on the parameter estimates nor on the time-dependent effects previously reported (see Table A2 and Figure A4 in the Supplementary File). The time-dependent coefficient for the second model mentioned here shows greater uncertainty, which is to be expected given the smaller sample size. Some app characteristics such as number of downloads are less predictive, which can be explained by the fact that we are pooling less popular games here. Next, we inspected for general model robustness by: (a) running the model with a less conservative estimate of the amount of in-game purchases by categorizing the first vending process as an in-game purchase; and (b) running the model while including mobile games which were unavailable on Google Play (excluding the game-level covariates). These model adjustments had no impact whatsoever on the trends already reported (see Table A2 and Figure A4 in the Supplementary File). As one could expect, the time-dependent coefficient for the less conservative estimate of purchase detection shows a less strong effect when compared with the more stringent definition, but the reversal effect is still clearly present.

4. Discussion and Conclusion

This study contributes to the field of media entertainment studies (and game studies more specifically) on multiple grounds. First, on a methodological level, this is to our knowledge the first study that applies log data to explore time-dependent relations between actual in-game purchase behavior and continual mobile game use. While there have been a couple of investigations into the impact of several app characteristics on survival time (e.g., Lee & Raghu, 2014), no study has specifically sought to establish a clear time-dependent relation between a particular behavioral antecedent and consequence. Moreover, we developed a detection algorithm to discern in-game purchases in log data. We urge future studies to examine the validity and robustness of this formulation of vending processes. Available validated algorithms might stimulate the field to sketch out the effect of in-game purchases in a more nuanced fashion, incorporating the longitudinal logic implied here. Finally, our proposed tolerance metric might prove to be a valuable tool for communication scholars in general. We believe that most—if not all—measures of mobile media consumption stemming from behavioral log data could profit from our tolerance-approach. Mobile media tend to be consumed in extremely short bursts and are prone to habitual activation (Oulasvirta, Rattenbury, Ma, & Raita, 2012), which makes the demarcation of detailed but relevant periods of media consumption a new challenge for the communication scientist.

Second and more fundamentally, on a theoretical level the study is the first to explore the relationship between in-game purchase behavior and continual mobile game use (RQ1), as well as how this relationship

changes over time (RQ2). Regarding our first research question, the results reveal a positive relationship between making in-game purchases and continual mobile game use. Specifically, compared to games where at least one in-game purchase is performed, games that allow microtransactions to be made but where players have not (yet) performed in-game purchases experience a 39% increase in risk of being removed from players' app repertoires. Moreover, and interestingly, games that do not make use of microtransactions and where players thus have no opportunity to make purchases see a 22% increase in risk when compared with games where purchases are available but not (yet) performed. In other words, the mere availability of microtransactions in mobile games decreases their risk of removal from players' app repertoires, and when players perform at least one in-game purchase, this risk of removal decreases even further.

There are two important notes to be made about the unveiled association. First, it is noteworthy that the mere availability of microtransactions in games decreases the risk of removal as well, when compared to games where no microtransactions are available. This might be explained by some underlying characteristic(s) shared among these games. For example, game developers who make use of this monetization strategy generally have access to a more steady stream of revenue and might therefore offer a higher initial production value, as well as updates over time (e.g., balancing their games regularly) and recurrent releases of new content (e.g., releasing new chapters of the story, character customization options), which might cause players to keep the game in their repertoire for a longer period of time. A second note pertains to the control variables included in the model. Interestingly, we found no significant association between a game being either free-to-play or pay-to-play and the risk estimates. This suggests that the behavioral effects of these two different revenue models on mobile gamers are distinct.

When it comes to our second research question, analyses show that the effect of in-game purchase behavior on continual mobile game use changes significantly over time. Specifically, a clear reversal effect is found: While the risk for removal decreases with 36% for mobile games with a performed in-game purchase during the first three weeks, it is associated with a 38% increase in risk after said time period. Results thus show that prior in-game purchase behavior has a negative impact on continual mobile game use later on in a game's life span. Mobile games in which microtransactions are available but where players have not made any in-game purchases, on the other hand, are more prone to long-term survival if they make it through the first few weeks.

At first glance, these results seem to be in line with the expectations posed previously: Performing in-game purchases might prolong the survival time of a mobile game initially (e.g., by taking away obstructions that have been artificially implemented in the gameplay, as well as the player frustration that results from this), but after a while this effect may turn sour (e.g., because the game will keep introducing barriers and looking for monetary investments, weakening the game experience and resulting in players becoming discontented). Importantly, since our results consistently show that the estimated risk score of apps with actual performed in-game purchases differs significantly from the risk score of apps where purchases are available but not (yet) performed, the reversal effect cannot be sufficiently explained by the frustrating experiences and negative player attitudes that might potentially result from the built-in barriers in microtransaction games. In both instances, players may be confronted with barriers that invite them to make an in-game purchase, but it is only after actually performing said purchase(s) that a differential survival curve and the reversal effect take place.

Given the exploratory nature of our inquiry, the question remains how one should interpret these findings. We call for future research to disentangle the causal mechanism behind this reversal effect. Nonetheless, we wish to give several pointers here. A potential explanation for why the act of performing in-game purchases triggers these differential patterns in continual mobile game use can be found in Self-Determination Theory (SDT; Ryan & Deci, 2000). Motivation is crucial in supporting continual behavior (Teixeira, Carraça, Markland, Silva, & Ryan, 2012), and SDT concerns itself with factors that can facilitate or undermine motivation, both intrinsic and extrinsic (Ryan & Deci, 2000). Within a digital game context, intrinsic motivation, especially, has proven to be essential (Ryan, Rigby, & Przybylski, 2006). Although games can be played for external reasons, such as receiving monetary rewards (e.g., professional gamers), most players are intrinsically motivated to do so: They play games because it is intrinsically satisfying to them (Ryan et al., 2006), because they are seeking enjoyment.

A sub theory within SDT, called Cognitive Evaluation Theory (CET; Deci & Ryan, 2000), addresses events and conditions that can reinforce or impede this intrinsic motivation. Specifically, CET states that factors that enhance a person's psychological needs of autonomy (i.e., a person's need for volition or free will when performing a task) and competence (i.e., a person's need for challenge and feelings of effectiveness) can support intrinsic motivation, while factors that thwart these needs can enfeeble it (Deci & Ryan, 2000; Ryan et al., 2006). Importantly, gaming research has shown that the satisfaction of these needs of autonomy and competence in a game also predicts players' enjoyment (Neys, Jansz, & Tan, 2014). Moreover, both players' intrinsic motivation to play games (Ryan et al., 2006) and their enjoyment of a game (Neys et al., 2014) have been shown to be positive predictors of their future play intentions.

When taking a look at in-game purchase behavior and its relation to continual mobile game use, then, it seems plausible that controlling behavior by implementing artificial barriers could impair players' sense of au-

tonomy and competence. Making a purchase may then lift these barriers and lead to unconstrained or even accelerated gameplay for a moment, which may lead to a temporary burst of enjoyment. This might explain why in-game purchase behavior (versus deciding not to make an in-game purchase) is positively related to continual game use in the short term. However, the decision to purchase may also cause players to feel as if they are forced to perform in-game purchases to be able to keep on playing the game the way they want to (diminishing their feelings of autonomy), and that their progress in the game is not due to their own skills and abilities, but rather because of their monetary investments (weakening their feelings of competence).

Additionally, in-game purchase behavior may also cause a shift in player motivation from the inherently satisfying aspects of a game (i.e., intrinsic motivation: "I play the game because I find it enjoyable") to external, monetary aspects (i.e., extrinsic motivation: "I play the game because I have already invested money in it, and I have to get my money's worth"), which research shows is much shorter-lived (e.g., Teixeira et al., 2012).

As a consequence of these two processes, the act of making purchases in a game may chip away players' intrinsic motivation for continued play one purchase at a time, resulting in a negative association with enjoyment and continual mobile game use in the long run. Players who do not make in-game purchases, on the other hand, may feel the frustrating impact of the microtransaction design initially (curbing their autonomy and intrinsic motivation in the short term), but in the long haul, they may feel that their achievements are their own (resulting in a stronger sense of competence) and that choosing not to make monetary investments allows them to play the game according to their own will (enhancing their autonomy and keeping the focus on the inherently satisfying aspects of the game instead of on its monetary facets). As such, their intrinsic motivation and enjoyment will be more durable, resulting in increased long-term survival for the mobile game(s) they play.

Two constraints, more specifically in data availability and modeling capacity, are worth noting here as they can inform future researchers to expand on these findings. First, the reliance on log data as the sole data source impaired our model significantly. Given the reported behavioral relationships in this research, one promising avenue is to use our findings to develop a testable conceptual model which not only includes behavioral measures, but also sheds light on the psychological antecedents and consequences of the reversal effect reported here. This will allow the field to contextualize the reversal effect and ultimately to formulate some sensible recommendations to strengthen players' game literacy. The combination of trace and survey data will prove to be indispensable in this regard (Stier, Breuer, Siegers, & Thorson, 2019). Second, although we are confident our analysis strategy yields valid and reliable results, the data structure employed here is ideally modeled by techniques that take into account the cross-classified nature of the data. Given the computational limitations of the frailty packages available in R, a Bayesian approach (using Stan, for example) might prove to be valuable to optimize the parameter estimates in future research.

Acknowledgments

We would like to thank the reviewers and the editors for their valuable comments and suggestions. This research has been made possible by the Mobile DNA Project of the imec-mict-UGent Research Group.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Alha, K., Koskinen, E., Paavilainen, J., Hamari, J., & Kinnunen, J. (2014). Free-to-play games: Professionals' perspectives. In *Proceedings of the 22nd International Academic Mindtrek Conference* (pp. 49–58). New York, NY: Association for Computing Machinery. Retrieved from https://dl.acm.org/doi/10.1145/3275116.3275133

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11*(3), 173–190.

Balakrishnan, J., & Griffiths, M. D. (2018). Loyalty towards online games, gaming addiction, and purchase intention towards online mobile in-game features. *Computers in Human Behavior*, *87*, 238–246.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken, NJ: John Wiley & Sons.

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication*, *18*(4), 508–519.

Boghe, K., De Grove, F., Herrewijn, L., & De Marez, L. (2020). *Scraping application data from the web—Addressing the temporality of online repositories when working with trace data*. Extended abstract presented at the 70th International Communication Association Conference.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227–268.

Ellis, D. A., Davidson, B. I., Shaw, H., & Geyer, K. (2019). Do smartphone usage scales predict behavior? *Inter-*

national Journal of Human-Computer Studies, *130*, 86–92.

Hamari, J. (2015). Why do people buy virtual goods? Attitude toward virtual good purchases versus game enjoyment. *International Journal of Information Management*, *35*, 299–308.

Hamari, J., Alha, K., Järvelä, S., Kivikangas, M. J., & Koivisto, J. (2017). Why do players buy in-game content? An empirical study on concrete purchase motivations. *Computers in Human Behavior*, *68*, 538–546.

Hamari, J., & Keronen, L. (2017). Why do people buy virtual goods: A meta-analysis. *Computers in Human Behavior*, *71*, 59–69.

Hamari, J., & Lehdonvirta, V. (2010). Game design as marketing: How game mechanics create demand for virtual goods. *International Journal of Business Science and Applied Management*, *5*(1), 14–29.

Hsiao, K., & Chen, C. (2016). What drives in-app purchase intention for mobile games? An examination of perceived values and loyalty. *Electronic Commerce Research and Applications*, *16*, 18–29.

Hsiao, C., & Chiou, J. (2012). The effects of a player's network centrality on resource accessibility, game enjoyment, and continuance intention: A study on online gaming communities. *Electronic Commerce Research and Application*, *11*, 75–84.

Jung, E., Baek, C., & Lee, J. (2012). Product survival analysis for the App Store. *Marketing Letters*, *23*(4), 929–941.

Lee, G., & Raghu, T. S. (2014). Determinants of mobile apps' success: Evidence from the App Store market. *Journal of Management Information Systems*, *31*(2), 133–170.

Luton, W. (2013). *Free-to-play: Making money from games you give away*. Boston, MA: Addison Wesley.

Margolin, D. B. (2019). Computational contributions: A symbiotic approach to integrating big, observational data studies into the communication field. *Communication Methods and Measures*, *13*(1), 1–19.

Naab, T. K., Karnowski, V., & Schlütz, D. (2018). Reporting mobile social media use: How survey and experience sampling measures differ. *Communication Methods and Measures*, *13*(2), 126–147.

Neys, J. L., Jansz, J., & Tan, E. S. (2014). Exploring persistence in gaming: The role of self-determination and social identity. *Computers in Human Behavior*, *37*, 196–209.

Oulasvirta, A., Rattenbury, T., Ma, L., & Raita, E. (2012). Habits make smartphone use more pervasive. *Personal and Ubiquitous Computing*, *16*(1), 105–114.

Prinja, S., Gupta, N., & Verma, R. (2010). Censoring in clinical trials: Review of survival analysis techniques. *Indian Journal of Community Medicine*, *35*(2), 217–221.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Ryan, R. M., Rigby, S. C., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344–360.

Scharkow, M. (2016). The accuracy of self-reported internet use: A validation study using client log data. *Communication Methods and Measures*, *10*(1), 13–27.

Schönrock-Adema, J., Heijne-Penninga, M., Van Hell, E. A., & Cohen-Schotanus, J. (2009). Necessary steps in factor analysis: Enhancing validation studies of educational instruments. The PHEEM applied to clerks as an example. *Medical Teacher*, *31*(6), 226–232.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/0894439319843669

Suissa, S. (2008). Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, *167*(4), 492–499.

Teixeira, P. J., Carraça, E. V., Markland, D., Silva, M. N., & Ryan, R. M. (2012). Exercise, physical activity, and self-determination theory: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(78). https://doi.org/10.1186/1479-5868-9-78

Teng, C., & Chen, W. (2014). Team participation and online gamer loyalty. *Electronic Commerce Research and Application*, *13*, 24–31.

Therneau T. (2020). A Package for Survival Analysis in R. R package version 3.2-3. [Computer software]. Rochester, MN: Mayo Clinic. Retrieved from https://CRAN.R-project.org/package=survival

Therneau, T. M., Crowson, C., & Atkinson, E. (2017). *Using time dependent covariates and time dependent coefficients in the cox model*. Retrieved from https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

Therneau, T. M., Grambsch, P. M. (2000). *Modeling survival data: Extending the cox model*. New York, NY: Springer.

Wijman, T. (2020). The world's 2.7 billion gamers will spend $159.3 billion on games in 2020; the market will surpass $200 billion by 2023. *Newzoo*. Retrieved from https://newzoo.com/insights/articles/newzoo-games-market-numbers-revenues-and-audience-2020-2023

Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E., & Groothuis-Oudshoorn, C. (2018). Time-varying covariates and coefficients in Cox regression models. *Annals of Translational Medicine*, *6*(7), 121.

About the Authors

Kristof Boghe is a PhD Student at Ghent University. His research interests lie in all things related to the computational turn in the social sciences and digital methods. He is fascinated with how the sheer abundance and availability of digital data can give us new insights on the inner workings of our (offline or online) social world. More specifically, he wants to unravel the implications of doing research in the age of big data on theory formation within communication sciences.

Laura Herrewijn is a Postdoctoral Researcher and Assistant Professor in Digital Marketing at the Department of Communication Sciences at Ghent University. Her research can be found on a crossroads between marketing communication and game studies, with a strong focus on virtual environments (such as those represented in digital games and virtual and augmented reality apps) and their potential for the integration of persuasive communication (e.g., advertising, educational content), along with consumer behavior (e.g., in-app purchases).

Frederik De Grove is an Assistant Professor at the Department of Communication Sciences (Ghent University), where he teaches Digital Methods. He successfully defended his PhD in 2014 and obtained a MA in Statistical Data Analysis in 2019 (Ghent University). In addition, he is Project-Coordinator for the GOA project NewsDNA (2018–2022). The main goal of this interdisciplinary project is to promote news diversity through algorithmic development.

Kyle Van Gaeveren started as a Junior Researcher at MICT in November 2018. He graduated as a Master in Sociology (Ghent University, 2011) and Social Work and Welfare Studies (Ghent University, 2014), and is currently obtaining his BA in Computer Science (Karel de Grote Hogeschool). He joined MICT on the Mobile DNA project with a central focus on analyzing log data of smartphone usage.

Lieven De Marez is a Professor at the Department of Communication Sciences at Ghent University. At the department, he is heading the multidisciplinary research group for Media, Innovation and Communication Technologies (MICT), affiliated to imec, the world-leading R&D and Innovation Hub in Nanoelectronics and Digital Technologies. Within imec, Lieven is the founding father of the Digimeter (http://www.imec-int.com/en/digimeter), a reference monitor on the adoption and use of digital technologies.

COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 231–238
DOI: 10.17645/mac.v8i3.3163

Article

Open-Source's Inspirations for Computational Social Science: Lessons from a Failed Analysis

Nathaniel Poor

Underwood Institute, Cambridge, MA 02139, USA; E-Mail: natpoor@gmail.com

Submitted: 15 April 2020 | Accepted: 7 July 2020 | Published: 13 August 2020

Abstract
The questions we can ask currently, building on decades of research, call for advanced methods and understanding. We now have large, complex data sets that require more than complex statistical analysis to yield human answers. Yet as some researchers have pointed out, we also have challenges, especially in computational social science. In a recent project I faced several such challenges and eventually realized that the relevant issues were familiar to users of free and open-source software. I needed a team with diverse skills and knowledge to tackle methods, theories, and topics. We needed to iterate over the entire project: from the initial theories to the data to the methods to the results. We had to understand how to work when some data was freely available but other data that might benefit the research was not. More broadly, computational social scientists may need creative solutions to slippery problems, such as restrictions imposed by terms of service for sites from which we wish to gather data. Are these terms legal, are they enforced, or do our institutional review boards care? Lastly—perhaps most importantly and dauntingly—we may need to challenge laws relating to digital data and access, although so far this conflict has been rare. Can we succeed as open-source advocates have?

Keywords
computational social science; fandom; games; online community; open source; Reddit

Issue
This article is part of the issue "Computational Approaches to Media Entertainment Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences, Germany), Tim Wulf (LMU Munich, Germany) and M. Rohangis Mohseni (TU Ilmenau, Germany).

1. Introduction

This article uses an autoethnographic approach to explore issues I encountered on a project about a gaming community and its online embodiment. I discuss how lessons from the open-source community could have helped me and, more broadly, can help computational social science (CSS) move past current tensions with application programming interface (API) availability, commercialism, and academic research for the greater social good. My study, part of a larger work, used data on more than 2.25 million posts to the online forum site Reddit, captured over two one-year periods, accessed via an API offered through Pushshift. The hypothesis for that project was that users move from one community to another when a new game in a series comes out (hence the two one-year time periods, related to two game releases and opportunities for community movement). My initial findings, despite a large amount of data and relevant theory, was that users do not in fact move from community to community—a null finding.

This current article explores how this null finding could have happened, even with so much data and good theory, and proposes a way forward for CSS as a whole in light of such occasional, but potentially enlightening, problems.

First, I touch on both open-source software and CSS as ideological in nature, in order to frame the autoethnographic case study of my initially failed research project. Then I discuss my study by explaining the site (Reddit) and the API (Pushshift). I present the topic in terms of the specific game series (Bethesda's Elder Scrolls franchise)

and the theories involving online communities and fans. I then move on to the analysis of why the study did not work and the solutions I arrived at.

These solutions serve as a jumping-off point for the main idea of this article. The solutions for my specific research problems can be contextualized within challenges that researchers have identified for CSS and that must be overcome (e.g., Bruns, 2019; Freelon, 2018; Halavais, 2019). But we can draw lessons from the history and current mature state of the open-source software ecosystem to help us move forward. The open-source world is an environment with a wide mix of commercial and freely available items, third parties who add value to information, teams of people with different expertise, and legal hurdles that have had to be overcome, similar to the CSS world itself.

1.1. Open-Source Software

Open-source software (DiBona, Ockman, & Stone, 1999; Raymond, 1999) is software for which the code (and the software itself) is available for anyone to copy (*libre*) as well as available for free (*gratis*). Sometimes it is referred to as 'free and open-source software,' or FOSS, although opinions differ about the definitions important to those involved in the effort (DiBona et al., 1999). Beyond this, the idea is that the code can be freely modified so that users can fix bugs or customize it as they need. The source code is available; it is open-source. This openness contrasts with commercial software, for which users must pay and for which they cannot access the source code. Unlicensed copies and the hacking of commercial software to get at its code are generally illegal. Note that all of these considerations are ideological, economic, and political. These are not just technologies, but rather they are sociotechnical systems.

Open-source is by now well established and does not garner the attention it once did. Many academics are aware of open-access journals and the open-access licenses governing articles in such journals. These stem from the licensing ideas in open-source software, which utilize copyright law to allow copying with certain requirements (such as attribution) instead of disallowing it. Open-access journals are in some ways the journal equivalent of open-source software. Open-source advocates have carefully reconsidered copyright within the existing legal framework to essentially turn the concept on its head: instead of copyright, there is copyleft. That is, instead of using copyright laws to restrict copying of works, advocates realized that licensing laws could be used to restrict the restriction that would put those works under lock and key.

Open-source, besides a type of legal license or ideological stance, is also a way of working. Many people freely donate time and expertise to work on large, distributed open-source projects. But not all licenses require that the end product always be free. Some allow modifications that users or open-source-related companies can charge for. Other companies, such as O'Reilly, publish guidebooks to open-source software, and they charge for those books just as they would charge for any other book.

People undertaking open-source work often use open-source tools to make more open-source code in turn, such as using open-source text editors and compilers to make open-source computer programs. Some CSS tools, such as Python, R, and some database applications, are open-source.

The software world is a mixed environment, much like the CSS world of accessing data and research. Some source code (or data) is free, other source code is not, and one must pay for access or not access it at all. Some people add value to the data, some people work on data for free, and teams of people with a variety of relevant skills often work on projects.

1.2. CSS as Ideology

CSS, generally, is the large-scale analysis of digital data relating to human behavior (Lazer et al., 2009). The size of the data set needed to achieve the CSS label varies depending on the perspective of the viewer; attainable sizes (in terms of collection, storage, and analysis) have increased over time. Data complexity also may present a challenge to deciding what is and what is not CSS, as may the type of analysis used.

The advent of CSS can, in hindsight, be seen as a sensible sociotechnical response to improvements in several technological areas: local computing power, easier programming languages, greater internet speeds, data access via APIs, and a greater number of people interacting with a greater amount of digital material online. As part of the social sciences, CSS relies on the belief that things can and should be measured, and that those measurements can accurately measure what we are trying to measure (Bulmer, 2001). This is ideological. Some of the technological advances needed for CSS are ideological as well: the arguments that computing languages *should* be easier than, for example, C, or that data *should* be collected and then made available by API. This point should not be overlooked, because sites from which we may want data can have restrictive, vague, and problematic terms of service that perhaps researchers should ignore in light of a greater social good (Fiesler, Beard, & Keegan, 2020). Additionally, when APIs are shut down and data access is curtailed, some CSS studies become impossible or at least difficult to undertake (Bruns, 2019; Freelon, 2018).

When the API is fronting data that has commercial value, as with Facebook and Twitter, or if the data come with privacy concerns, making the data available freely and for free becomes problematic in different ways (Bruns, 2019; Fiesler et al., 2020; Halavais, 2019). Much like with the open-source community in its early days of growing popularity, a tension exists between those who want information to be freely available for a greater

social good and those who want to commercialize it (DiBona et al., 1999; Raymond, 1999). This tension can be approached by understanding commercial restrictions and economic gain for the few, on the one hand, and academic access and research for the greater good, on the other—although both sides share data privacy concerns, albeit for different reasons, working toward different outcomes.

The API, however, was not the problem I encountered in my research's null finding. The data were available, in fact, because one person believed they should be available. Without that API, my project might have been impossible.

2. The Research Project

2.1. Reddit and Pushshift

Reddit is a website that serves as an online space for thousands of different forums, called subreddits, many of which function as online communities (Panek, Hollenbach, Yang, & Rhodes, 2018) and, importantly, as online fan communities (Gunderman, 2020). Some subreddits receive hundreds or thousands of posts per day. Founded in 2005, Reddit is similar to, and draws from, older online bulletin-board systems like Usenet, AOL, Slashdot, and modem-based BBS systems. The people who run Reddit generally take a hands-off approach to site governance, which has led to some problems (Massanari, 2017). But the users make each specific subreddit and determine its rules. Like many online spaces, users may create a username (and thus an identity of sorts) if they wish and may post to whichever subreddits they like, or they might just lurk and read posts without commenting. But if they do post, they create both text (the post) and associated metadata (such as who posted, where, and when). These digital trace data were what I wanted, but Reddit does not make it easily available. Instead, an individual, Jason Baumgartner, has taken it upon himself to collect all of it, billions of posts, and make it available via an API at the website Pushshift.io (Gaffney & Matias, 2018). Note that he does this for free and solicits donations to help the effort.

2.2. The Elder Scrolls, Fans, and Online Gaming Communities

The Elder Scrolls franchise is a series of fantasy adventure computer games reminiscent of *The Hobbit* (Tolkien, 1937), with elves and wizards and magic swords. The initial game in the series, The Elder Scrolls: Arena (Bethesda Softworks, 1994), was released in 1994. The Elder Scrolls V: Skyrim (Bethesda Softworks, 2011) was released in November 2011, was remastered in 2016 for newer game consoles, and was recoded and released for the Nintendo Switch in 2017. In short, Skyrim is extremely popular and has sold millions of copies worldwide. A massively multiplayer game set in the Elder Scrolls universe, The Elder Scrolls Online (Bethesda Softworks, 2014), similar to the better-known World of Warcraft (Blizzard Entertainment, 2004), was released in 2014. The releases of these two games—Skyrim and The Elder Scrolls Online—were the points in time in which I was interested.

Many people who buy Elder Scrolls games are more than just purchasers or players of the game. They are fans, as is true of many people and many cultural products (Fiesler, 2007; Jenkins, 1992). Fans, and people more generally, form communities and online communities; currently many communities have both online and offline components to varying degrees (Poor & Skoric, 2014; Wellman, Boase, & Chen, 2002). Fan communities can be robust and can survive moves from one platform to another (Fiesler, Morrison, & Bruckman, 2016; Pearce, 2009).

Some fans of games like the Elder Scrolls go beyond just buying the game. They participate in actively creating and changing the game worlds, such as by coding modifications, or mods, to those games when possible (usually on the Windows/PC platform). Game modders form their own subculture within gaming fans of a game; and the many necessary interactions among modders can lead to strong community ties (Poor, 2014). More generally, fans are also well known for creating fictional stories about the objects of their fandom, called fanfic (Jenkins, 1992).

It is through communication—for fans, perhaps communication takes the form of fanfic, discussion of how to make a mod, or discussion of the game in general—that humans form community (Carey, 1989; Dewey, 1927; Iyer, Cheng, Brown, & Wang, 2020). This phenomenon is not found just in our online behavior (Kraut & Resnick, 2011). It is a fundamental capability that evolved in us over millions of years (Gamble, Gowlett, & Dunbar, 2014; Tomasello, 2010). Fandom is not solely denoted by communication: Actually buying the objects related to that fandom is an important and a heavily intertwined part of how fan identities are established and maintained (Hills, 2003). To some extent, fans are "ideal consumers [who] automatically buy the latest works" (Cavicchi, 1998, p. 62).

Altogether, being a fan of one Elder Scrolls game or of the franchise overall, buying the new game in the franchise upon its release, and then discussing it with other fans on Reddit seems like a sensible path to many fans of the series based on the above-mentioned theory. That was my main hypothesis in my failed study.

2.3. Data and Results

Using Python, I scraped the third-party Reddit API at Pushshift, ending up with data for more than 2.25 million posts spanning two full years. For the first one-year period when Skyrim (Bethesda Softworks, 2011) was released, I obtained data on 979,582 posts: who posted, when, and to which of the several game-related sub-

reddits. To study correlations in posting behavior, I winnowed it down to the three months before and the three months after the new game was released, using data on 772,873 posts. For the second new game (The Elder Scrolls Online [Bethesda Softworks, 2014] superseded Skyrim as the newest game in the franchise), I scraped data on 1,296,146 posts, which I similarly winnowed down to data on 861,040 posts spanning the three months before and the three months after its release.

I ran correlations between number of posts on one subreddit during the three months before a game release and number of posts on another subreddit during the three months after the same game release, per user. This captured before-release and after-release posting behavior. The first correlation, from The Elder Scrolls IV: Oblivion (Bethesda Softworks, 2006) to The Elder Scrolls V: Skyrim (Bethesda Softworks, 2011), was 0.16. The second, from Skyrim to The Elder Scrolls Online (Bethesda Softworks, 2014), was 0.04. Clearly, this null result was headed for the file drawer (Rosenthal, 1979). How do you get a null result with 2.25 million data objects and good theory?

2.4. Problems and Solutions

Perhaps I did not have quite the right data. Maybe I needed more specific sales data or player data, which exist but cannot be accessed. Maybe I should have included the text of the posts, not just the metadata, and performed textual analysis to gauge the sentiment of users toward the new game, or else to determine where their level of devotion lay. Were they fans of the old game specifically, or of the series generally? Possibly I had aggregated data at an analytically unsound level, encountering a Simpson's paradox as can befall such work on Reddit (Lerman, 2018).

Perhaps I should have added surveys to the digital data, as some researchers have suggested (Stier, Breuer, Siegers, & Thorson, 2019). Perhaps I was too narrow in my focus, looking only at the fans or users as they participated on Reddit but missing their activity on other platforms (Menchen-Trevino, 2013). Researchers have noted how Skyrim (Bethesda Softworks, 2011) fans (Puente & Tosca, 2013) and people in general (Baym, 2007) use more than one online space for their online activity.

Possibly my analysis was not very good. Initially I had wanted a path-analysis model (given how people use multiple subreddits over time); but after consultation with a colleague I went with much simpler correlations. Maybe I had the wrong theory—perhaps consumerism was not the driving force here (indeed, community attachment turned out to be the story). How could I have avoided this null finding?

My eventual solution was to add a co-author who has more expertise in the platform and qualitative methods than I do, and who also works in game studies. Her partner and occasional co-author is also a data scientist, one far more skilled than I am at using big data techniques. We ended up looking at the data again and reinterpreting it, which led us to consider new theories and, in turn, led us to go over the data again in an iterative approach. Our further thinking led us to consider how to get data from multiple platforms about the players of interest, which in some cases seems impossible because those data aren't made available by the companies in question. Each of these solutions, which worked for this specific project, should also be viewed in a wider and more general context within academic research while keeping in mind the issues faced by open-source advocates and practitioners.

3. Solutions in a Larger Context

In summary, the solutions to the challenges we had to resolve were as follows: 1) Use an interdisciplinary team with expertise across methods, theory, and topic, 2) use a continually iterative understanding of the theory, data, methods, and findings, and 3) acknowledge the limitations that may stem from a data environment that includes free data and protected or unattainable data.

More broadly, two additional issues hover over research, issues with which both computational social scientists and open-source advocates must deal. They must: 1) Work within a potentially restrictive legal environment in a creative manner to move work forward (for open-source there is copyleft, for CSS there is working around or through terms of service [Fiesler et al., 2020; Halavais, 2019]), and 2) challenge that restrictive legal environment directly (as suggested by Puschmann [2019], and as successfully undertaken by Christian Sandvig at the University of Michigan [American Civil Liberties Union, 2020]).

3.1. Team Work

The advanced questions we can ask now, building on decades of research and methods, call for advanced methods and understanding. Advanced methods require more than just large amounts of raw computing power; they call for both quantitative and qualitative methodological approaches (Menchen-Trevino, 2013), a broad understanding of theories, topical expertise, and computational skills. In short, CSS done well requires teams (Lazer et al., 2009).

Large, complex data sets require more than complex statistical analysis to come to human answers. Humans are messy and beautiful, and qualitative methods are much better positioned to capture and understand the beautiful mess than are quantitative methods (Law, 2004). Used together, however, they can present enlightening pictures of human behavior—hence the somewhat recent move toward mixed methods (Creswell, 2009) and the understanding that both types of methods complement each other (Strauss & Corbin, 1998).

This necessary variety is similar to work in FOSS, where large projects require teams with varied exper-

tise depending on the project, which can include coding languages (Python), analysis languages (R), databases (MongoDB, MySQL), all parts of an operating system (Linux), web browsers (Firefox), and graphics software (GIMP).

3.2. Continually Iterative

Open-source projects take time to put together, test, and release. Creators might roll out new releases. That is, people have constantly worked on them, considered issues, and revised along the way. Although an end result emerges, the ongoing process is a vital part of the overall effort. The same is true, or should be, for many CSS projects.

The way we write up research makes it seem as if it has followed a nice, linear narrative that happens to fit the journal article format rather well. First, we read some literature, and that makes us think of some hypotheses and some methods to test them, and only then do we happen to find (or create) the perfect data and arrive at publishable results. But much research does not work this way. Preregistration is one important step for certain types of studies (Nosek, Ebersole, DeHaven, & Mellor, 2018); it is an important acknowledgment of this issue and of the file drawer problem (Rosenthal, 1979).

Researchers working on qualitative and combined methodologies have engaged with this issue and have espoused the usefulness of an iterative approach (Davidson, Edwards, Jamieson, & Weller, 2019; Muller, Guha, Baumer, Mimno, & Shami, 2016). For instance, in a grounded theory approach, researchers might iteratively build categories from data, revisiting the data again and again as they recognize more categories (Strauss & Corbin, 1998). This process can be especially important for CSS projects, where understanding the often large and diverse data set takes more than one pass.

3.3. Mixed Data Environment

In addition to mixed-methods approaches with quantitative and qualitative data (Creswell, 2009; Nelson, 2017), computational social scientists work in an environment with a variety of available data. Some are free, such as Wikipedia data. Some data have slightly restrictive license requirements, such as some APIs that require registration and user tokens or authentication (Bruns, 2019). Some data require payment before one can access them. Some data repositories, such as Pushshift, hope for donations. Other data might be curated (Gruzd, 2016). Some data have been hacked and released (Poor, 2017) or unwisely released (Resnick, 2016). Other data are only available to in-house researchers, or not at all.

This situation is similar in part to the environment in which FOSS programmers work, in that code is available under a variety of licenses, ranging from free to restricted to simply unavailable.

3.4. Creative Solutions

Turning copyright on its head into copyleft was a highly creative solution to the problems that FOSS advocates wanted to address. Copyright, at least in the US context—even with fair use exceptions—is almost always used to deny people the right to copy. FOSS advocates needed a way to allow copying, but with certain restrictions, allowances, and requirements, such as making the changed code available for free or giving credit to previous coders.

Researchers in CSS need access to data but cannot always get it (Bruns, 2019; Freelon, 2018). One problem can arise from a site's terms of service, which may disallow data scraping even when it is technologically possible (Fiesler et al., 2020; Halavais, 2019). Although researchers can ignore the terms of service and scrape a website anyway—hoping to avoid rate limits, throttling, and getting blocked completely—researchers can also approach users directly (Halavais, 2019). Another approach could be to claim fair use doctrine (under American law) as conferring a right to copy the information for academic research purposes.

3.5. Challenge the Existing Structure

FOSS advocates took on the commercial software industry and succeeded. FOSS software is widespread, from some of our own CSS tools to the Linux operating system to the Apache web server (which has been the most widely used web server for many years). Even some large for-profit corporations like IBM support FOSS. But the effort has not always been easy, and the FOSS world is no stranger to lawsuits in which FOSS licensing terms have been upheld by the courts or where settlements have been reached in favor of the FOSS litigant (e.g., Neuburger, 2009; Smith, 2009; Stricklett, 2020).

Most professors, however, are not used to acting as social agitators or legal advocates, except perhaps in social work or law. But FOSS advocates needed to face this challenge in order to legitimize their work. Similarly, academic researchers may need to face established legal and economic structures in order to legitimize important research efforts, although to date few have (e.g., American Civil Liberties Union, 2020). One approach is to educate and work with legislative bodies (Puschmann, 2019); another is to use the courts. Although professors are supposed to have some legal protections for their work, few may want to risk a legal challenge with an unclear outcome to pursue one research project when they could do other research instead. This reluctance may have to change. Additionally, institutions such as universities and associations (e.g., the Association for Computing Machinery) have greater resources compared to individuals and may have to lead in this area, whether pushed by members or guided by leadership.

Note that such a legal effort, like FOSS's work and its now-established legal precedent, requires a team of

people. Taking a legal effort to court is daunting. The recent and successful challenge to the Computer Fraud and Abuse Act in the US was led by Christian Sandvig (American Civil Liberties Union, 2020), but Sandvig is not a lawyer. He is a communication studies professor at the University of Michigan. Other plaintiffs in that legal action included researchers from the University of Illinois and Northeastern University, as well as one commercial publisher, although the American Civil Liberties Union and its legal team were behind the case, and it was Sandvig who gave his name to the case: *Sandvig v. Sessions* (2018). The findings of the case are narrow, but they may eventually be seen as an important and inspirational first step on a longer journey.

3.6. Remaining Issues

One place where this comparison of FOSS with CSS falls short is that in the open-source world, the programmers make the code, the content, and the data they work with. In CSS, researchers mostly do not make the data; they scrape it or use an API to access it. Hence the idea of researchers creating data, via surveys and interviews perhaps in a mixed-methods study, may be of great importance (Bruns, 2019; Freelon, 2018). Beyond creating data in this way, the next step parallel to the FOSS example is to make the data available. Indeed, the idea of making data available has been gaining traction among researchers, although important ethical issues arise when sharing human-centric digital data (Fiesler et al., 2020), which is a problem when considering whether social media and other companies should make data available at all.

Could academics and industry work together (Puschmann, 2019)? Some commercial software companies do support FOSS work by their employees, but the academy/industry relation in social media is somewhat different. Some companies such as Facebook have their own internal research departments, and overall, their motivation to share data with academics is unclear. Perhaps their data drive their advertising revenue and are locked down, but using employee time and company resources to aid academic research also subtracts from the bottom line.

4. Conclusions and the Way Forward

The initial findings for my study—that Reddit users were not moving from the subreddit for the old Elder Scrolls game to the subreddit for the new Elder Scrolls game—was accurate if unexpected, despite having data on more than 2,25 million posts for the overall study across two one-year periods. The literature and my own personal experience had led to a well-grounded, albeit unsupported, hypothesis. My mistake was that I had tried to do too much on my own. In the FOSS world, there is a saying: 'with enough eyeballs, all bugs are shallow.' I needed a team of experts. Eventually we iterated over the research from start to finish, and then over it again (theory, data, methods, conclusions), reaching better conclusions. We debated what we could do with the data we could access, and how we could access it, and wondered about the data we could not get. This picture is familiar to FOSS practitioners.

CSS practitioners may not be prepared to be part of a movement or a revolution in the way that many FOSS advocates have seen themselves (DiBona et al., 1999), but some are already taking steps in that direction. We have to engage with potential legal issues with the terms of service, and we might choose to ignore the terms of service altogether. We might have to deal with intransigent institutional review boards which are there not to help us but instead to protect the university (Halavais, 2019), and whose word is essentially university law. We may have to work with legislative governmental bodies, as suggested by Puschmann (2019). Attempting to challenge national laws through the courts and reframe them in our favor is a daunting step that requires several years, a solid case, and a top-notch legal team (American Civil Liberties Union, 2020). So far, few have dared go this route. More may have to do so if we are to flourish as a field. The alternative is to sit passively while big data companies profit from their advertising and guide the laws in ways that favor their own commercial interests, to the disadvantage of academic research for the greater social good.

Acknowledgments

I thank Dr. Johannes Breuer, Dr. Tim Wulf, and Dr. Rohangis Mohseni for their efforts in organizing this thematic issue and for their encouragement for this submission, and the anonymous reviewers who helped guide this article toward its current form.

Conflict of Interests

The author declares no conflict of interests.

References

American Civil Liberties Union. (2020). Federal court rules 'big data' discrimination studies do not violate Federal anti-hacking law. *American Civil Liberties Union*. Retrieved from https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti

Baym, N. (2007). The new shape of online community: The example of Swedish independent music fandom. *First Monday*, *12*(8).

Bethesda Softworks. The Elder Scrolls IV: Oblivion [Video game]. (2006). Rockville, MD: Bethesda Softworks.

Bethesda Softworks. The Elder Scrolls Online [Video game]. (2014). Rockville, MD: Bethesda Softworks.

Bethesda Softworks. The Elder Scrolls V: Skyrim [Video game]. (2011). Rockville, MD: Bethesda Softworks.

Bethesda Softworks. The Elder Scrolls: Arena [Video game]. (1994). Rockville, MD: Bethesda Softworks.

Blizzard Entertainment. World of Warcraft [Video game]. (2004). Irvine, CA: Blizzard Entertainment.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information Communication and Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bulmer, M. (2001). Social measurement: What stands in its way? *Social Research*, *68*(2), 454–480.

Carey, J. W. (1989). *Communication as culture: Essays on media and society*. New York, NY: Routledge.

Cavicchi, D. (1998). *Tramps like us: Music and meaning among Springsteen fans*. Oxford: Oxford University Press.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. London: Sage.

Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2019). Big data, qualitative style: A breadth-and-depth method for working with large amounts of secondary qualitative data. *Quality and Quantity*, *53*(1), 363–376. https://doi.org/10.1007/s11135-018-0757-y

Dewey, J. (1927). *The public and its problems*. Denver, CO: Swallow Press.

DiBona, C., Ockman, S., & Stone, M. (Eds.). (1999). *Open-sources: Voices from the open-source revolution*. Sebastopol, CA: O'Reilly.

Fiesler, C. (2007). Everything I need to know I learned from Fandom: How existing social norms can help shape the next generation of user-generated content. *Vanderbilt Journal of Entertainment and Technology Law*, *10*, 729–762.

Fiesler, C., Beard, N., & Keegan, B. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the 2020 International Conference on Web and Social Media*.

Fiesler, C., Morrison, S., & Bruckman, A. S. (2016). An archive of their own: A case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2574–2585). New York, NY: ACM Press. https://doi.org/10.1145/2858036.2858409

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1007/978-981-13-0402-6_6

Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0200162

Gamble, C., Gowlett, J., & Dunbar, R. (2014). *Thinking big: How the evolution of social life shaped the human mind*. London: Thames & Hudson.

Gruzd, A. (2016). Social media data stewardship. *Social Media Lab*. Retrieved from https://socialmedialab.ca/2016/03/21/defining-social-media-data-stewardship-smds

Gunderman, H. C. (2020). View of fan geographies and engagement between geopolitics of Brexit, Donald Trump, and Doctor Who on social media. *Transformative Works and Cultures*, *32*. https://doi.org/10.3983/twc.2020.1675

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information Communication and Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Hills, M. (2003). *Fan cultures*. New York, NY: Taylor & Francis.

Iyer, S., Cheng, J., Brown, N., & Wang, X. (2020). When does trust in online social groups grow? In *Proceedings of the fourteenth international AAAI conference on web and social media* (pp. 283–293). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Jenkins, H. (1992). *Textual poachers: Television fans and participatory culture*. New York, NY: Routledge.

Kraut, R. E., & Resnick, P. (2011). *Building successful online communities*. Cambridge, MA: MIT Press.

Law, J. (2004). *After method: Mess in social science research*. New York, NY: Routledge.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L. L., Brewer, D., . . . Alstyne, M. V. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science*, *1*(1), 49–58. https://doi.org/10.1007/s42001-017-0007-4

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research. *Policy & Internet*, *5*(3), 328–339. https://doi.org/10.1002/1944-2866.POI336

Muller, M., Guha, S., Baumer, E. P. S., Mimno, D., & Shami, N. S. (2016). Machine learning and grounded theory method: Convergence, divergence, and combination. In *Proceedings of the 19th international conference on supporting group work* (pp. 3–8). https://doi.org/10.1145/2957276.2957280

Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Neuburger, J. (2009). Jacobsen v. Katzer: Open-source Software project gains key rulings in copyright infringement litigation. *New Media and Technology Law Blog*. Retrieved from https://newmedialaw.proskauer.com/2009/12/16/jacobsen-v-katzer-open-source-software-project-gains-key-rulings-in-copyright-infringement-litigation

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

Panek, E., Hollenbach, C., Yang, J., & Rhodes, T. (2018). The effects of group size and time on the formation of online communities: Evidence from Reddit. *Social Media + Society*, *4*(4), 1–13. https://doi.org/10.1177/2056305118815908

Pearce, C. (2009). *Communities of play*. Cambridge, MA: MIT Press.

Poor, N. (2014). Computer game modders' motivations and sense of community: A mixed-methods approach. *New Media and Society*, *16*(8), 1249–1267. https://doi.org/10.1177/1461444813504266

Poor, N. (2017). The ethics of using hacked data: Patreon's data hack and academic data standards. In M. Zimmer & K. Kinder-Kurlanda (Eds.), *Internet research ethics for the social age: New challenges, cases, and contexts* (pp. 277–280). New York, NY: Peter Lang.

Poor, N., & Skoric, M. M. (2014). Death of a guild, birth of a network: Online community ties within and beyond code. *Games and Culture*, *9*(3), 182–202. https://doi.org/10.1177/1555412014537401

Puente, H., & Tosca, S. (2013). The social dimension of collective storytelling in Skyrim. In *Proceedings of the 2013 DiGRA international conference*.

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information Communication and Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Raymond, E. S. (1999). *The Cathedral and the Bazaar: Musings on Linux and Open-source by an accidental revolutionary*. Sebastopol, CA: O'Reilly.

Resnick, B. (2016). Researchers just released profile data on 70,000 OkCupid users without permission. *Vox*. Retrieved from https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Sandvig v. Sessions, No. 16-1368 (D.D.C. Mar. 30, 2018).

Smith, B. (2009). FSF settles suit against Cisco. *Free Software Foundation*. Retrieved from https://www.fsf.org/news/2009-05-cisco-settlement.html

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*. https://doi.org/10.1177/0894439319843669

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage.

Stricklett, S. (2020). Google v. Oracle: An expansive fair use defense deters investment In original content. *IPWatchdog*. Retrieved from https://www.ipwatchdog.com/2020/01/19/google-v-oracle-expansive-fair-use-defense-deters-investment-original-content/id=117951

Tolkien, J. R. R. (1937). *The Hobbit*. London: George Allen & Unwin.

Tomasello, M. (2010). *Origins of human communication*. Cambridge, MA: MIT Press.

Wellman, B., Boase, J., & Chen, W. (2002). The networked nature of community: Online and offline. *IT & Society*, *1*(1), 151–165.

About the Author

Nathaniel Poor is a Computational Social Scientist who primarily researches online communities, often ones related to games. He is the President and Founder of the Underwood Institute, a non-profit involved in data for good and code for good efforts.

Media and Communication (ISSN: 2183-2439)

Media and Communication is an international open access journal dedicated to a wide variety of basic and applied research in communication and its related fields. It aims at providing a research forum on the social and cultural relevance of media and communication processes.

www.cogitatiopress.com/mediaandcommunication

COGITATIO