Where No One Has Gone Before: A Meta-Dataset of the World’s Largest Fanfiction Repository

Kodlee Yin, Cecilia Aragon, Sarah Evans, Katie Davis
University of Washington, Seattle, USA
kodlee@frult.me, {aragon,sarahe,kdavis78}@uw.edu

Figure 1. An overview of fanfiction story metadata by fandom/category. The category Books contains the most popular fandom (Harry Potter) by number of stories, but the category Anime is much deeper, with multiple fandoms generating more stories.

ABSTRACT
With its roots dating to popular television shows of the 1960s such as Star Trek, fanfiction has blossomed into an extremely widespread form of creative expression. The transition from printed zines to online fanfiction repositories has facilitated this growth in popularity, with millions of fans writing stories and adding daily to sites such as Archive Of Our Own, Fanfiction.net, FIMfiction.net, and many others. Enthusiasts are sharing their writing, reading stories written by others, and helping each other to grow as writers. Yet, this domain is often undervalued by society and understudied by researchers. To facilitate the study of this large but often marginalized community, we present a fully anonymized data release (via differential privacy) of the metadata from a large fanfiction site (to protect author privacy, story, profile, and review text is excluded, and only metadata is provided). We use visual analytics techniques to draw several intriguing insights from the data and show the potential for future research. We hope other researchers can use this data to explore further questions related to online fanfiction communities.

Author Keywords
Fanfiction; online communities; youth.

ACM Classification Keywords
H.5.3 Group and organization interfaces: Web-based interaction.
INTRODUCTION
Writing, sharing, and discussing fanfiction is a hugely popular online activity, and the site we call FanFic.com (anonymized to comply with terms of service) is a vast online fanfiction repository. As of June 5, 2016, there were 6,402,752 stories, 166,739,153 story reviews, and 1,455,952 authors on the site. To put this last figure in context, there were 113,000 active editors on English Wikipedia on the same date. Yet the number of ACM papers published on Wikipedia far exceeds the number of papers studying fanfiction: 1,581 Wikipedia papers compared to a handful of papers that deal with fanfiction. We do not claim that fanfiction has had the impact or importance of Wikipedia; nevertheless, the phenomenon of fanfiction is clearly an understudied online context relative to its popularity.

It is not simply its popularity that makes fanfiction worthy of investigation; it is what people are doing on these sites—and how they are benefiting—that merits empirical inquiry. Authors (often young people) are publishing stories that are several hundred thousand words in length, sometimes exceeding the length of the original works on which they were based. Readers are offering encouragement and constructive feedback on stories, which authors use to improve their writing [5,9]. All of this happens in a positive, supportive community atmosphere that stands in stark contrast to the negativity and even hate speech that is found on so many online sites (e.g., Reddit, 4chan, comments posted on news sites).

Previous work has documented the benefits that people gain through their participation in online fanfiction communities. Authors experience mentorship from the community, grow as writers, gain recognition for their work, and form meaningful connections with other fans [2-5,9-18]. Most existing research on fanfiction has consisted of relatively small-scale, ethnographic investigations. These studies offer valuable insight into users’ experiences and interactions in online fanfiction communities. However, they cannot provide a comprehensive view of the full scope of activity on a fanfiction repository. The current study provides such a view, constituting the first fully anonymized data release of a fanfiction site’s complete metadata, consisting of six million stories across ten thousand fandoms. Previous work that undertook quantitative analysis only sampled the data [7,19,21]. We utilize differential privacy [8] for the data release so that researchers can explore questions related to online fanfiction communities without violating the privacy of the authors. A secondary contribution of this work is the demonstration of a visual analytics approach to draw insight from this vast corpus of human-generated data.

PREVIOUS WORK AND BACKGROUND
History of Fanfiction
Fanfiction is a type of transformative work in which fans of a variety of different media—including television shows, movies, comic books, anime, and video games—write stories that are based on the original media but expand upon or alter it. Fanfiction authors may consider plot trajectories different from the original story, explore the untold background stories of characters, or place the characters in an “alternate universe” that is completely different from the original setting. Fanfiction stories can vary in length from “one shots” of less than 1000 words to novel-length stories.

Fanfiction developed within the context of popular television shows of the 1960s, such as Star Trek and The Twilight Zone [11]. Yet, one might go considerably further back in history and regard the works of Shakespeare and Virgil as early examples of fanfiction. Well-known works of contemporary fanfiction include Wicked: The Life and Times of the Wicked Witch of the West (based on Baum’s The Wonderful Wizard of Oz), Pride and Prejudice and Zombies (based on Austen’s Pride and Prejudice), and The Wind Done Gone (based on Mitchell’s Gone with the Wind). For every popular example of fanfiction, however, there are thousands more fanfiction stories that are unknown to the general public.

Research on Fanfiction
Research on fanfiction began nearly 25 years ago with Jenkins’ seminal investigation of fan culture [11]. This work was pivotal in countering negative stereotypes of fans and fan culture by describing fan communities as rich sites of participatory culture. Jenkins challenged the common view that fanfiction was unoriginal, derivative work, instead positioning fanfiction stories as highly creative acts of manipulating and extending source material in unexpected, boundary-pushing ways.

Since Jenkins’ foundational work, research on fanfiction has addressed issues of legality [20,22,23], identity [1,3,14], sexuality [11,24], and educational value [4,6,9,12,15-18]. For example, Black [2,3,4] explored English Language Learners’ (ELL) participation in online fanfiction communities, focusing in particular on language acquisition and identity formation. More recently, Campbell et al. [5] conducted an in-depth nine-month ethnographic investigation of Fanfiction.net and FIMFiction.net, which included participant observations and interviews with authors. In a follow-up study, they conducted a thematic analysis of 4,500 reader reviews on Fanfiction.net [9].

Selecting ten original canonical works from Fanfiction.net, Milli and Bamman [19] employed natural language processing (NLP) methods to investigate differences in the characters emphasized and the treatment of gender in fanfiction stories compared to the original works. They found that secondary characters tended to be mentioned more frequently than primary characters in fanfiction stories. In addition, they found a small but significant increase in the proportion of female character mentions in fanfiction stories compared to the original works. This investigation represents the first to apply computational methods to the study of fanfiction data, illustrating the potential to derive empirical insights from such an approach. However, the data used by Milli and Bamman
constituted a fairly small proportion of the total corpus of fanfiction published on the site. The authors called for more quantitative analyses of fanfiction texts, and we hope that by making a large set of story metadata publically available and demonstrating its richness, we will be encouraging this type of research.

METHODS

Fanfic.s.com has various views for its website, none of which provide a complete snapshot. The site is organized hierarchically; the top-level subdivision is into nine categories: Anime/Manga, Books, Cartoons, Comics, Games, Misc, Movies, Plays/Musicals, and TV Shows. Each category is subdivided into fandoms (e.g. in the Books category, Harry Potter and Twilight are the two most popular fandoms). Finally, within each fandom, paginated lists of 25 stories per page, which we call fandom pages, provide the title, synopsis, author, content rating, language, and genre(s) (e.g., Romance, Drama) of each story, alongside statistics about each story such as word count, number of chapters, reviews, number of favorites and followers. Optional data includes the last time the story was updated, and characters from the fandom used in the story. Collectively, we define this information as the Fanfic.s.com metadata. It does not contain the story texts themselves, but is a rich and dynamic repository of information about the millions of writers and readers of fanfiction on this site.

We collected the data via a combination of Apache HttpClient and jsoup. The list of fandoms was stored in a MySQL relational database. We processed the paginated fandom pages and stored the raw HTML results as-is for later processing. In total, we scraped and processed 10,339 fandoms on Sunday November 20, 2016. The data was stored in a normalized MySQL database with appropriate tables such as “genre,” “story,” “story_genre.” Variables were aggregated and Laplacian noise added to ensure differential privacy even among outliers.

FINDINGS

The dataset is accessible at http://research.fruit.me. It contains metadata describing 6,807,100 stories across 1,516,335 authors in 44 languages. Each row contains an anonymized user ID, story ID, six quantitative, and six categorical variables as documented on the release site. The stories span 10,294 fandoms and use 46,337 unique characters 9,308,807 times. The 20 genres were applied on 6,159,491 stories, leaving 647,609 uncategorized.

The largest categories were Anime/Manga, TV Shows, and Books with 1,905,055, 1,553,815, and 1,442,290 stories each, respectively. The top genres were Romance, Humor, and Drama with 3,423,862, 1,296,042, and 1,147,377 stories each, respectively. Finally, the top fandoms were Harry Potter, Naruto, and Twilight with 713,814, 387,218, and 212,929 stories, respectively.

We have provided graphs and Tableau visualizations of a few of our most interesting findings. Figure 1 illustrates the differing “shapes” of fandom popularity within categories. Although the category Books contains the single most popular fandom (Harry Potter) by number of stories, the category Anime is much deeper, with multiple fandoms generating more stories overall.

We also observe that the hypothesis, based on our previous ethnographic investigations, that a significant portion of the authors on this site are English-speaking students, appears to be confirmed by the uptick in posting during the northern hemisphere summer months, with a brief blip in late December (Figure 2). Note the steady growth until 2013; we have no explanation for the drop-off after that year, and further research is encouraged.
A simple table of category-aggregated metrics (Table 1) reveals interesting facts at a glance: Anime is the most popular category, but stories in Books and Plays receive more reviews per story on average, indicating, as has been shown in previous research [5, 9], more engagement and support of authors in those categories. Stories are significantly longer in the Books category than they are in any other category.

We extracted multiple visualizations from the data, only a small fraction of which are reproduced here due to space constraints. We found a classic power-law distribution in the word counts of stories with a peak under a thousand words. A matched-genre heat map table for the Harry Potter fandom (Table 2) reveals both understandable and intriguing correlations. It is not surprising that Romance contains more stories than any other genre, but it is intriguing that Western has the least. It is also interesting that Humor/Romance was the most popular combination of genres, comprising 7% of all Harry Potter stories out of 400 possible combinations.

CONCLUSION
With this work, we hope to make generally available a fully anonymized, rich dataset about a highly popular but understudied phenomenon. Even a cursory glance through the data uncovers fascinating details about differences among fandoms, varying proportions of stories between categories, how genre types vary by language, and more. We hope this dataset will pique researchers’ curiosity and spur further study of this topic.

ACKNOWLEDGMENTS
We thank the reviewers for their helpful comments.
REFERENCES


