Comic Book Movies Affecting Comic Book Sales

Final Report for PHY 408

Jack Farrell
1003978840

April 22, 2021

Contents

1 Introduction

In the last two decades, movies based on comic books have become some of the most popular in Hollywood. On the other hand, due in part to a speculative boom leading to a crash[1] and in part due to the emergence of digital media, the sales of comic books themselves have declined from their peak values in the 1990s. In theory, the enormous success of the super-hero genre on the big screen should translate to increased interest in the comic book source material. Determining the extent to which this link exists is the focus of our project.

[1] Very interesting — see *e.g.* [3]

In this report, we use Fourier Transforms and the *cross-correlation* of two time series to analyze the total monthly earnings of comic books and comic book movies and understand their relationship. We limit our focus to movies in the *Marvel Cinematic Universe (MCU)*, a collection of films featuring comic book characters like *Iron Man*, *Thor*, *Captain America*, and *Spider-Man*. These projects feature somewhat faithful adaptations of comic-book characters and stories, and they also represent some of the most profitable and successful comic-book movies, with over 22 Billion (USD) box-office earnings worldwide [1].

Our analysis is guided by the question: **To what extent do the releases of MCU films affect the sales of Marvel comic-books?** We hypothesize the following: on one hand, on a broad scale, the prevalence of Marvel adaptations during the time period of the MCU's activity should cause a general upward trend of Marvel comic-book sales. Additionally, though, fluctuations in comic-book sales around the broad trends should be correlated with the release of MCU movies.

In terms of the time series, the above question is really about the similarity of two signals. But since correlation does not imply causation, our results (if any) will really speak only to the *consistency* of our hypotheses with the data — we will not be able to verify or falsify it.

2 Analysis

2.1 Data Sources

This project takes data from two sources. For the comic book sales figures, we pull from "Comichron" [2], an internet database featuring monthly comic book sales figures. For the movie earnings, we pull from "Box Office Mojo" [1], an internet database giving various data related to movies, including box-office earnings over time.

2.2 Raw Data

We focus on the *total* earnings of all comics published by Marvel and the *total* earnings of all movies in the MCU (domestically: USA) — in other words, we analyze the general relationships between Marvel movies and Marvel comics holistically. The reason for this choice is that the sales of individual series of comics (*e.g. The Amazing Spider-Man*) are volatile and change with many factors including the writer and artist. On the other hand, we expect the publisher-wide sales figures to depend on less variables. Our datasets cover all months between Jan. 2008 and Dec. 2019, roughly bookending the MCU (so far!).

Fig. (1) plots the total Marvel Comics earnings over the MCU's history (again, 2008–2019 inclusive). As we might expect, the MCU dataset (purple trace, right panel) shows a constant zero signal interrupted by sharp peaks corresponding to the release of movies. On the other hand, the comics data shows a broad complicated trend with superimposed high-frequency fluctuations. Our first step will be to filter the data to convince ourselves that there really is a trend.

2.3 Filtering

To make the trend more clear, we filter the data. First, we estimate the trend using a least-squares fit to a degree 5 polynomial. We used the `numpy.polyfit()` function. Then, we subtract the

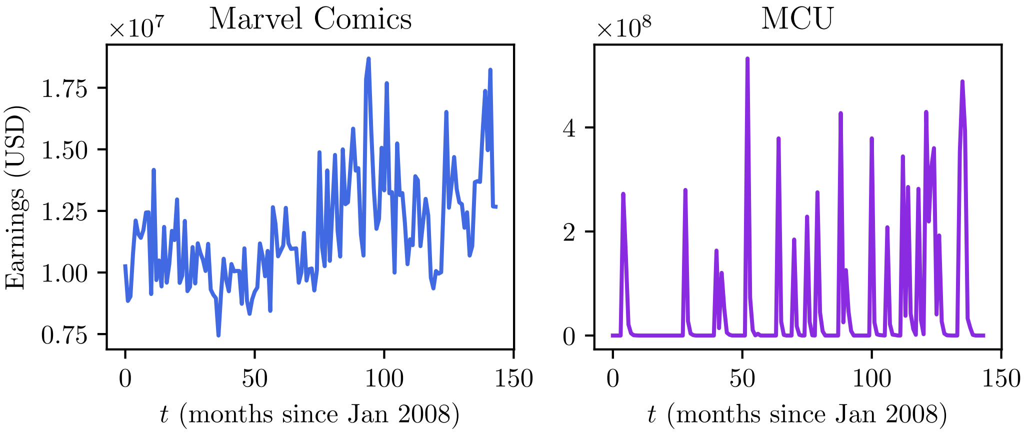

Figure 1: *Left panel* Total Marvel Comics earnings as a function of time t in months since Jan. 2008. *Right panel* Total Marvel Cinematic Universe domestic box office earnings as a function of time t in months since Jan. 2008.

trend, leaving just the fluctuations around the estimated trend. We give the raw data, trendline, and detrended data in Fig. (2). Then, we filter the data using the *Fourier Transform* of our time series

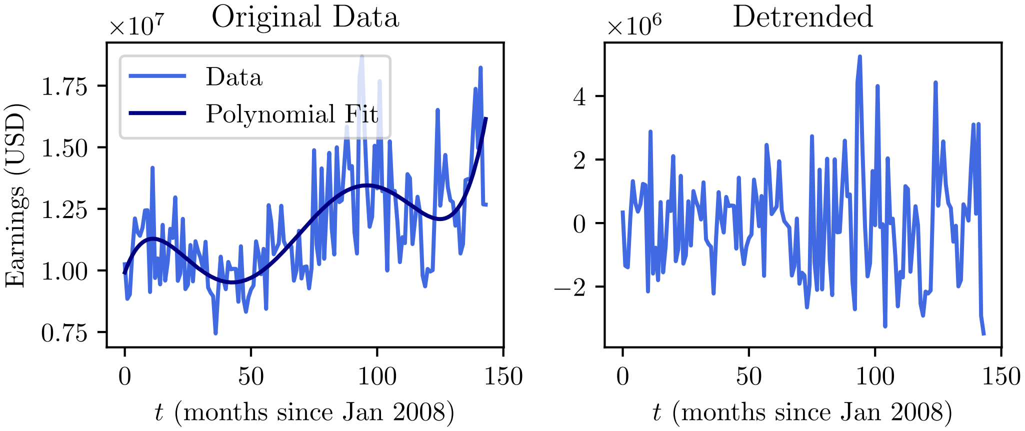

Figure 2: *Left panel* Raw comic book earnings and a degree 5 polynomial fit. *Right panel* Subtracting the polynomial trendline from the raw data to obtain a 'detrended' time series.

f, defined by:

$$F_k = \Delta t \sum_{j=0}^{N-1} f_j \mathrm{e}^{-\mathrm{i}2\pi jk}, \tag{1}$$

where N is the length of the time series. We plot the Fourier Transform in Fig. (3). Based on the amplitude of the peaks, which give roughly the 'component' of the signal that has the corresponding frequency, we find the highest amplitude peaks have frequencies lower than 0.1 month^{-1} in absolute value. As such, to smooth out the data, we set all Fourier components corresponding to frequencies

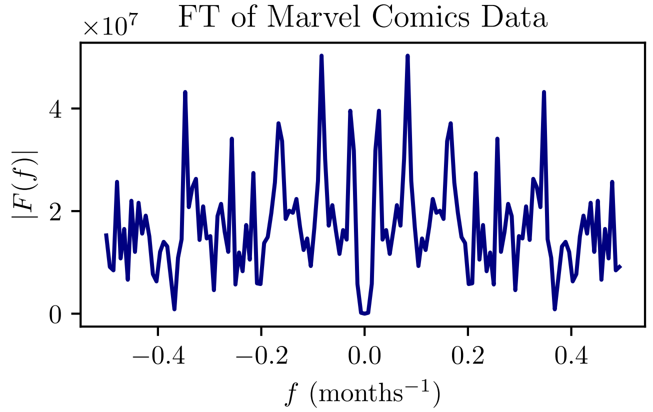

Figure 3: Amplitude of the fourier transform of the Marvel Comics Earnings time series as a function of frequency f.

greater than that value to zero, effectively truncating our Fourier Transform. Taking an inverse Fourier Transform and adding back in the polynomial trend, we obtain the smoother trace shown in Fig. (4). As we expected because of the polynomial fit (and from eyeballing the raw data), there is

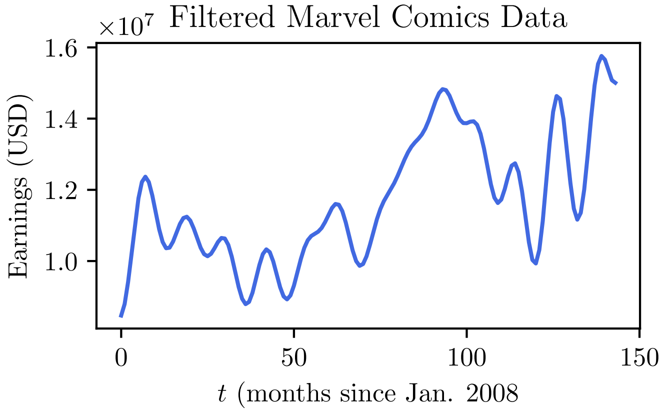

Figure 4: Result of filtering the Marvel Comics data by truncating the Fourier Transform

a general upward trend over the course of our roughly ten-year window of analysis. Interestingly, the increase seems to correspond with the *number* of MCU releases, but we do not elaborate on the quantitative aspects of this relationship in this report, preferring instead the correlation analysis of the following subsection, which makes a different point.

2.4 Correlation

Having considered (briefly) the broad affects of increasing MCU releases on Marvel comics data in the form of showing the explicit trend in the latter, we now turn to studying *fluctuations* around the general trend. To understand how these oscillations relate to the MCU signal, the tool we use is the *cross-correlation* between time series f and g:

$$C_{fg}(\tau) = \int_{-\infty}^{\infty} \mathrm{d}\tau f^*(t) g(t+\tau). \tag{2}$$

Which, roughly, gives the correlation between the two time series when g is shifted by an amount τ. Peaks in C_{fg} when it is graphed as a function of τ give the extent to which g is related to f when g is shifted by an amount τ.

We first subtracted the data just as we had done for Fig. (2). We also normalized the resulting detrended data by dividing by the maximum (absolute value) entry. We do the same for the MCU earnings data: making the two series more comparable in magnitude gives more reasonable values for the cross-correlation.

Then, with the two normalized datasets (and the comics series detrended), we compute the cross-correlation. Fig. (5) gives the cross-correlation between the MCU and Marvel Comics earnings time series. The $\tau-$axis is arranged a positive value of τ corresponds to shifting the comics time series (in the positive direction) by an amount τ.

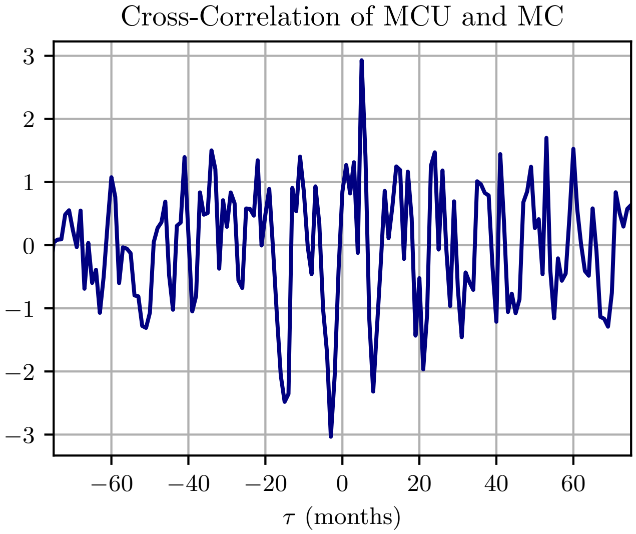

Figure 5: Cross-correlation of the MCU earnings dataset and the Marvel Comics (MC) dataset as a function of τ, the shift between the time series. Peaks give region of high correlation.

3 Discussion

3.1 Results

Our results have some implications in terms of answering our guiding questions. For one, the trend displayed in the filtered data of Fig. (4) shows a clear increase over time, which is at least consistent with our expectations. More importantly, though, as we noted in the previous section, the clear upward trend revealed by the filtering process motivates studying the fluctuations around that trend. But the results of our correlation study (Fig. (5) are messy and, in some ways, unconvincing. Still, we can extract some information from the plot. For instance, we do notice one peak that stands out at $\tau \approx 6$ months. This peak suggests that the comic book data around 6 months *earlier* is related to the MCU data, which makes sense; for big-budget movies like those in the MCU, the advertising campaign for the films would start around this time, including the release of several 'trailers' and television spots. As such, our results could suggest that it is the beginning of this process that affects comic book sales more than the actual releases of the films. There is also a peak at $\tau = 0$, corresponding to the release month of the films, but it does not stand out from the background noise in the plot.

Of course, correlation famously does *not* imply causation. At very least, though, we can say that the correlation data described above is *consistent* with our hypothesis that MCU movies affect Marvel Comics sales.

3.2 Control

To make our claim more convincing, it helps to have a 'control' dataset. If our hypothesis is correct, the sales figures for publishers outher than Marvel should not display the same behaviour at $\tau \approx 6$ months. We choose "IDW Publishing" — while "DC Comics" may be a more obvious choice, it would not be a proper control given that its comics fall into the same genre as Marvel's. Performing the same analysis described above on the monthly sales figures for IDW gives the results shown in Fig. (6). Here, the noise is comparable even to the peak near $\tau \approx 5$, and there are even some stronger peaks near $\tau \approx -55$. As such, we claim that the correlation peak we discussed in Sec. (2.4) is "real": it really does represent a fact about the affect of comic book movie advertising on comic book sales.

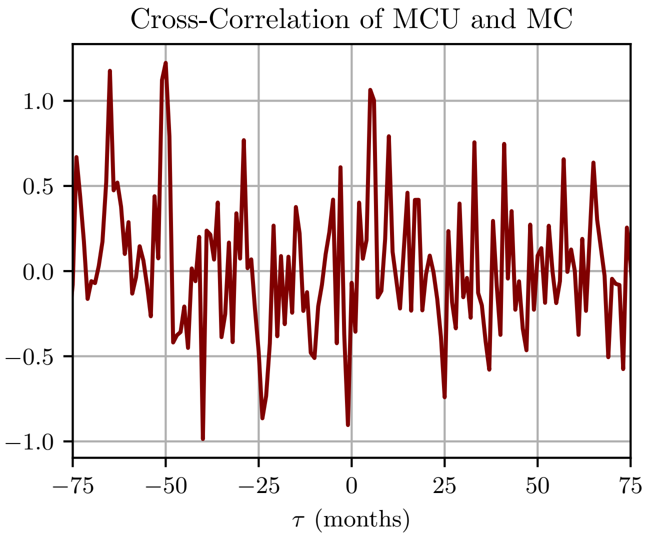

Figure 6: Cross-correlation of the MCU earnings dataset and the Marvel Comics IDW Comics earnings dataset as a function of τ, the shift between the time series. Peaks give region of high correlation.

3.3 Sources of Error

The analyses described in the previous sections are subject to several errors. To start, the noisiness of the data means that the peaks we identify in the cross-correlation are challenging to distinguish from the background. While our analysis of the 'control' dataset in the previous section makes the claims more plausible, the correlation is still not strong.

Secondly, that the time axis is only accurate to the nearest month gives us a resolution problem. The availability of higher-resolution (daily) comics earnings figures could show us more convincing trends, since it might be that the 'spike' in comic book sales lasts for less than a month, meaning it would get lost in our monthly graphs.

Of course, one additional major source of error is the number of variables that affect comic book data — this analysis does not consider factors like seasonal oscillations (summer is a high-earning month for comic books), inflation etc. A more sophisticated analysis could certainly attempt to model these variables; the annual oscillation, in particular, could be removed by way of a notch filter.

4 Conclusion

The analysis described in this report does not quite confirm or deny our hypothesis nor give a convincing answer to our guiding question about the effects of comic book movies on comic book sales. Still, we found two pieces of plausible evidence that certainly motivate future investigations into this topic. First, in Sec. (2.3), we found that filtering the monthly comic book sales data revealed a strong visual connection between the broad increase in comic book sales over our ten-year window of study and the release of MCU films over the same period. Even more convincingly, in Sec. (2.4), using techniques involving the cross-correlation, we obtain results suggesting that the lead-up to releasing MCU films (advertising campaigns, plausibly) could be measurably related to comic book sales.

The outcomes of this study give us several ideas for future projects, and two main directions stand out. First, it would be interesting to perform a similar analysis using *all* comic films compared to *all* super-hero comic books, instead of focusing just on Marvel — we might find an even stronger correlation. Secondly, we could take the opposite approach, analyzing the sales figures of individual comic-book series compared with specific film franchises (*e.g.* Avengers comic book v.s. Avengers film series). For popular comic book series, the correlation would probably not be strong (readership may 'saturate' at some value). But for smaller characters who are given movie franchises, (*e.g.* Doctor Strange, maybe), we might expect the movie to drive interest in the comics.

Overall, this project shows an interesting application of the Fourier, filtering, and correlation techniques we studied this semester. The question of how big-screen blockbusters affect the source material is, of course, more general than comic books and an import question in its own right.

References

[1] Box office mojo. https://www.boxofficemojo.com/.

[2] Comichron. https://www.comichron.com/.

[3] Jonathan V. Last. The crash of 1993. https://www.washingtonexaminer.com/weekly-standard/the-crash-of-1993, 2011. *The Washington Examiner*.

A Supplemental

A.1 Code

I guess it's not necessary for this assignment, but, if you're interested in the code or data, I'm happy for you to access it at: https://github.com/jfarrellhfx/final-report-phy408

A.2 Preprocessing

The process of obtaining the data in the first place and 'wrangling' it to get the total earnings v.s. time datasets we wanted was actually one of the most complicated aspects of this project, but it is not relevant to the report. Since I couldn't find the datasets I wanted online, I had to 'scrape' the data myself from the two source websites I've listed. I'd never done ths before, so it was actually a great learning experience.

If interested, the script I wrote to 'scrape' the online databases is `get_data.py`, which calls the functions we define in `scrapers.py`. Please note that this step requires a couple of extra Python packages not used in the course. The data wrangling and analysis steps are described in the notebook `analysis.ipynb`