Analyzing Character Networks in Crossover Fan Fictions of Archive of Our Own

Thomas Schmidt
Media Informatics Group
University of Regensburg, Germany
thomas.schmidt@ur.de

Johannes Hoffmann
Media Informatics Group
University of Regensburg, Germany
johannes-maximilian.hoffmann
@stud.uni-regensburg.de

Christian Wolff
Media Informatics Group
University of Regensburg, Germany
christian.wolff@ur.de

Abstract

We present results of the application of social network analysis for the characters of crossover fan fictions which are fan fictions merging the characters of multiple fandoms. We scraped a corpus of fan fictions from the platform Archive of our Own consisting of over 82,000 fan fictions. We use the tool *Gephi* to perform social network analysis and create a graph consisting of 98,014 nodes and 2,742,150 edges. We explore the graph looking at quantitative parameters but also by qualitatively-descriptively interpreting the visualized networks. One of the main results are that we found that original characters are an important part of crossover fan fictions and that characters of fandoms mostly cluster according to genre or national background.

1 Introduction

Fan fictions are narrative written works created by fans (mostly hobby authors) using characters, content and plot elements of existing popular media like books, movies, comics or games to write new stories (Hellekson and Busse, 2006). With the rise of the world wide web, the production of such fan fictions has increased significantly with platforms like fanfiction.net[1] or Archive of Our Own (AO3)[2] hosting over 7 million stories or 8 million respectively. Researchers in the humanities examined different facets of this phenomenon like the motivation of the authors (Hellekson and Busse, 2006; Thomas, 2011; Duggan, 2020), the cultural influence of this medium (Van Steenhuyse, 2011; Thomas, 2011; Jamison and Grossman, 2013) or the quantitative dominance of slash fan fiction (fan fictions with a focus on homo-romantic content) (Hellekson and Busse, 2006; Tosenberger, 2008; Duggan, 2017). Since this platforms offer great amounts of accessible narrative prose texts with rich metadata, researchers in natural language processing (NLP) are using fan fictions corpora for the analysis and evaluation of various machine learning tasks (Kim and Klinger, 2019; Muttenthaler et al., 2019; Liu et al., 2019; Vilares and Gómez-Rodríguez, 2019; Zhang et al., 2019). From a digital humanities (DH) perspective, various research has been pursued like the analysis of gender stereotypes (Fast et al., 2016), the role of user feedback (Frens et al., 2018; Pianzola et al., 2020b), national specifics of fan fictions (Schmidt et al., 2021e), intertextuality (Büchler, 2018; Kleindienst and Schmidt, 2020), cultural evolution (Pianzola et al., 2020a) or the general analysis of the fan culture based on the metadata (Milli and Bamman, 2016; Yin et al., 2017).

In this paper, we focus on the specific genre of crossover fan fictions: fan fictions in which characters or settings of two or more different fandoms (the franchise a fan fiction is about e.g. *Harry Potter*) co-occur, for example stories in which *Harry Potter* meets *Sherlock Holmes*. We argue that this genre is especially interesting for literary and fan studies to analyze how fans alter and combine original source material. We analyze this phenomenon with computational social network analysis (SNA) which is a method to model social structures using networks and graph theory. Actors are nodes of a graph with edges representing relationships (Otte and Rousseau, 2002). In DH, this method has been applied, for example, for the analysis of characters

[1] https://www.fanfiction.net/
[2] https://archiveofourown.org/

in novels (Grayson et al., 2016; Rahul et al., 2021) , TV shows (Zhang et al., 2018), to analyze collocations (Schmidt et al., 2020b) or, in the case of fan fictions, to visualize authors and readers networks (Carvallo and Parra, 2020).

In our setting, we regard the fictional or non-fictional characters of fan fictions as nodes and the co-occurrence of two characters as edge between these nodes. The weight of the edges is defined by the absolute number of shared appearances of this character combination. We analyze the resulting character networks based on quantitative parameters of graph theory and SNA but also qualitatively-descriptively by analyzing and interpreting the visualized networks.

2 Corpus

We acquired fan fictions by scraping a corpus of crossover fan fictions of the platform AO3. AO3 is one of the largest fan fiction platforms and permits the scraping of its content according to the terms of service.[3] We scraped the crossover fan fictions of the platform in June 2021. Filtering all fan fictions (over 8 million) by the "Crossover"-tag results in around 82,000 works. We scraped the HTML-pages of these works via a Python script and transformed the content including text and the metadata of the stories in a JSON-format. Table 1 summarizes some general statistics of the corpus.[4]

The corpus consists of 82,250 works by 37,965 authors written in 35 languages with the majority of the works being English (92.2%). Measured in words, the corpus amounts to over 1 billion words. The most popular fandoms are *Harry Potter*, *Supernatural* and the *Marvel Cinematic Universe*. While the corpus is of great value for text based analysis or investigations concerning the rich metadata (e.g. fandoms, relationship types), we will focus on the character metadata as main component for the network analysis.

The characters assigned to fan fictions as metadata tags are regarded as the nodes for the SNA. Please note that AO3 emphasises the normalisation of metadata by assigning workers for this task. Thus differences in spelling of character names, while certainly still existing, are mostly normalized. The corpus features 98,014 unique characters of which 61.1% are part of only one work (see figure 1).

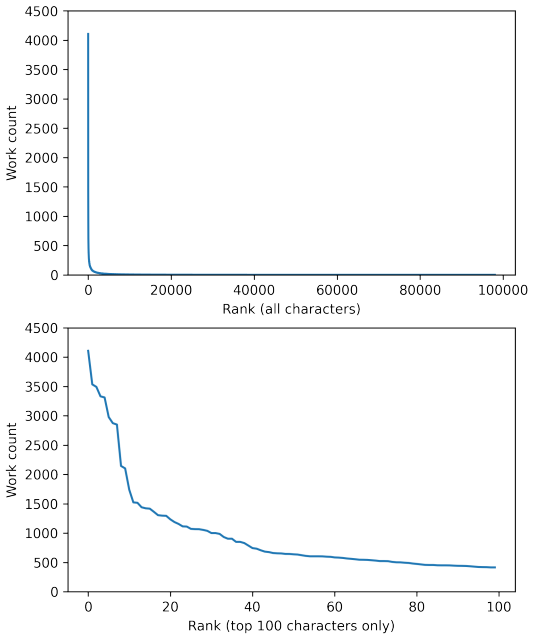

Figure 1: Line graph for the proportion of works including a specific character ordered by rank. The y-axis is the number of works the specific character is part of. The x-axis lists the characters ranked by frequency (e.g. 20 is the twentieth most frequent character). Top: includes all characters; bottom: only includes the 100 highest ranked characters.

3 Network Analysis and Preliminary Results

For the network analysis, we use the tool *Gephi*[5], which is a well-known tool for this taks also applied in other DH research (Grandjean, 2016; Carvallo and Parra, 2020). We calculated the shared appearances between each unique characters and transformed the data into the necessary *Gephi*-format. We analyze various graph-based metrics via *Gephi* and explore the visualization of the network. In the following, we present some preliminary results of our analysis. More results and additional material will be made available on GitHub.[6]

The network consists of 98,014 nodes, each representing one character, and 2,743,150 edges, each representing at least one shared appearance between the two connected nodes/characters. Compared to other research in DH (Grayson et al., 2016;

[3] https://archiveofourown.org/tos_faq

[4] Due to legal constraints the corpus is currently only available via request (thomas.schmidt@ur.de). More information about the corpus will be made available on GitHub: https://github.com/lauchblatt/CrossoverFanFictions

[5] https://gephi.org/

[6] https://github.com/lauchblatt/CrossoverFanFictions

Metric	Value
Number of fan fictions	82,050
Number of words	1,056,351,915
Average number of words per fan fiction	12,874
Median number of words per fan fiction	2,987
Maximum number of words	2,407,284
Number of fandoms	17,732
Most frequent fandoms	*Harry Potter* (8.4%), *Supernatural* (6.9%), *Marvel Cinematic Universe* (5.3%)
Number of unique characters	98,014
Most frequent characters	Dean Winchester (5.0%), Sam Winchester (4.3%) (both *Supernatural*), Tony Stark (4.3%) (*Marvel Cinematic Universe*)
Most frequent relationship types	general (40.3%), male-male (33.2%), female-male (26.2%)

Table 1: General statistics about the crossover fan fiction corpus.

Carvallo and Parra, 2020; Rahul et al., 2021), this is a very large network. The characters with the most shared appearances are "Dean Winchester" and "Sam Winchester" (3,201) (*Supernatural*), followed by "Sherlock Holmes" and "John Watson" (2,619) (*Sherlock Holmes*) and "Steve Rogers" and "Tony Stark" (2,026) (*Marvel Cinematic Universe*).

The degree of a node is determined by the number of edges to other nodes this node has. The average degree of all nodes in the network is 55.97 (*Median (Mdn)* = 18, *Standard deviation (SD)* = 166.72), the weighted degree (sum of the weight of the edges) is 90.98 (*Mdn* = 19, *SD* = 492.23). The character tags with the highest degree are "Original Characters" (15,484), "Original Female Character(s)" (10,792) and "Original Male Character(s)" (8,914). These are characters created and invented by the authors themselves, oftentimes depicting the narrator. This shows that, actually, authors are not focusing so much on fandom-merging in crossover fan fictions but on the integration of original characters into existing fandoms. The analysis of the weighted degree supports this findings but also highlights the dominance of Marvel characters in our corpus: "Tony Stark" (38,916), "Steve Rogers" (35,588) and "Original Characters" (31,455) have the highest weighted degree.

To calculate modules/clusters (collections of nodes strongly intertwined) we use the default modularity algorithm by *Gephi* (Blondel et al., 2008). The algorithm detects 1,229 modules in the network. Investigating these modules shows, unsurprisingly, that mostly individual fandoms represent clusters. Each module contains 79.75 nodes on average (*Mdn* = 2, *SD* = 617.9). 1,120 (91.13%) of all modules contain less than 10 nodes highlighting the frequency of individual stories with unique sets of characters.

The clustering coefficient (CC) (Watts, 1998) is a metric that measures to which degree a node tends to cluster and ranges from 0 to 1. A node with a CC of 1 only shares edges with other nodes where all of these neighbouring nodes share an edge with each other while a node with a CC of 0 only shares edges with other nodes where none of these neighbouring nodes share an edge with each other. The average CC of all nodes is 0.81 (Mdn = *1*, SD = *.28*). 58,697 (59.98%) of the nodes have a CC of 1 and 2,764 (2.82%) one of 0. The overall CC is therefore rather high with many nodes interacting in their cluster. Indeed, there are many clusters determined by only one single story that lead to these results. "Original Characters" (.007), "Original Female Character(s)" (.011) and "Original Male Character(s)" (.014) have the lowest clustering coefficient and therefore cluster the least, showing again that the majority of crossover fan fictions connects specifically to originally invented characters of the writers.

The network visualization is performed with the *OpenOrd* layout implementation (Martin et al., 2011) included in *Gephi* since it is recommended

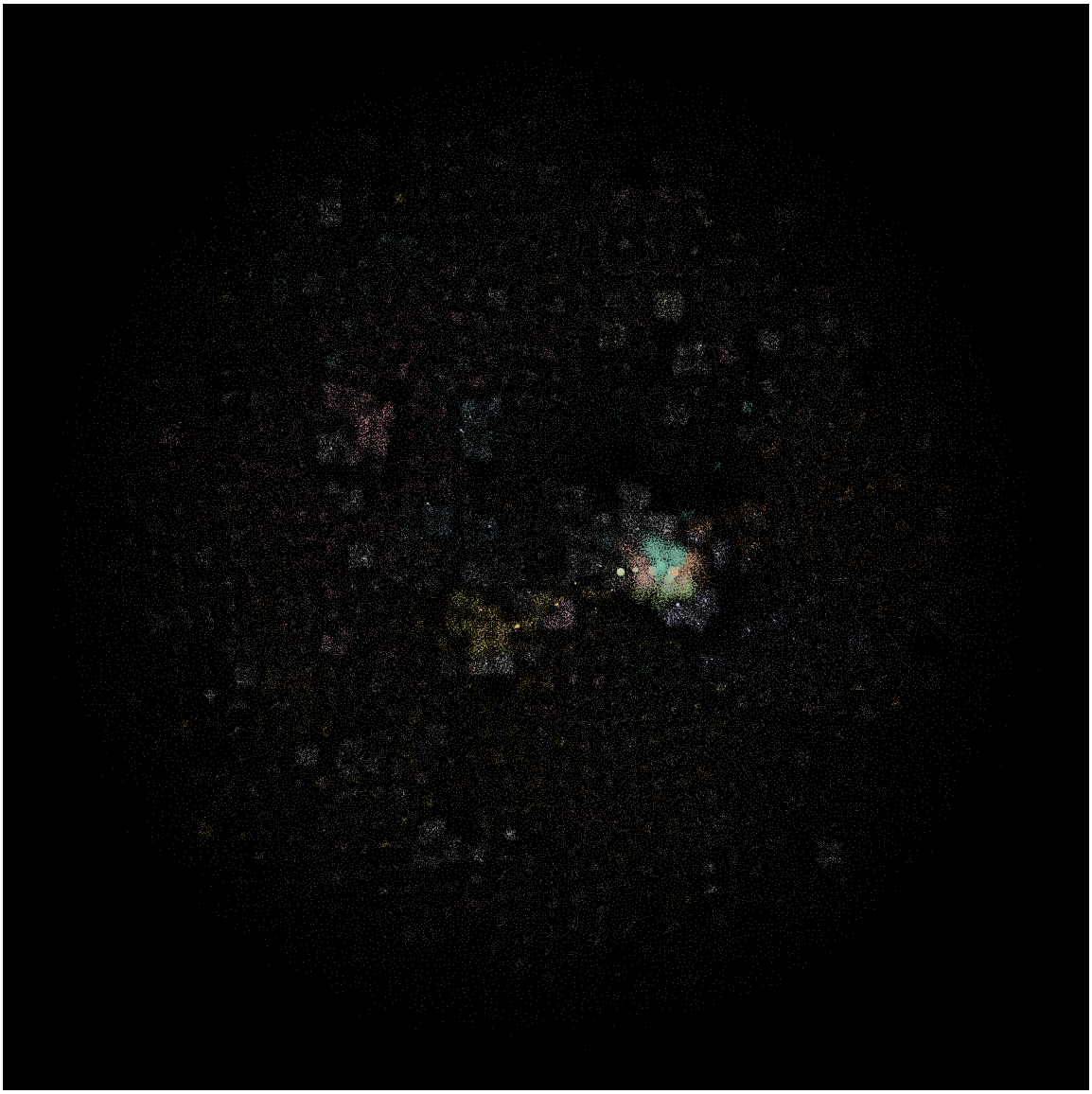

Figure 2: Visualization of the overall network. Edges are hidden. Different modules/clusters are in different colors. Size of nodes is determined by frequency of a character.

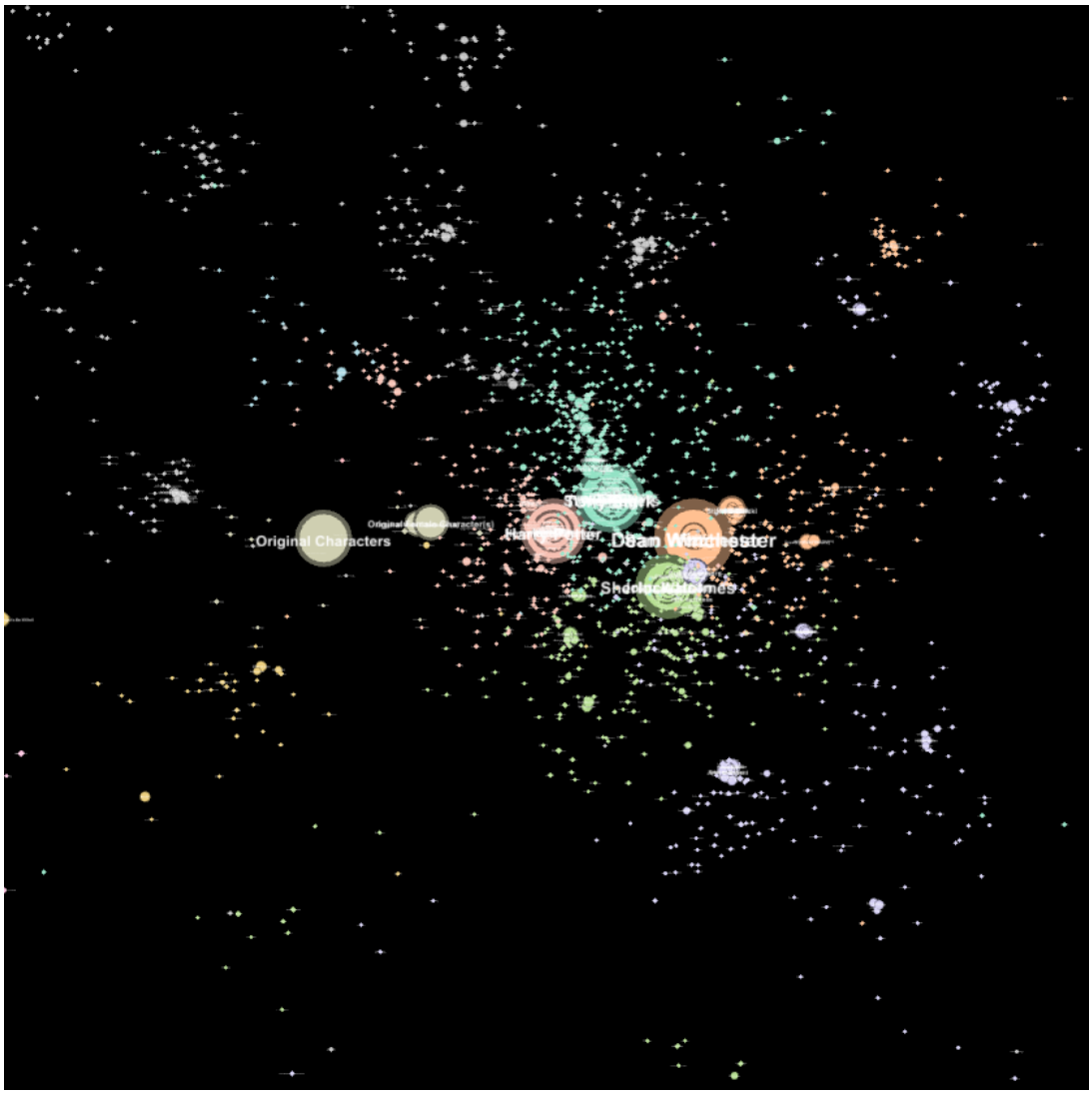

Figure 3: Visualization of the main clusters in the center of the network containing the most popular fandoms. Red is the *Harry Potter* fandom, cyan *Marvel Cinematic Universe*, orange *Supernatural*, green *Sherlock*, beige the "Original Characters". On the upper bottom left (yellow) is the anime and *My Hero Academia* cluster.

for networks of this size. Figure 2 shows the overall network hiding edges for better reading. The size of the nodes is determined by the overall frequency of a character. The overall network shows the large amount of individual clusters. However, the largest fandoms and most important character-nodes cluster in the middle of the network (see figure 3 for a zoomed-in view). We found that the largest nodes of each of this sub-clusters is defined by a large node representing the main character surrounded by the side characters. Furthermore, the close proximity of these clusters shows that crossovers are also mostly created among the most popular general fandoms.

A more in-depth analysis of the main clusters showed that the cyan colored *Marvel* cluster also contains a lot of nodes of other superhero fandoms like *Batman*. The orange colored cluster represents the fandoms *Supernatural* and *Teen Wolf* showing a strong connection of these two series. Finally, the green colored *Sherlock* module includes other British content as nodes of the fandoms *James Bond* and *Doctor Who*. The cluster for original characters is more to the center showing the interconnection with most fandoms. Another very large module of the network is actually determined by the fandom of the anime *My Hero Academia* (upper bottom left in figure 3) and shows connections with other anime and animated series.

4 Discussion

We were able to make some interesting findings about crossover fan fictions via the quantitative analysis of character network metrics as well as the descriptive interpretation of the visualizations. We found that original characters are an important part of this genre and also showed that the main fandoms cluster strongly with each other and that individual fandom clusters are determined by genre (*Supernatural/Teen Wolff* or the anime cluster) or national background (*Sherlock Holmes, James Bond, Doctor Who*).

However, since the sheer size of the network hindered the analysis and interpretation, we want to continue our work by filtering the network of very rare fandoms and small clusters. Additionally, the textual and metadata content of the corpus offers a lot more analysis possibilities besides SNA. We want to continue the analysis of this special type of fan fictions via other popular methods in DH like sentiment analysis(Schmidt and Burghardt, 2018a,b; Schmidt et al., 2021a), emotion analysis (Schmidt et al., 2021b,c) or topic modeling (Moßburger et al., 2020; Schmidt et al., 2020a). Further, we see potential analyzing the transformation process between the source material and the fan fiction (Kleindienst and Schmidt, 2020) also concerning movie based material which might offer exciting multimodal ways of analysis regarding current advances of digital film studies (Schmidt et al., 2019, 2021d; Schmidt and Wolff, 2021).

References

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Marco Büchler. 2018. The Re-Creation Of Harry Potter: Tracing Style And Content Across Novels, Movie Scripts And Fanfiction. *DH 2018*, page 4.

Andrés Carvallo and Denis Parra. 2020. Analyzing Network Effects on a Fanfiction Community. *arXiv:1909.02886 [cs]*. ArXiv: 1909.02886.

Jennifer Duggan. 2017. Revising Hegemonic Masculinity: Homosexuality, Masculinity, and Youth-Authored Harry Potter Fanfiction. *Bookbird: A Journal of International Children's Literature*, 55(2):38–45. Publisher: Johns Hopkins University Press.

Jennifer Duggan. 2020. Who writes Harry Potter fan fiction? Passionate detachment, "zooming out," and fan fiction paratexts on AO3. *Transformative Works and Cultures*, 34.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, page 9.

John Frens, Ruby Davis, Jihyun Lee, Diana Zhang, and Cecilia R. Aragon. 2018. Reviews matter: How distributed mentoring predicts lexical diversity on fanfiction.net. *ArXiv*, abs/1809.10268.

Martin Grandjean. 2016. A social network analysis of twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3.

Siobhán Grayson, Karen Wade, Gerardine Meaney, Jennie Rothwell, Maria Mulvany, and Derek Greene. 2016. Discovering structure in social networks of 19th century fiction. In *Proceedings of the 8th ACM Conference on Web Science*, pages 325–326, Hannover Germany. ACM.

Karen Hellekson and Kristina Busse. 2006. *Fan Fiction and Fan Communities in the Age of the Internet: New Essays*, illustrated auflage edition. McFarland and Company, Inc., Jefferson, N.C.

Anne Jamison and Lev Grossman. 2013. *Fic: Why Fanfiction Is Taking Over the World*, illustrated auflage edition. Smart Pop, Dallas, Texas.

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, Wincing Leia, and a Seriously Great Friendship: Learning to Classify Emotional Relationships of Fictional Characters. *Proceedings of NAACL-HLT 2019*.

Nina Kleindienst and Thomas Schmidt. 2020. Investigating the Transformation of Original Work by the Online Fan Fiction Community: A Case Study for Supernatural. In *Digital Practices. Reading, Writing and Evaluation on the Web*, Basel, Switzerland.

Chen Liu, Muhammad Osama, and Anderson de Andrade. 2019. DENS: A Dataset for Multi-class Emotion Analysis. *arXiv:1910.11769 [cs]*. ArXiv: 1910.11769.

Shawn Martin, W. Michael Brown, Richard Klavans, and Kevin W. Boyack. 2011. OpenOrd: an open-source toolbox for large graph layout. In *Visualization and Data Analysis 2011*, volume 7868, pages 45 – 55. International Society for Optics and Photonics, SPIE.

Smitha Milli and David Bamman. 2016. Beyond Canonical Texts: A Computational Analysis of Fanfiction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2048–2053, Austin, Texas. Association for Computational Linguistics.

Luis Moßburger, Felix Wende, Kay Brinkmann, and Thomas Schmidt. 2020. Exploring Online Depression Forums via Text Mining: A Comparison of Reddit and a Curated Online Forum. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 70–81, Barcelona, Spain (Online). Association for Computational Linguistics.

Lukas Muttenthaler, Gordon Lucas, and Janek Amann. 2019. Authorship attribution in fan-fictional texts given variable length character and word n-grams. In *CLEF*.

Evelien Otte and R. Rousseau. 2002. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28:441 – 453.

Federico Pianzola, Alberto Acerbi, and Simone Rebora. 2020a. Cultural accumulation and improvement in online fan fiction. preprint, Open Science Framework.

Federico Pianzola, Simone Rebora, and Gerhard Lauer. 2020b. Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers' comments in the margins. *PLOS ONE*, 15(1):e0226708.

Rahul, Ayush, Divya Agarwal, and Devika Vijay. 2021. Genre Classification using Character Networks. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 216–222.

Thomas Schmidt and Manuel Burghardt. 2018a. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Thomas Schmidt and Manuel Burghardt. 2018b. Toward a Tool for Sentiment Analysis for German Historic Plays. In *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018*, pages 46–48, Lausanne, Switzerland. Laboratoire laussannois d'informatique et statistique textuelle.

Thomas Schmidt, Manuel Burghardt, and Christian Wolff. 2019. Toward Multimodal Sentiment Analysis of Historic Plays: A Case Study with Text and Audio for Lessing's Emilia Galotti. In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019)*, volume 2364 of *CEUR Workshop Proceedings*, pages 405–414, Copenhagen, Denmark. CEUR-WS.org.

Thomas Schmidt, Johanna Dangel, and Christian Wolff. 2021a. Senttext: A tool for lexicon-based sentiment analysis in digital humanities. In Thomas Schmidt and Christian Wolff, editors, *Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021)*, volume 74, pages 156–172. Werner Hülsbusch, Glückstadt.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021b. Emotion Classification in German Plays with Transformer-based Language Models Pretrained on Historical and Contemporary Language. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 67–79, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Thomas Schmidt, Katrin Dennerlein, and Christian Wolff. 2021c. Using Deep Learning for Emotion Analysis of 18th and 19th Century German Plays. In Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels-Oliver Walkowski, Joëlle Weis, and Ulrike Wuttke, editors, *Fabrikation von Erkenntnis. Experimente in den Digital Humanities*.

Thomas Schmidt, Alina El-Keilany, Johannes Eger, and Sarah Kurek. 2021d. Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies. In *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*, Krasnoyarsk, Russia.

Thomas Schmidt, Johanna Grünler, Nicole Schönwerth, and Christian Wolff. 2021e. Towards the Analysis of Fan Fictions in German Language: Exploration of a Corpus from the Platform Archive of Our Own. In *2nd International Conference of the European Association for Digital Humanities (EADH 2021)*, Krasnoyarsk, Russia.

Thomas Schmidt, Philipp Hartl, Dominik Ramsauer, Thomas Fischer, Andreas Hilzenthaler, and Christian Wolff. 2020a. Acquisition and analysis of a meme corpus to investigate web culture. In *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Conference Abstracts*, Ottawa, Canada.

Thomas Schmidt, Florian Kaindl, and Christian Wolff. 2020b. Distant reading of religious online communities: A case study for three religious forums on reddit. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*, pages 157–172, Riga, Latvia.

Thomas Schmidt and Christian Wolff. 2021. Exploring Multimodal Sentiment Analysis in Plays: A Case Study for a Theater Recording of Emilia Galotti. In *Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021)*, pages 392–404, Amsterdam, The Netherlands.

Bronwen Thomas. 2011. What Is Fanfiction and Why Are People Saying Such Nice Things about It? *Storyworlds: A Journal of Narrative Studies*, 3.

Catherine Tosenberger. 2008. Homosexuality at the Online Hogwarts: Harry Potter Slash Fanfiction. *Children's Literature*, 36(1):185–207. Publisher: Johns Hopkins University Press.

Veerle Van Steenhuyse. 2011. The Writing and Reading of Fan Fiction and Transformation Theory. *CLCWeb: Comparative Literature and Culture*, 13(4).

David Vilares and Carlos Gómez-Rodríguez. 2019. Harry Potter and the Action Prediction Challenge from Natural Language. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2124–2130, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven H. Watts, Duncan J.and Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Kodlee Yin, Cecilia Aragon, Sarah Evans, and Katie Davis. 2017. Where No One Has Gone Before: A Meta-Dataset of the World's Largest Fanfiction Repository. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6106–6110, Denver Colorado USA. ACM.

Lingfei Zhang, Chunfang Li, Luyu Fan, and Minyong Shi. 2018. Analysis of Character Relationship in TV Series Based on Complex Network. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 723–727.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating Character Descriptions for Automatic Summarization of Fiction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7476–7483.