Novelty and Success of Fanfictions

January 9, 2018

It has been believed that successful creative works strike a balance between novelty and familiarity. The classical inverted-U theory, also known as the Wundt-Berlyne curve, proposes that the perceiver's hedonic value first increases with novelty, until a certain threshold; afterwards, further increasing novelty will lead to a drop in hedonic value [1]. Many experiments have been performed in seeking for empirical evidence for this theory [8] [16]. However, it is still not clear where this assumed "maximum novelty" lies in the novelty – familiarity continuum.

The contemporary popular culture has nurtured many creative works that are adaptations, remakes, and "remixes" [13] of existing works. In particular, the movies and TV industries have produced many stories that are variations of previously successful stories. For example, the stories of Sherlock Holmes has had more than 25,000 adaptations [4]. The DC Universe has "rebooted" at least 3 times, and the Marvel Universe has more than 1,000 parallel universes [3], each telling a somewhat different story with a similar set of characters. As an extreme example, the origin story of the Spider-man has been re-played in three movies since 2002 [23]. Market reception indicates that audience welcome this kind of stories: out of the 10 highest-grossing movies of 2016, 8 are sequels, remakes, or part of a movie universe [21].

This trend is also reflected in a new type of creative works — fan works. Known formally as transformative works, they are creative works made by fans based on one or more original works ("canons"), and are often centered around certain characters or story lines[6]. For example, a story written by a contemporary fan about Sherlock Holmes in his retirement is considered a fan work. Although fan works contain multiple media types such as art, music and games, one of the most common type is creative writing—fanfictions. People interested in such activities often connect and interact with each other, forming communities known as fandoms[20]. Fandoms have developed into very large communities, especially in the Internet. The data source of our study, ArchiveOfOurOwn.org, is an online fanfiction archive that allows users to upload their fictions, and categorizes them based on fandoms. Established in 2009, it has 1,313,000 users and 3,423,000 pieces of fanfictions by November 2017 [14].

Fanfictions as a unique cultural phenomena has drawn attention from media studies and cultural studies [18]. However, most of the existing studies focus on the identity of fanfiction writers[2], the practice of fanfiction writing [12], and the interaction between fans [9]. Relatively less studies have focused on fanfictions themselves as creative works, especially using quantitative methods.

The authors and readers' passion about fanfictions alludes to a balance between novelty and familiarity: they desire to see familiar characters and elements, but in novel stories. By analyzing people's perception of fanfictions, we may gain a better understanding of this balance. The hedonic value of readers when reading fanfictions can be evaluated by how much they like a specific fanfiction. In our dataset, the readers can click "Kudos" (an equivalent of "likes") if they enjoy reading a fiction, thus the number of Kudos can be used as a metric for hedonic value.

Researches are only starting towards a quantitative definition of novelty. Many have employed a network approach, defining a creative work's novelty by looking at how they reference previous works [5][19][10]. In some cases, tags of a work can be used to evaluate its novelty[17]. To quantify the novelty of fictions, we employ three methods based on language modeling, topic modeling and tag analysis. Used widely in natural language processing and information retrieval[11][15], a language model represents documents as distributions over words. We can then compare documents by calculating the distance between these distributions. Topic models, on the other hand, represents a document as a distribution over topics. Therefore we construct an "average" fiction as the average of these probability distributions, and define novelty as "the distance to average": more specifically, a fiction is more novel if it has a large distance from the average of previous fictions, and vice versa.

While the text of fanfictions directly capture their contents, the tags generated by authors (see Methods) provides a summary of the fictions. Authors often use tags to describe the genre and plot of their fictions, for example, "Angst" or "Alternative Universe". Although a user is free to use any tag, when she starts to enter a tag, the website will prompt with tag suggestions, avoiding multiple synonyms. The tag set of a fiction provides an abstraction of its content. Besides the language modeling approach, we also ran a separate analysis on the tag set.

With our analysis, we show a negative relationship between novelty and success, indicating that fanfiction readers prefer more familiarity in general, discouraging novelty.

Results

We first present some descriptive statistics of our data.

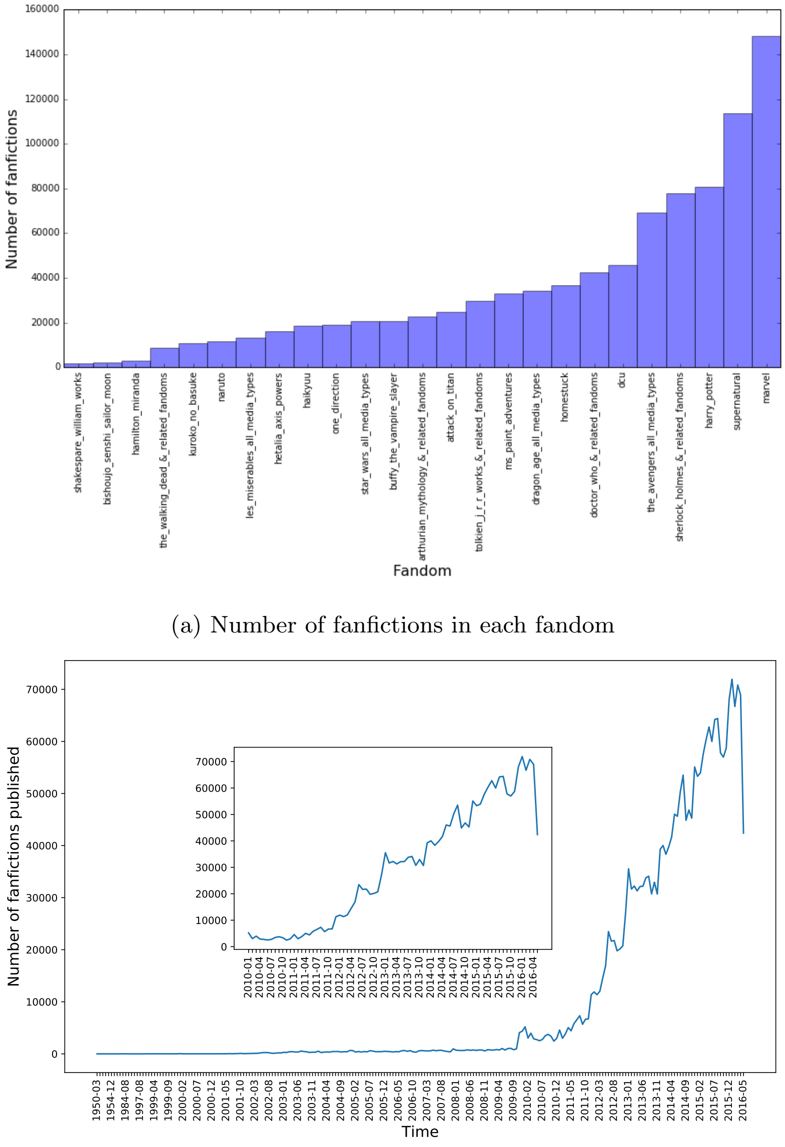
Statistics of the fanfiction dataset

Figure 1a shows the number of fictions in each of the 25 fandoms that we study. Our data collection includes fandoms in each of the broad categories based on medium: movies, TV shows, books, anime & manga, and musicals (see Methods). The fandoms with the most fictions are *Marvel*, *Supernatural*, and *Harry Potter*. Figure 1b shows the time distribution of fictions published. As AO3 was established in December 2009, the fictions earlier than this time might be migrated from other platforms, thus may not correctly reflect the online community. We therefore only run our analysis using the fictions published in January 2010 and later. After the establishment, the site first experienced a slow growth until around 2012, when the number of fictions published begin to rise radically. At its peak in early 2016, more than 70,000 fictions were published in a month.

The length distribution of the fanfictions is presented in Figure 2a . Since many fictions have multiple chapters, we consider length as the average number of words in each chapter. A great variation of average length can be observed. The variation of length in documents can have a significant effect on the distance between documents, biasing towards shorter distance between longer documents. A similar issue has been noticed in vector space models very early on [?], where many normalization methods have been proposed to correct it. Instead of applying a normalization, in our analysis, we sample a fixed length (1,000 words) from each document.

Figure 1b shows the logged cumulative distribution of Kudos. Following a long-tailed distribution, a small number of fictions receive the majority of Kudos, while most fictions receive little or no Kudos. Because of this pattern, we use the log of Kudos to evaluate the success of fictions.

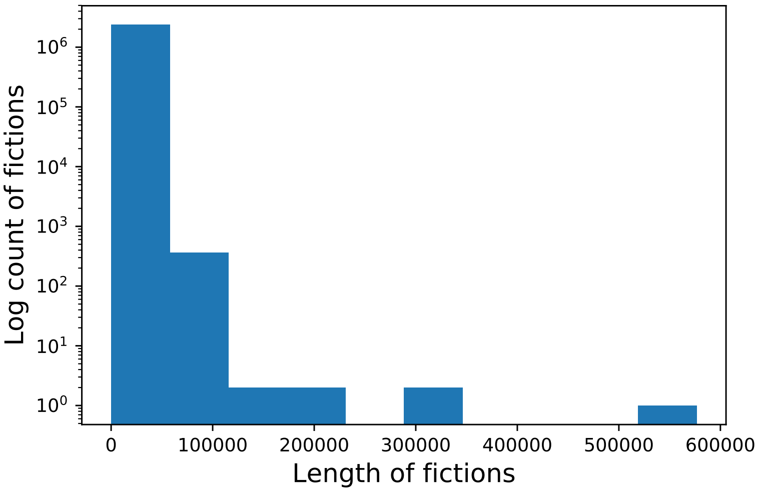
In Figure 3, we observe the tag set length of the fictions. The majority of fictions has under 20 tags, with a few having as many as 150 tags. Because the users have the liberty to create any tag, many of the tags will not be reused. To reduce the influence of infrequent tags, we therefore discard the tags that appear only once.



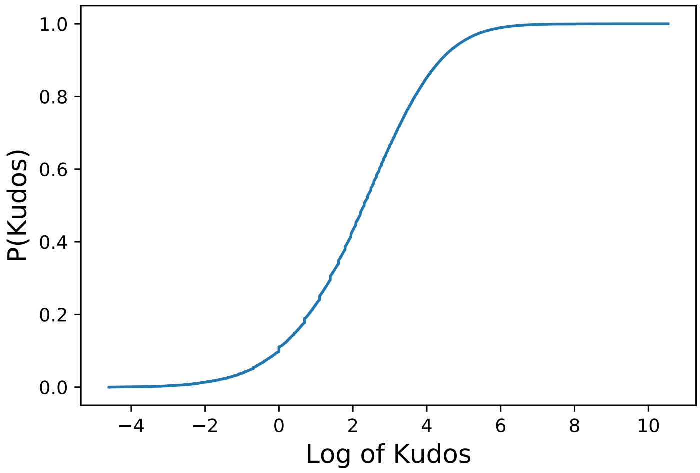
(a) Number of fanfictions in each fandom

(b) Number of fanfictions published each month

Figure 1: Statistics of the size and temporal distribution of our fanfiction dataset



(a) Length distribution of fictions. For multi-chapter fictions, we show the average length of each chapter.



(b) Linear-log cumulative distribution of Kudos. For multi-chapter fictions, we use the average number of Kudos per chapter. A long-tail distribution is observed, where a small portion of fictions receive the majority of Kudos.

Figure 2: Statistics on the length and Kudos of fictions

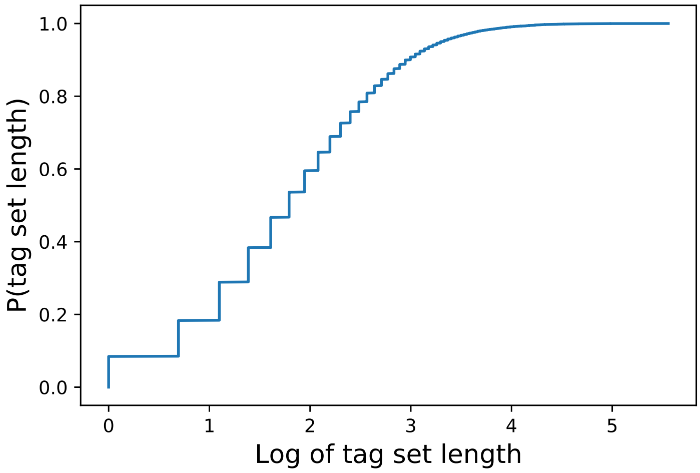


Figure 3: Linear-log cumulative distribution of tag set length.

Document novelty and success

We create two distinct measurements for the novelty of fanfiction documents, and employ them to identify the relationship between novelty and the number of Kudos that they receive. A negative correlation is observed, consistent across both measurements. Figure 4a shows the result from the unigram language model. Under this model, a fixed number of words (1,000 in our experiment) is sampled from each document in a bag-of-words approach, where the ordering of words is not preserved. We then calculate the probability distribution over the unigrams. To account for unseen unigrams, we apply the Simple Good-Turing smoothing method [7], which assigns a small non-zero probability to those unigrams.

The novelty score of each fiction is computed by comparing it to an “average" fiction. To define the “average" fiction, we take a set of fictions consisting of all fiction published within the past 6 months before the target fiction is published. Then we calculate the average probability distribution of the fictions in this set. We compute the novelty score of the fiction as the cosine distance between the target fiction and the average fiction:

$$n = 1 - \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

From Figure 4a, we observe that as the novelty score increases between 0 and 1, the z-score of Kudos steadily decrease from 0.1 to -0.3: that is, from above average to below average.

Besides the language model, we also present the results of running a topic modeling algorithm to detect the topics of fanfictions. The *Latent Dirichlet Allocation* (LDA) [**?**] has been widely used for such tasks. As an unsupervised learning algorithm, it models a topic as a probabilistic distribution over a set of words, and a document as a probabilistic distribution over a set of topics. We are thus able to compute the distance between documents by comparing their topic compositions. Similar to the unigram model, we construct an “average” fiction as the average topic distribution of fictions in a set consisting of fictions published during a given time period, and define the novelty score as the cosine distance between a fiction and the average fiction. Figure 4b shows the results of the LDA model. Consistent with the result of the unigram model, as novelty scores increase, the z-score of Kudos decrease from positive (above average) to negative (below average).

Tag novelty and success

We adapt the method from [17] to compute a novelty score for each tag. This method considers the “surprise" of observing a tag, given a set of tags from all fictions in the past. To control for the size of the size of this set, we use a limited length of past history, only considering the fictions published within 6 months of the target fiction. First, we compute the probability $P(t)$ of observing a tag t over a set of tags S:

$$P(t) = \frac{S_t}{S} \tag{2}$$

Where S_t is the size of the subset of fictions that includes the tag t, and S is the size of the whole set. Naturally, we can use $-logP(t)$ as a measurement of the surprise of tag t. The novelty of a fiction is then defined as the average surprise of its tags:

$$N = -\frac{1}{S_T} \sum_{t \in T} logP(t) \tag{3}$$

Where S_T is the number of tags of a fiction. Higher surprise of tags will therefore indicate more novelty. In Figure 5, we observe once again that as novelty increase, the z-score of Kudos experiences a fluctuating but overall steady decline.

Discussion

We have proposed three ways to quantify the novelty of fanfictions. We do this by modeling the texts of the fictions, and using document similarity measurements to calculate the distance between a fiction and the “average” fiction of its time and fandom; and also by considering the author-generated tags of the fictions, calculating the “surprise” of observing these tags. All methods have presented a negative correlation between a fanfiction’s novelty and the Kudos that it receives, indicating that more novelty leads to less hedonic value.

Our results show a pattern diverting from the Wundt-Berlyne curve: instead of a reversed-U shape, we observe the hedonic value consistently decreasing when novelty increases (Fig 4a, 4b, 5). This may be explained by the nature of fan works – because fans desire to see familiar characters and stories, it is reasonable for them to prefer fanfictions that stay close to the original works. Presumably, a similar psychology may explain the abundance of sequel and spin-off films in today’s film market. A possible extension of this work is therefore to investigate if the novelty-Kudos pattern holds for pop culture products other than fanfictions.

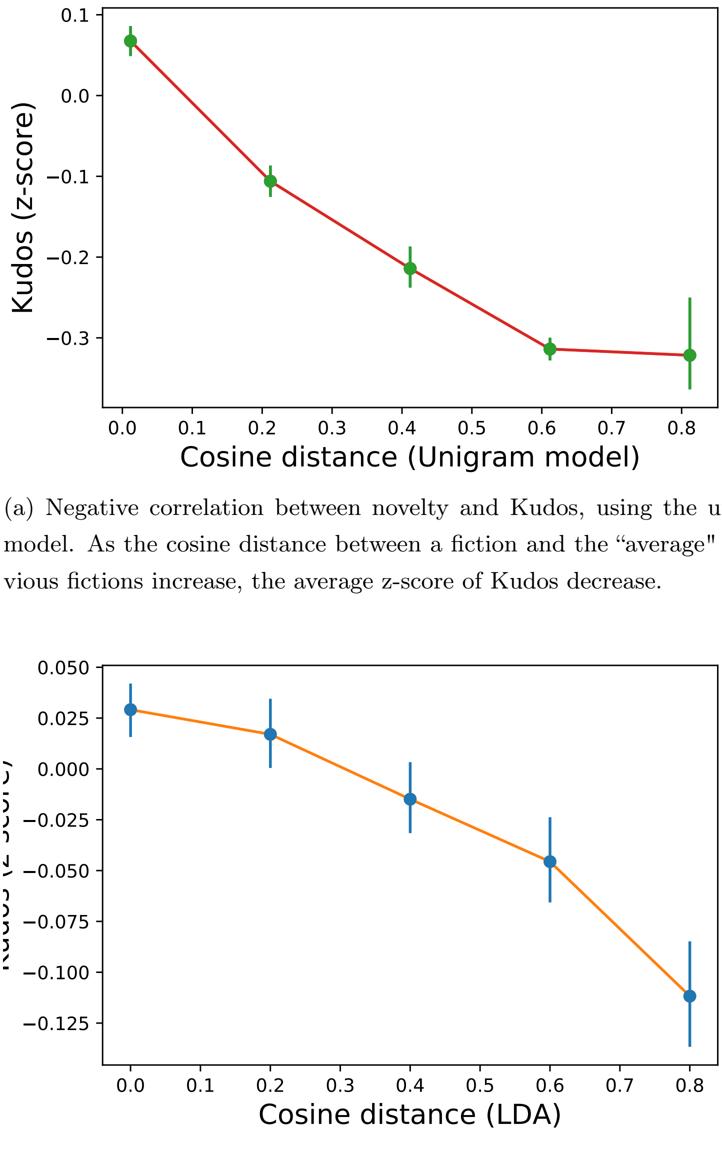
Methods

Data collection

We collected fanfictions from 50 fandoms on AO3, according to its list of most popular fandoms as of March 2016 [1]. To avoid duplication, we remove fandoms with heavy overlapping (e.g.: we keep *Marvel* and removed *Marvel movies*). Fandoms that cover diversed topics (e.g.*k-pop*) are also removed. Finally, we only kept the fictions written in English. This leaves us with 904,760 fictions from 25 fandoms. In our analysis, only samples of the data are used.

Besides the work texts, we also collected metadata including 23 fields. We only used information contained in some of these fields. Table 1 gives the names and descriptions of these fields.

[1] from this list: http://archiveofourown.org/media



(a) Negative correlation between novelty and Kudos, using the unigram model. As the cosine distance between a fiction and the "average" of previous fictions increase, the average z-score of Kudos decrease.

(b) Negative correlation between novelty and Kudos, using topic modeling. The cosine distance is calculated between the topic distribution of a fiction and the average of topic distribution of previous fictions. As cosine distance increases, the average z-score of Kudos decrease.

Figure 4: Relation between novelty and Kudos on a document level

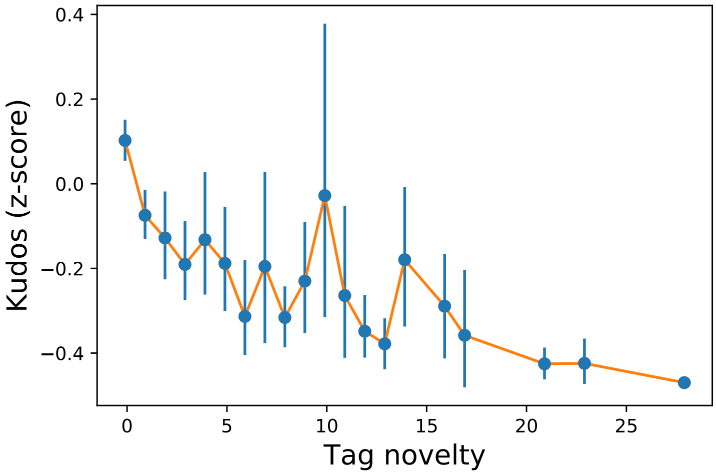


Figure 5: Tag novelty and Kudos.

Table 1: Metadata of the fictions

Fields	Description	Usage
Text	The fiction texts.	All text analysis are carried out on these texts.
Title	Titles of the fictions.	Used to identify the fictions.
Fandoms	Describes which fandom(s) the fiction belongs to.	Used to categorize the fictions.
Author	The author of the fiction.	Used for identifying the fictions and for text analysis.
Kudos	The number of “likes" that the fiction receives.	Used for evaluating the fiction’s success.
Publish Date	The date the fiction was published.	Used for temporal analysis.

Language model

We model the fictions with a unigram language model. The Simple Good-Turing smoothing[7] is applied to improve the model, and to assign non-zero probabilities to previously unseen unigrams. When creating the set of unigrams, we also remove the rare unigrams that appear in less than 5 documents and left out fictions with less than 500 words.

Topic modeling

The Python library Gensim's LDA [?] is used to train topic models on our data. We set the number of topics to 40 and use default values for other parameters.

Computing tag novelty

References

[1] Berlyne, D. E. Novelty, complexity, and hedonic value. *Attention, Perception, & Psychophysics 8*, 5 (1970), 279–286.

[2] Black, R. W. Language, culture, and identity in online fanfiction. *E-learning and Digital Media 3*, 2 (2006), 170–184.

[3] Database, M. Multiverse/universe listing, 2017.

[4] Doyle, A. C. *The New Annotated Sherlock Holmes: The Complete Short Stories: The Adventures of Sherlock Holmes and The Memoirs of Sherlock Holmes (Non-slipcased edition)(Vol. 1)(The Annotated Books).* WW Norton & Company, 2007.

[5] Elgammal, A., and Saleh, B. Quantifying creativity in art networks. *arXiv preprint arXiv:1506.00711* (2015).

[6] Fanlore. Transformative work, 2015. [Online; accessed 14-December-2015].

[7] Gales, W. A. Good-turing smoothing without tears. *Journal of Quantitative Linguistics* (1995).

[8] Hargreaves, D. J. The effects of repetition on liking for music. *Journal of research in Music Education 32*, 1 (1984), 35–47.

[9] Hills, M. The expertise of digital fandom as a 'community of practice' exploring the narrative universe of doctor who. *Convergence 21*, 3 (2015), 360–374.

[10] Iacopini, I., Milojević, S., and Latora, V. Network dynamics of innovation processes. *ArXiv e-prints* (July 2017).

[11] Jurafsky, D. *Speech & language processing.* Pearson Education India, 2000.

[12] Magnifico, A. M., Curwood, J. S., and Lammers, J. C. Words on the screen: broadening analyses of interactions among fanfiction writers and reviewers. *Literacy 49*, 3 (2015), 158–166. LIT-OA-2015-003.R1.

[13] Manovich, L. What comes after remix. *Remix Theory 10* (2007), 2013.

[14] of Our Own, A. Archive of our own, 2017.

[15] PONTE, J. M., AND CROFT, W. B. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1998), SIGIR '98, ACM, pp. 275–281.

[16] SLUCKIN, W., COLMAN, A. M., AND HARGREAVES, D. J. Liking words as a function of the experienced frequency of their occurrence. *British Journal of Psychology 71*, 1 (1980), 163–169.

[17] SREENIVASAN, S. Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. *Scientific reports 3* (2013).

[18] THOMAS, B. What is fanfiction and why are people saying such nice things about it? *Storyworlds: A Journal of Narrative Studies 3*, 1 (2011), 1–24.

[19] WANG, D., SONG, C., AND BARABÁSI, A.-L. Quantifying long-term scientific impact. *Science 342*, 6154 (2013), 127–132.

[20] WIKIPEDIA. Fandom— Wikipedia, the free encyclopedia, 2015. [Online; accessed 14-December-2015].

[21] WIKIPEDIA. 2016 in film, 2017.

[22] WIKIPEDIA. Fanfiction.net, 2017.

[23] WIKIPEDIA. Spider-man in film, 2017.

[24] YUNG, H. Market transformation in transformative works: The effects of introducing incentives in markets for fanfiction.

[25] ZHAO, A. Predicting popularity of fanfiction stories based on title and summary.