SONG POPULARITY AS A CONTAGIOUS PROCESS

SONG POPULARITY AS A CONTAGIOUS PROCESS IN GREAT BRITAIN

By

DORA P. ROSATI, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

of Master of Science

McMaster University

MASTER OF SCIENCE (2015) McMaster University
(Mathematics) Hamilton, Ontario

TITLE:	SONG POPULARITY AS A CONTAGIOUS PROCESS IN GREAT BRITAIN
AUTHOR:	Dora P. Rosati, B.Sc. (McMaster University)
SUPERVISOR:	Dr. David J.D. Earn
NUMBER OF PAGES:	ix, 63

Abstract

Determining the mechanisms by which a song becomes popular is a complex problem. The rapidity with which some songs gain popularity often leads to them being described as 'contagious' or 'infectious'. Upon closer examination, we find that the download time series for many popular songs do resemble epidemic curves derived from case report data for infectious diseases. This correspondence suggests an interesting link between the fields of infectious disease research and music research: perhaps ideas from epidemiological modelling might be useful in investigating how certain songs 'spread' through human populations, and perhaps employing disease epidemic models might help to better understand the mechanisms underlying song popularity. Download data were obtained from MixRadio based on song downloads through Nokia cell phones in Great Britain over a seven year period and aggregated at various timescales. Songs were characterized by fitting a standard epidemic model to song download time series. The fits estimate standard epidemiological parameter values for each song, providing new insights about popularity of music. In addition, we propose and analyze a new model that is better suited to 'song transmission' and comment on how this model might be used to study song popularity in the future.

Acknowledgments

I would like to thank my supervisor Dr. David Earn for all of his guidance, suggestions and contributions. His insightful advice in the preparation of this thesis and throughout the duration of my degree has been invaluable. I would also like to thank Dr. Ben Bolker for his time and expertise spent helping with various fitting issues, and Dr. Matthew Woolhouse and his research group for feedback and for insightful discussions about factors affecting song download trends. I am grateful to the members of Dr. Earn's research group for useful discussions and suggestions, and to Cody Koykka for helpful edits and comments.

I would like to most sincerely thank my parents and my brother Mark for their continual belief in me and for their unending love and support. I am also grateful to Amy Tapley, for always being on my side, to Niel Van Engelen, for his patience, support and encouragement, and to all of the friends and family who have been encouraging or interested in my work during this process.

Contents

List of Figures

List of Tables

Declaration

I hereby declare that the work presented here is my own, completed under the supervision of Dr. David Earn, and that it is, to the best of my knowledge, original. All sources of data and software have been acknowledged within the body of the thesis.

1 Introduction

Music is ubiquitous in society; everyone listens to it and most people have certain styles that they prefer. As a result, there exists an enormous variety of music and a vast number of songs for listeners to choose from. In spite of this, there is a relatively small selection of songs that most people are able to recognize at a given time. It is quite remarkable given this huge variety and number of songs that so few of them become enormously popular. How does a song become popular and how is it that certain songs become so much more popular than others? What are the underlying social mechanisms that drive these processes?

Many parallels can be drawn between the outbreak of an infectious disease and the release of a new hit song. An epidemic of an infectious disease can have a huge effect on a population. The disease sweeps through the population, passing from person to person with various social interactions facilitating its transmission. Eventually it reaches some peak prevalence and dies down as the susceptible pool is exhausted and/or the population begins to recover. Often by the time the epidemic is over, a large proportion of the population will have been infected with the disease. There is a period after the release of a hit song during which the song 'spreads' rapidly through the population, from person to person and through the media, until it reaches some peak popularity and its appeal gradually diminishes. For hit songs, by the time their period of extreme popularity is over, a large proportion of the population will have spent time listening to them.

Is it possible that the same social processes that allow for an infectious disease to spread through a population also play a role in how songs become popular? A popular song will often be referred to as 'catchy', as if, like the flu or measles, it could be caught. Perhaps there is more truth to this terminology than has been previously recognized. The time series for song download data presented in this thesis are similar in shape to time series for infectious diseases. This resemblance suggests that it is possible that there are social mechanisms underlying song popularity that are similar to the social mechanisms that drive the spread of an infectious disease, and has acted as our motivation to use standard epidemiological models to study how songs become popular. We seek to elucidate the underlying social processes that drive song popularity and to investigate how they are similar to and different from disease transmission.

2 Background

2.1 Song Popularity Research

Song popularity is a topic of great interest and has been the subject of much research. While it is possible that there are specific musical characteristics that can be considered popularity indicators, there are also underlying social processes that affect how a song gains popularity. Disentangling the influence of musical and social aspects on a song's popularity is a difficult task.

Previous research has found both support for [14, 35] and evidence against [36] the idea that musical features of a song can predict a song's popularity. Nunes and Ordanini [35] used audio information to show that songs that were number 1 hits on the Billboard Hot 100 Charts had distinctly different instrumentation than songs that never climbed above the 90 position on these charts, and Dhanaraj and Logan [14] found that audio and lyric information about a song could each be used to generate better than random predictions about whether or not a song would be a hit. However, Pachet and Roy [36] found that it was not yet possible to predict a song's popularity based on audio information about the song, regardless of whether this information was derived from an audio signal or from human input.

Several studies [8, 25, 38] have found that information from social media sites, social music sites or Peer-to-Peer networks can effectively be used to predict song popularity, which hints at the underlying social processes that may drive song popularity. Bischoff *et al.* [8] built a model that predicted song popularity based on various *Last.fm* tags relating to user listening habits and previous popularity of the artist in question. Schedl *et al.* [38] used *Last.fm* play count data to predict popularity of artists in specific countries. They compared this method with predicting artist popularity based on a) user posts from *Twitter*, b) information from shared folders in *Gnutella*, and c) the number of pages returned by search engines that were related to an artist in a specific country. Kim *et al.* [25] also looked at *Twitter* posts as a predictor of song popularity and found that hashtags that related to music listening behaviour of users could be used to forecast rankings of songs on Billboard charts.

There has also been neural imaging work done that examined the influence of a song's overall popularity on adolescents' rankings of that song [6]. The functional magnetic resonance imaging (fMRI) data collected in this study suggested that teenagers are more likely to change their evaluation of a song to more closely align with its overall popularity rating as a result of the anxiety created by a difference between their opinion and the opinion of others. It was also found that neural activity in specific regions of the brain while listening to songs significantly corre-

lated with sales data for that song over the next three years, even though subjective ratings of the songs from participants did not [7].

2.2 Epidemiological Modelling

The model used to investigate song download data in this study was the susceptible-infectious-recovered (SIR) model. This is a standard epidemiological model that is commonly used to track the spread of infectious disease [2, 11, 21]. It is a compartmental model, meaning that the population being studied is divided into several compartments depending on their infection status (see Figure 1). Each individual is either 'susceptible' to the disease in question, 'infectious' or 'recovered' (hence SIR), at which point it is assumed that they cannot become reinfected. The rate at which individuals move between these three compartments is represented by the set of ordinary differential equations:

$$\frac{dS}{dt} = -\beta SI \tag{1a}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{1b}$$

$$\frac{dR}{dt} = \gamma I \tag{1c}$$

where β is the transmission rate and γ is the recovery rate for the disease being modelled. Two more easily interpretable parameters that can be studied from this model are the mean infectious period, given by $1/\gamma$, and the basic reproduction number, given by $\mathcal{R}_0 = \beta/\gamma$. The basic reproduction number tells us in a wholly susceptible population, how many individuals on average would be infected by one infectious individual. The final size Z of an epidemic can also be calculated based on $\mathcal{R}_0$ [24, 31]. This tells us what proportion of the initially susceptible population will have been infected at some point during the epidemic by the time the epidemic is over. The formula for final size Z is [24]:

$$Z = 1 - e^{-\mathcal{R}_0 Z} \tag{2}$$

This is not the first time epidemiological models have been considered as a tool for studying popular songs. A similar idea was employed by Tweedle and Smith? [40], who studied the effects of positive and negative media attention on 'Bieber Fever'. However, while they were working with epidemiological models, their study was entirely theoretical – they did not apply the ideas to any data. The focus of the

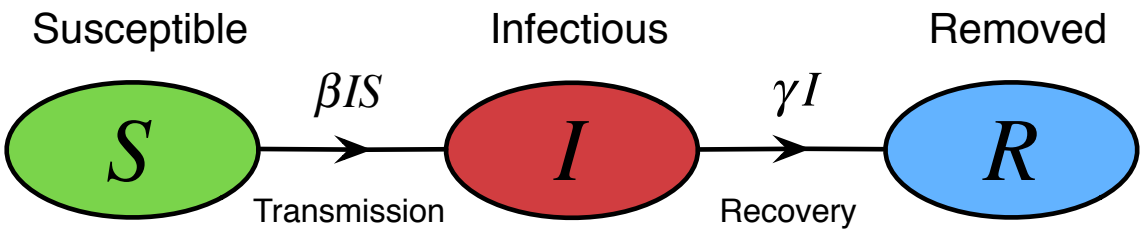


Figure 1: A flow chart representing how the SIR model tracks movement of individuals among the three disease state compartments. Note that the third class can be labelled as 'Recovered' or 'Removed'. The 'Removed' label is used to indicate that individuals are no longer part of the transmission process, whether this is because they have recovered or died or been removed from the system in some other way – from a modelling perspective the distinction is not significant.

study was also on the excessive popularity of an individual artist with a specific demographic, making it a different phenomenon from that being studied here.

3 Description of the Data

The data used here were obtained through a data sharing agreement with MixRadio. The database contains information on nearly 1.4 billion individual downloads through Nokia cell phones in 33 countries between 2007 and 2014, including track titles, artist names and artist genre classifications. Minute-by-minute download counts can be extracted for individual tracks. The database also contains various metadata about users, such as a user ID, total number of downloads, user country, and user start and end dates. A field has also been added to classify each user as a certain type of 'XHead' where X is the genre from which the user has downloaded the most tracks (not necessarily $> 50\%$ of their total downloads) [3]. The size of the database and type of data that it contains make it an excellent tool for studying various cultural and social questions relating to music popularity. Some exploratory work in these areas has already been done [43].

We chose to focus our investigation on the top 1000 downloaded songs in Great Britain (GB). The list of top 1000 downloaded songs was determined by considering downloads by all users in GB between 2007 and 2014. This gave a sizeable amount of download data to analyze (the database contains information on 60 221 294 downloads in GB by 552 784 users from 63 genres), and focusing on one country eliminated the issue of different countries adopting the MixRadio service at different times. After further narrowing down our song data set as described below in §4.1, we were left with 542 songs (when considering all users in GB). Table 1 shows how many songs

Genre	No. Songs
Country and Western	1
Reggae	4
Electronica	6
Metal	10
Soul/R&B/Funk	47
Indie/Alternative	48
Rap/Hip Hop	60
Dance	72
Rock	85
Pop	209

Table 1: The number of songs in each genre that appear in the song data set.

from each genre appear in our song data set.

3.1 Some Descriptive User Statistics

Signing up for the MixRadio service entitled a user to one year of unlimited free music downloads. After this period, the user could continue to download music for 77–99 pence per track, with prices being set by the artist. The database contains information on how many downloads each user in GB made in total. There are some users in GB who have download counts so high that it is likely they were downloading as much free music as possible (possibly using a computer algorithm of some kind) rather than downloading tracks on a preference basis. Downloads by such extreme users are not representative of song transmission, meaning users who downloaded more than 5000 tracks could be disregarded in our analysis. However, as Figure 2 shows, the proportion of users in GB who made more than 5000 downloads is very small. This means it can also likely be assumed that the contribution of downloads by users with more than 5000 downloads to a single song's popularity is negligible. There are also users who have made a very small number of downloads. Most noticeably, as Figures 3 and 4 show, many users in GB only downloaded a single track. It could be argued that users who made very few downloads were only trying the MixRadio service on a whim and that their downloads are therefore caused by mechanisms other than those that drive song popularity. However it could also be argued that even if users with few downloads were only looking to try out the MixRadio service, they would still download a song that they had been in some way influenced to

like. Conversely, it could be that they would simply choose to download one of their favourite songs, rather than a song that was currently popular. Based on this reasoning, we decided not only to look at the trends present in all downloads of the songs in our data set, but also the trends present when considering only downloads made by the subset of users who have made between 5 and 5000 downloads. It is unclear whether any differences should be expected in this population. The top 1000 songs downloaded by this subset of users has 37 different songs when compared to the list of top 1000 songs downloaded by all users in GB. We decided that this difference was small enough to be ignored and therefore used the list of top 1000 songs downloaded by all users throughout this work.

We can also examine the number of downloads that users in GB made by genre. Figure 5a shows the number of downloads by genre when considering all users in GB, while Figure 5b shows the number of downloads by genre when only considering users who have made between 5 and 5000 downloads.

As mentioned above, each user in the database is labelled as an XHead, where X is the genre from which they have downloaded the greatest number of tracks. Figure 6 shows the number of XHeads by genre X in GB when considering all users, and when considering only users who have made between 5 and 5000 downloads. These plots also show only the ten genres that appear in our song data set.

Since we are using disease epidemic models to study song popularity, it would be useful to have a measure of user 'lifespan', *i.e.,* the time for which they used the MixRadio service and were therefore 'part of the population'. Unfortunately, we do not have the dates on which users signed up for and stopped using the MixRadio service but rather the dates of their first and last downloads. These are referred to in the database as user start and end dates. Since users may not have downloaded a track on the same day that they signed up for the MixRadio service, and similarly may not have stopped using the service on the same day that they downloaded their last track, the user start and end dates can only be used to estimate a lower bound on user lifespans. We therefore refer to this lower bound as 'pseudo-lifespan'. The distribution of user pseudo-lifespan is shown in Figure 7. It is worth noting that users who only downloaded one track will have a pseudo-lifespan of 0. The 164 089 users with a pseudo-lifespan of 0 are not shown in Figure 7.

3.2 Some Descriptive Genre Statistics

There are four descriptive genre parameters that can be compared for the genres that appear in our song data set. These are:

1. the number of XHeads in GB from each genre X,

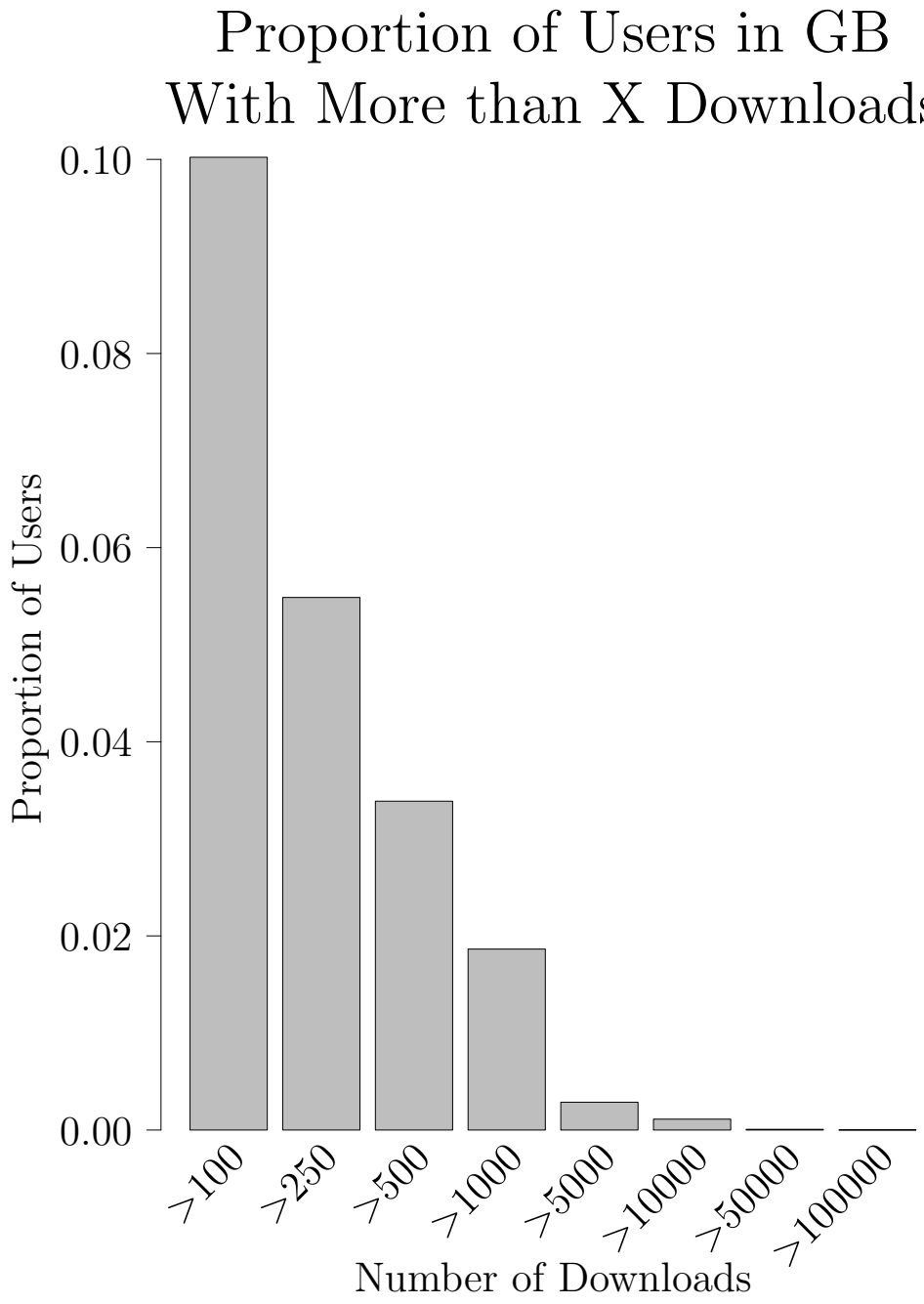


Figure 2: The proportion of users in GB who have made more than various numbers of downloads.

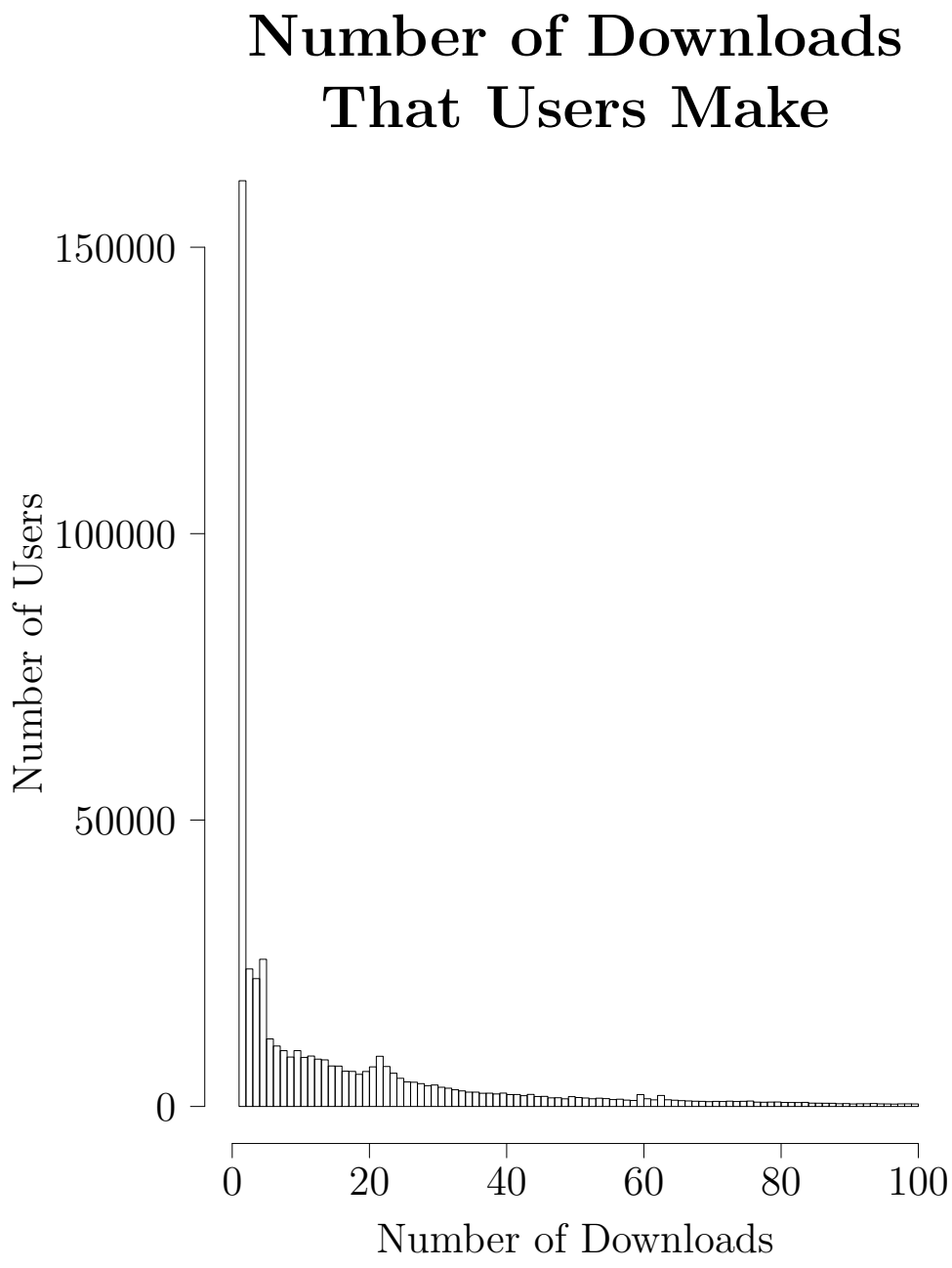


Figure 3: The number of users in GB who have made various numbers of downloads (bin width is 1). The same information can be seen plotted on a log scale in Figure 4.

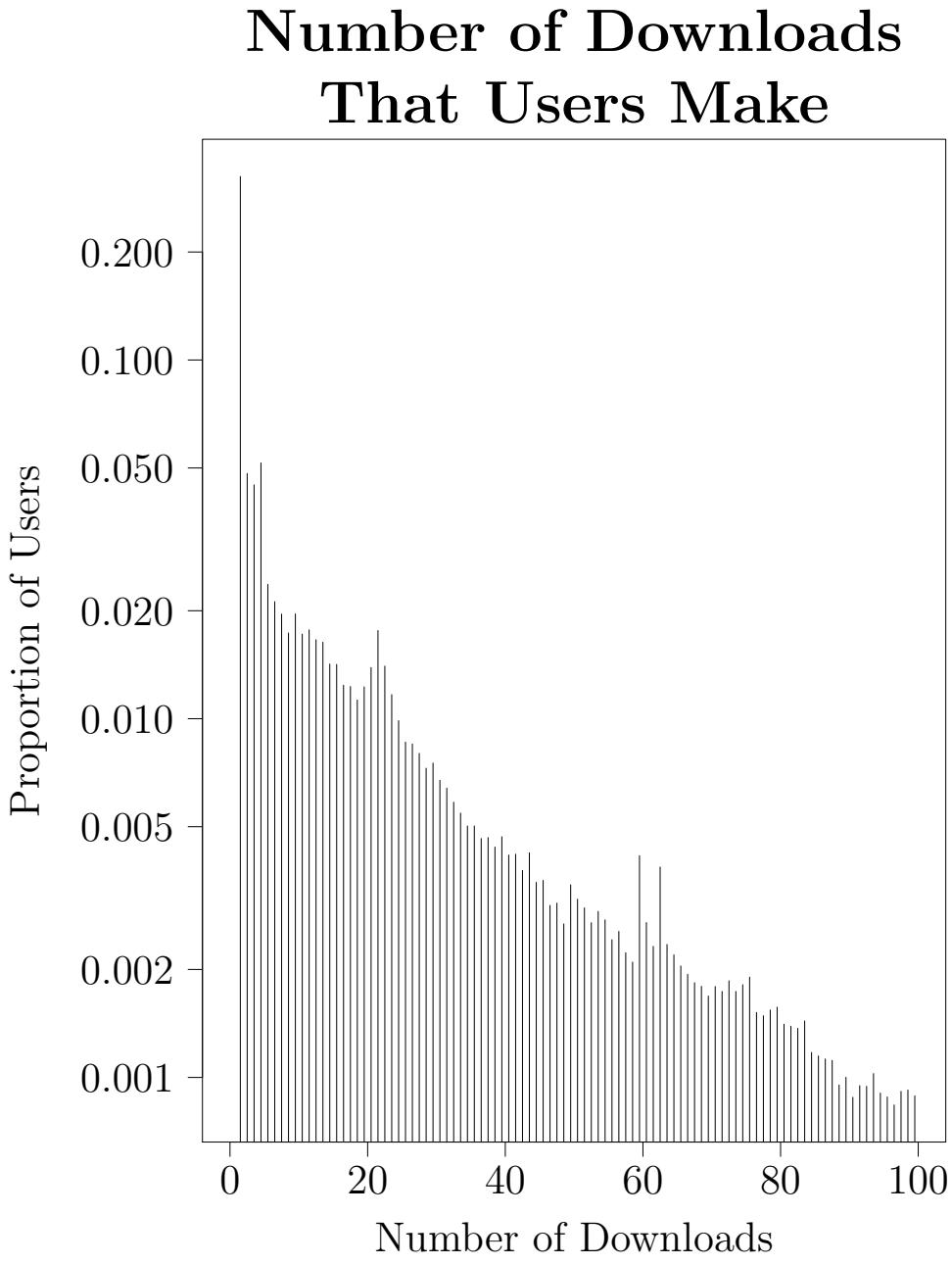


Figure 4: The proportion of users in GB who have made various numbers of downloads, where proportion is plotted on a log scale.

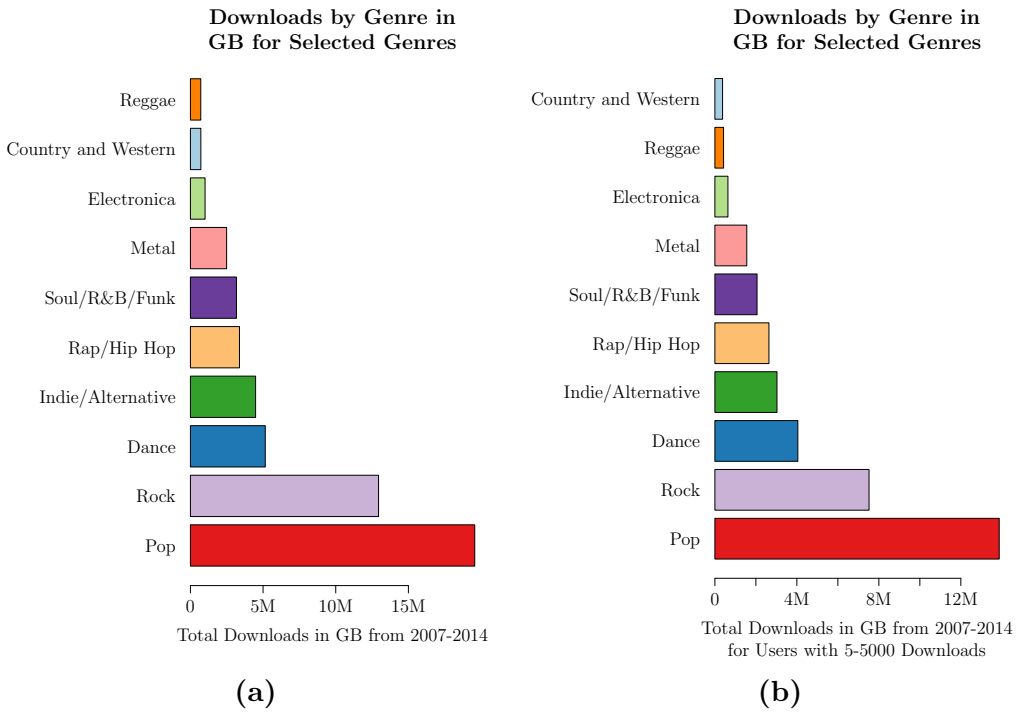
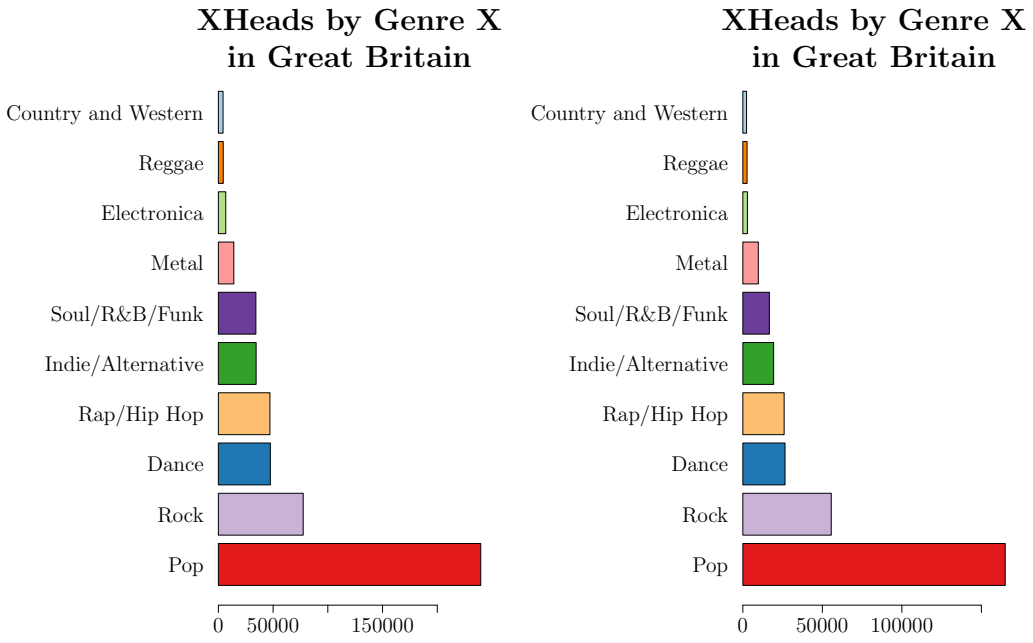


Figure 5: Downloads by genre plotted in order of number of downloads.



(a) Number of XHeads when considering all users in GB.

(b) Number of XHeads in GB who have made between 5 and 5000 downloads.

Figure 6: XHeads by genre plotted in order of number of XHeads.

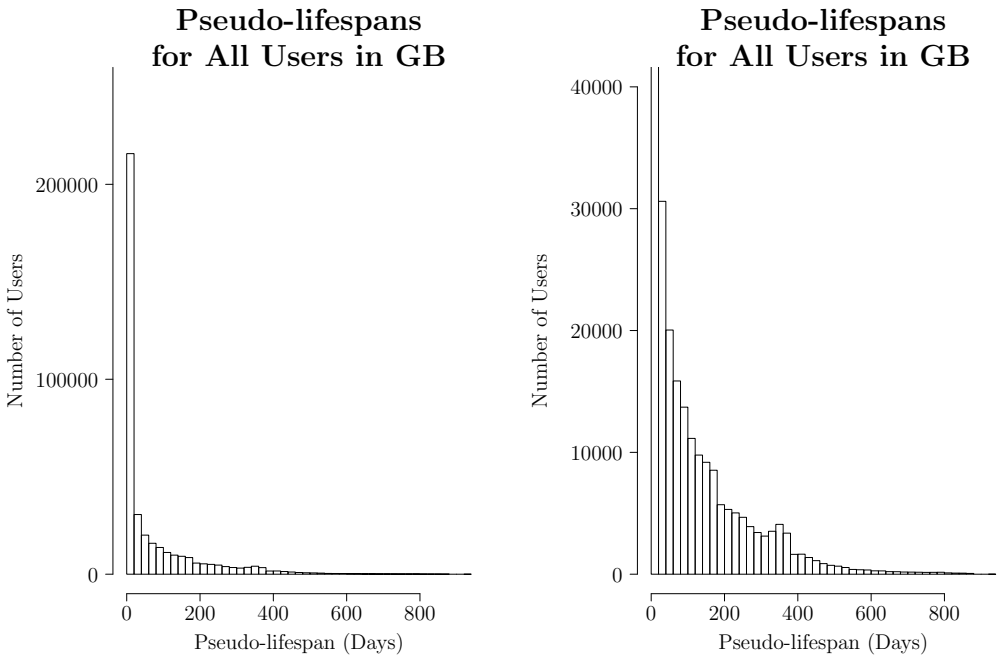


Figure 7: Distribution of pseudo-lifespans of users in GB. The second plot shows the same data as the first plot but on a smaller scale for the number of users.

2. the number of songs from each genre in our song data set,
3. the total downloads in GB in each genre, and
4. the number of downloads in each genre when only considering the songs in our data set (this will be referred to as the *filtered total downloads*).

The plots in this section present data that consider all users. Data that consider only users that made between 5 and 5000 downloads display similar trends.

There is a very tight correlation between the number of songs and the *filtered total downloads*, as can be seen in Figure 8a. A correlation is not entirely surprising since we are looking at the number of songs in each genre that made it into our sample set plotted against the number of downloads in each genre when considering only songs in that same sample set. However, the strength of this correlation is surprising and suggests that these two measures can be thought of as essentially equivalent. It is interesting to note that although there is a correlation between number of songs and total downloads (see Figure 8b), it is not as strong as the correlation between number of songs and filtered total downloads. The correlation between filtered downloads and total downloads is similar to this one (see Figure 8c).

The number of XHeads in each genre is positively correlated with both the number of songs in our sample set and the total downloads in each genre (see Figures 8d and 8e). The correlation with number of songs is stronger, and given this correlation it is not surprising to see that the number of XHeads also correlates with the filtered total downloads (see Figure 8f).

3.3 Song Download Time Series

The inspiration for this investigation came from the striking resemblance of song download time series to incidence curves derived from case report data for infectious diseases. Figure 9 shows examples of the download time series for four songs in the database. The downloads for these songs have been aggregated at different time scales for reasons that will be discussed in §4.1. The bottom two songs (shown in Figures 9c and 9d) are examples of songs with 'missing beginnings'. Some songs appear to gain popularity so quickly that a time series of downloads aggregated at the daily level does not display an initial increase in downloads. However, aggregating the entire time series at a finer timescale makes the download curve very noisy. This presented an interesting aggregation problem, which will be further discussed in Sections 4.1 and 4.2. Note that all of the plots in Figure 9 are created using downloads from all users in GB.

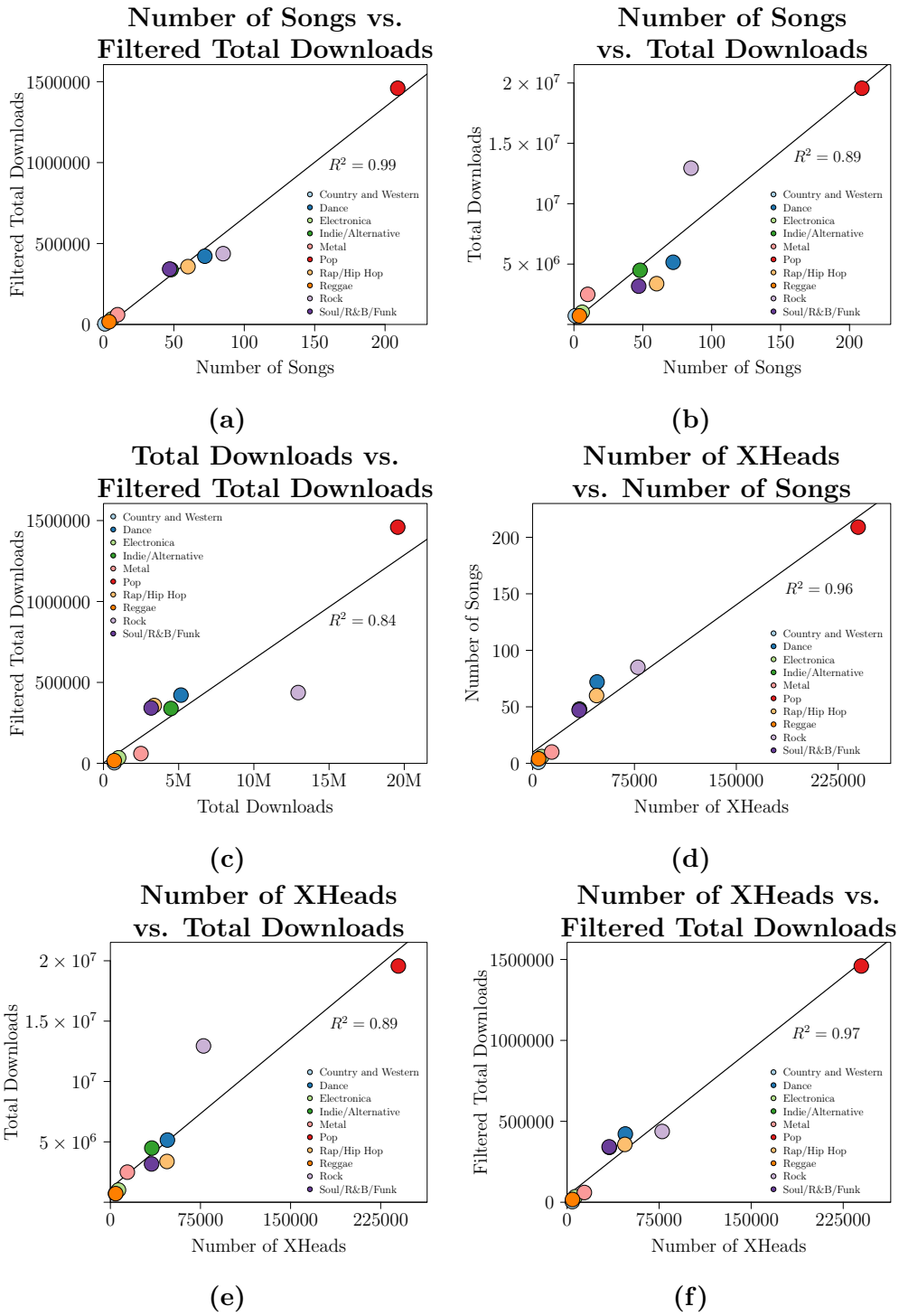


Figure 8: These scatter plots illustrate the correlations between various descriptive genre statistics.

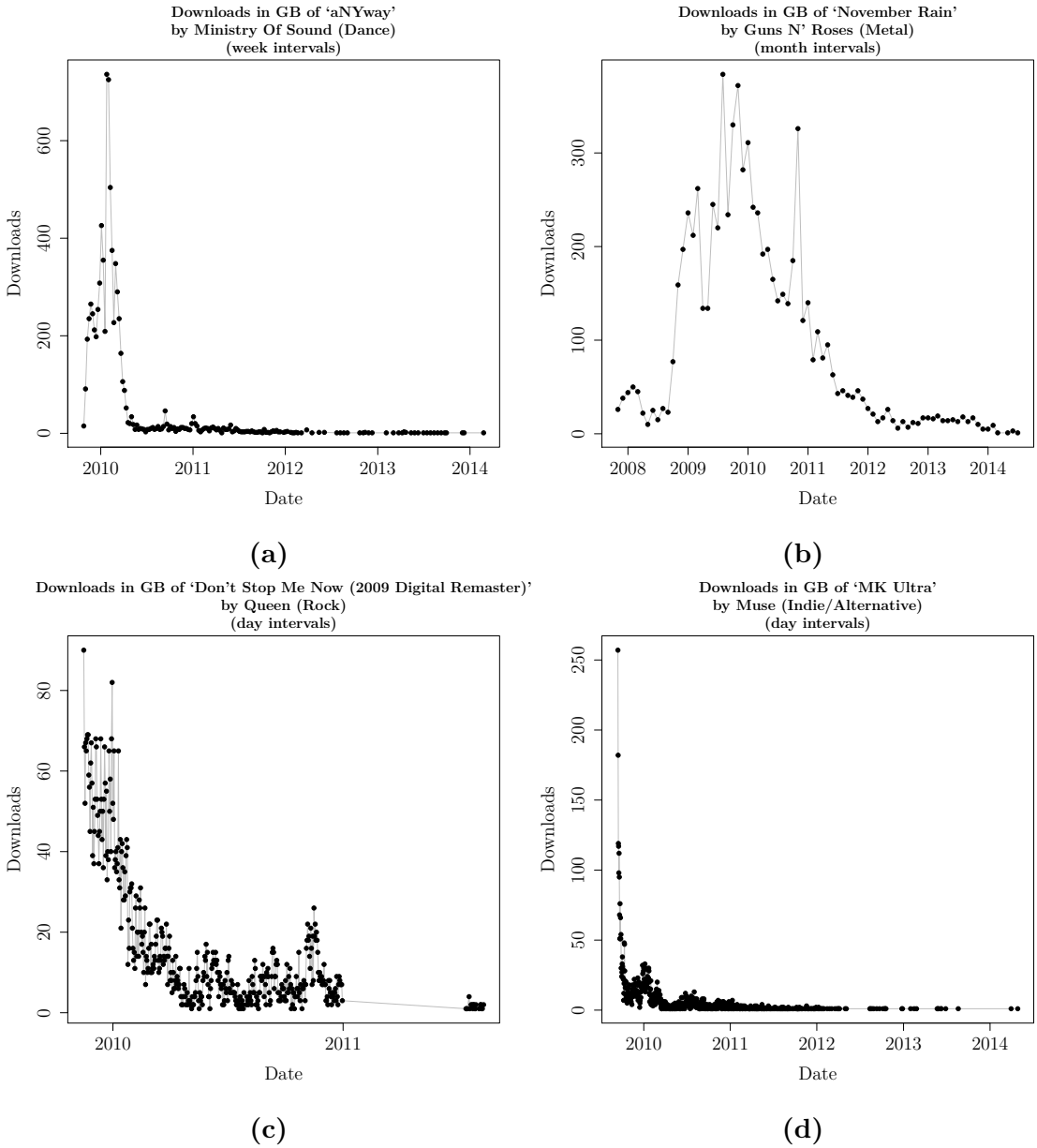


Figure 9: Four song download time series. The first shows weekly downloads, the second monthly downloads and the last two both show daily downloads.

4 Applying the SIR Model to Song Spread

If the SIR model is interpreted in the context of popular songs, individuals are classified as being 'susceptible to', 'infected with' or 'recovered from' a song. The mean infectious period $1/\gamma$ gives us a measure of the average time period for which an individual will continue to enjoy listening to a song, during which they may tell others about this song, thus 'spreading' it through the population. The basic reproduction number $\mathcal{R}_0$ gives a measure of the average number of people in a wholly susceptible population who will be influenced to download a new song by one individual who is actively listening to and talking about this song.

4.1 Fitting the Model to the Data

The SIR model was fitted to the top 1000 songs in GB. These songs were defined to be the 1000 songs with the highest total number of downloads in the database between 2007 and 2014. The database was queried using open-source MySQL implementation of SQL [42] to determine which 1000 songs had the highest download counts for users in GB. Minute-by-minute download counts for these songs were extracted, then aggregated at coarser timescales using the `R` statistical programming language [37]. The finest timescale used was daily since aggregating at timescale finer than this yielded noisy download time series and poor fits. For download time series that displayed the 'missing beginning' discussed in §3.3, the beginning of the time series was aggregated at a finer timescale than the rest in order to produce a time series that the SIR model could be fitted to. Finer aggregation was conducted up to the point where the peak number of daily downloads occurred. An example of two songs for which this was done can be seen in Figure 10. The SIR model was fitted to each of the resulting time series using the package `fitsir` in `R` [9] (this package employs least-squares fitting to match solutions of the SIR model to a given time series). Epidemiological parameter estimates from the fitted curves were extracted.

At this point, the number of songs being considered was restricted to those that yielded reasonable fits. Christmas and holiday songs were eliminated since their time series gave a pattern similar to seasonal epidemics, which the simple SIR model (Equation (1)) cannot generate [2, 21]. A minimum possible value for $\mathcal{R}_0$ was also calculated for songs in the data set as follows.

The final size was assumed to be:

$$Z(\mathcal{R}_0) = \frac{\text{Total Downloads}}{S_0}, \tag{3}$$

where 'Total Downloads' is the total number of downloads for a given song and

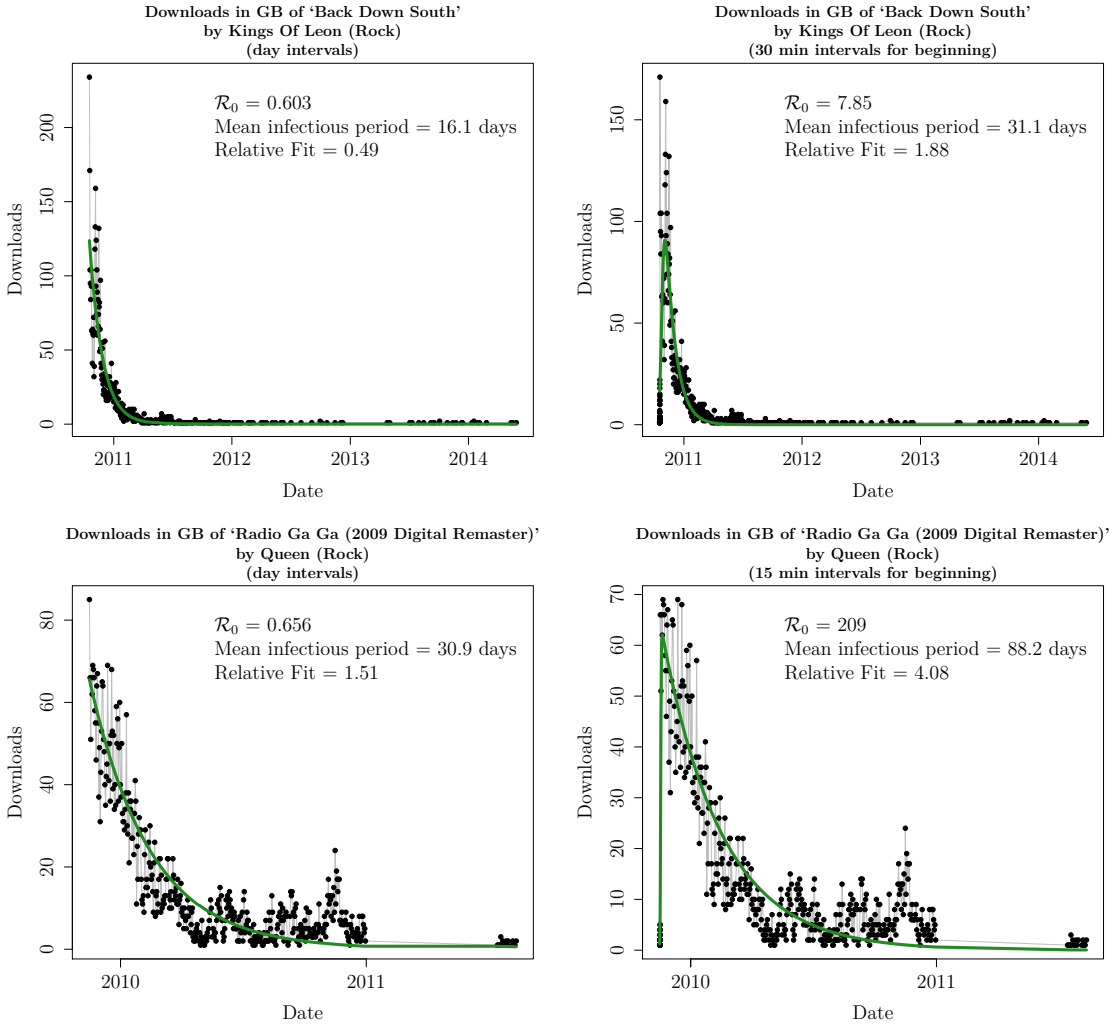


Figure 10: The beginning of the download time series for each of these two songs was aggregated at a finer time scale than the rest of the time series (which was aggregated at the daily download level). The black dots connected by the grey line represents aggregated downloads while the green line represents the fitted SIR model. In each case, the plot on the left shows the fitted model before the beginning was aggregated at a finer timescale and the plot on the right shows the fitted model after.

$Z(\mathcal{R}_0)$ (see Equation (2)) is a strictly increasing function of $\mathcal{R}_0$. The total number of users in GB was used as the maximum possible S_0 and the lowest total number of downloads for a song in the top 1000 songs in GB was used as the minimum possible Total Downloads. This gave a minimum possible Z for our song data set, which was then used to find a minimum possible $\mathcal{R}_0$ ($\mathcal{R}_{\min}$) by rearranging Equation (2) to get:

$$\mathcal{R}_0 = -\frac{\ln(1-Z)}{Z}. \tag{4}$$

Any songs with an $\mathcal{R}_0$ less than the minimum possible $\mathcal{R}_{\min}$ were also not included in further analysis. Finally, we calculated a crude measure of goodness of fit by finding the average relative distance between the model trajectory and the song download data points. Visual analysis determined that songs for which this measure was larger than 120 gave a poor fit, so these songs were also excluded.

This left a set of 542 songs for which the SIR model was considered to give a reasonable fit (140 of these were 'missing beginning' songs). A final size was calculated for each song based on the extracted $\mathcal{R}_0$ (using Equation (2)). This final size Z is a proportion of the initial susceptible population S_0, *i.e.,* the number of individuals initially susceptible. Equation (3) could therefore be used to calculate an estimate for S_0 for each song based on knowledge of Total Downloads and estimated final size. Thus, the parameters extracted from the fits for each song and the genre specific parameters discussed in §3.2 were compared in various ways.

4.2 Results and Discussion

4.2.1 All Users

Examples of songs for which the SIR model has been fitted to the download time series can be seen in Figures 11 and 12. Each of these figures displays aggregated downloads with black dots connected by a grey line and the fitted epidemic curve in green. The mean infectious period $1/\gamma$ and basic reproduction number $\mathcal{R}_0$ extracted from the SIR fit for each song are also listed in these figures. Figure 11 shows songs for which the SIR model gave a reasonable fit and demonstrates the ability of the epidemic model to capture the details of many of the song download time series. Figure 12 shows some songs for which the SIR model did not give a reasonable fit. The first two plots in Figure 12 are examples of songs that were not included in further analysis and demonstrate that there are some songs in our sample set for which the SIR model fails to capture the download trends that are present (the first was excluded because $\mathcal{R}_0 < \mathcal{R}_{\min}$ and the second because the relative fit measure was greater than 120). The bottom two plots in this figure were included and demonstrate

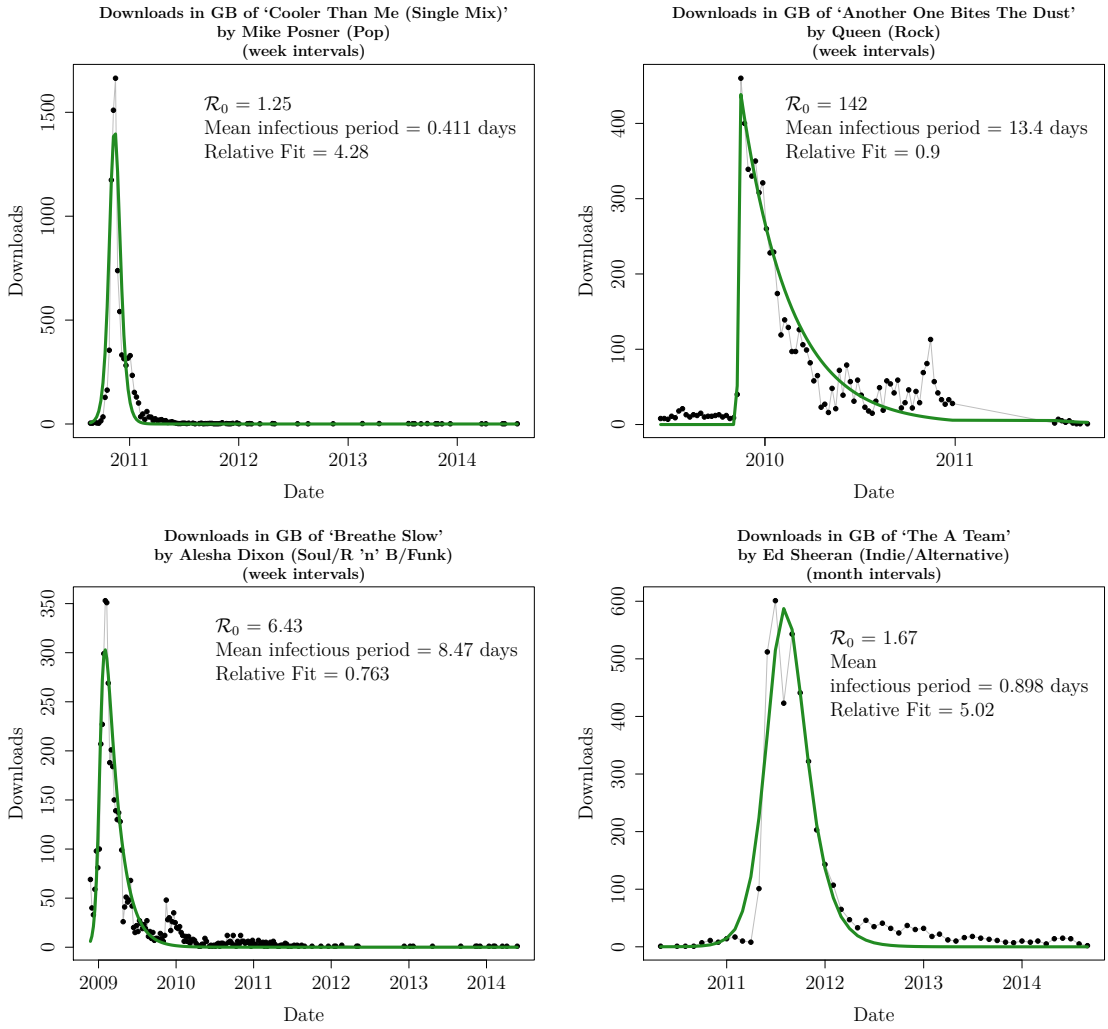


Figure 11: Four songs for which the SIR model yielded a good fit. The extracted mean infectious period $1/\gamma$ and basic reproduction number $\mathcal{R}_0$ is shown for each song as well.

that for some songs that fell within our criteria for inclusion the SIR fits do not look impressive by eye.

The distribution of various parameters that were extracted (*i.e.*, fitted) or calculated (from the extracted parameters) can be seen below. Figure 13 shows the distribution of $\mathcal{R}_0$. 531 of the 542 songs in our song set had a basic reproduction number of at most 100 and 490 songs had a basic reproduction number of at most

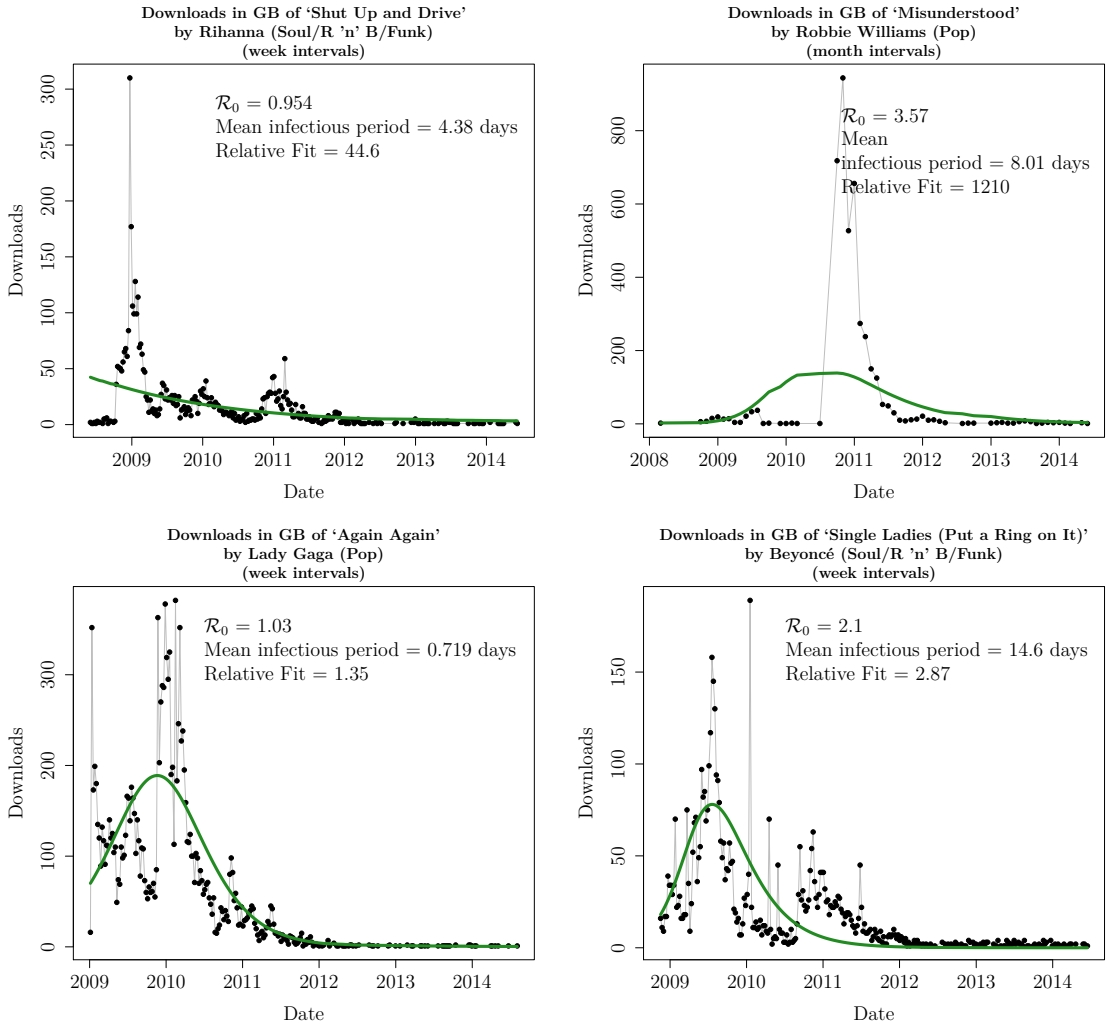


Figure 12: Four songs for which the SIR model did not yield a very good fit. The songs in the top two panels were excluded from further analysis. The fits for the songs in the bottom two panels, while less convincing to look at, met our criteria for inclusion.

Genre	No. Songs	Mean $\mathcal{R}_0$	Median $\mathcal{R}_0$
Country and Western	1	1.11	1.11
Reggae	4	3.14	3.18
Electronica	6	1.83	1.05
Metal	10	2.67	3.13
Soul/R&B/Funk	47	21.87	1.94
Indie/Alternative	48	5.34	1.39
Rap/Hip Hop	60	13.13	1.36
Dance	72	2.12	1.69
Rock	85	18.64	3.33
Pop	209	7.48	1.51

Table 2: Mean and median values of $\mathcal{R}_0$ by genre for the songs in our data set.

10. The maximum extracted $\mathcal{R}_0$ was 651.5 for the song 'Gravity' by 'Pixie Lott'. Figure 14 shows the distribution of the mean infectious period $1/\gamma$. 515 songs had mean infectious periods of at most 100 days and 442 had mean infectious periods of at most 25 days. The longest mean infectious period was 373.3 days for the song 'Between Two Lungs' by 'Florence + the Machine'. The distribution of initial growth rate r is shown in Figure 15. Figure 16 shows the distribution of calculated final sizes Z for the 542 songs. No genre patterns emerge in the histograms for these or any other extracted or calculated parameters. The mean and median values by genre for $\mathcal{R}_0$, the mean infectious period, the initial growth rate and final size are listed below. Tables 2 and 3 show the average values of $\mathcal{R}_0$ and the mean infectious period and Tables 4 and 5 show the average values of the initial growth rate and final size.

In the basic SIR model, the initial growth rate r is given by $r = \beta - \gamma$. $\mathcal{R}_0 = \beta/\gamma$, so $r = \gamma(\mathcal{R}_0 - 1)$. This means that a linear relationship exists between initial growth rate and the basic reproduction number. If the fitted γ were always the same, we would therefore expect to see a perfect linear correlation between these two parameters. More generally, variance in γ would account for scatter in the linear correlation between r and $\mathcal{R}_0$. As can be seen in Figure 17, there is a positive correlation between $\mathcal{R}_0$ and r, but it is surprisingly very weak. Also, points that lie at smaller values of $\mathcal{R}_0$ and r appear to follow a different trend than the overall linear correlation displayed. Note that in the scatter plot in Figure 17 (and in all subsequent scatter plots), each point represents one song in our data set and the colour of the point corresponds to the genre of that song.

An unexpected strong linear correlation exists between transmission rate β and

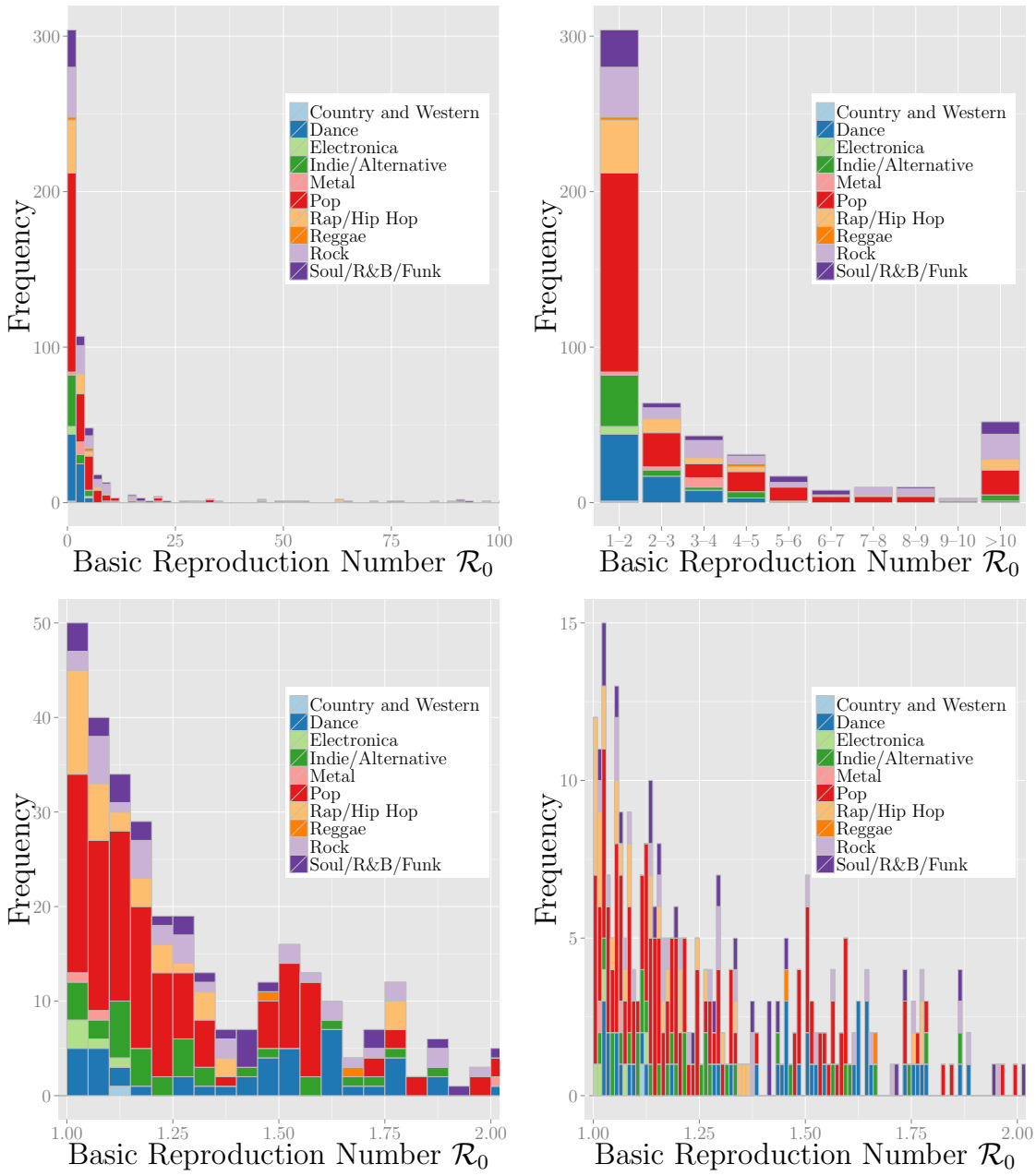


Figure 13: Distribution of the basic reproduction number $\mathcal{R}_0$ extracted from the SIR model fitted to the 542 songs in our sample set. 11 songs had an $\mathcal{R}_0$ greater than 100 and 52 songs had an $\mathcal{R}_0$ greater than 10.

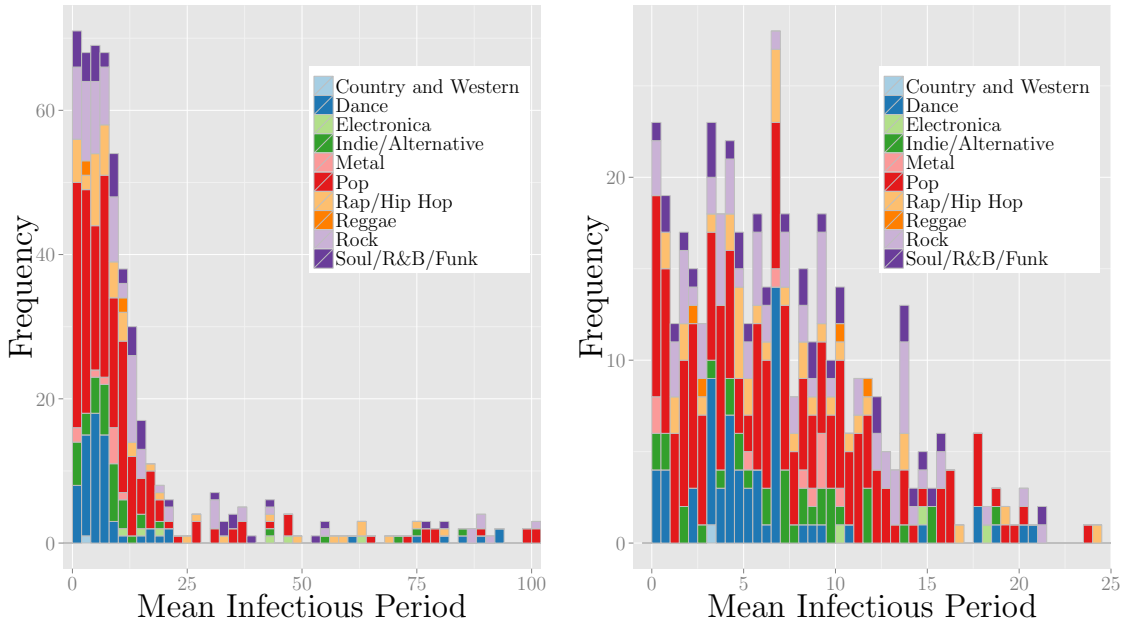


Figure 14: Distribution of the mean infectious period $1/\gamma$ extracted from the SIR model fitted to the 542 songs in our sample set. 27 songs had a mean infectious period greater than 100 and 100 songs had a mean infectious period greater than 25.

Genre	No. Songs	Mean $1/\gamma$	Median $1/\gamma$
Country and Western	1	3.179	3.179
Reggae	4	6.760	6.660
Electronica	6	32.832	30.988
Metal	10	6.820	8.694
Soul/R&B/Funk	47	29.336	10.279
Indie/Alternative	48	51.302	9.036
Rap/Hip Hop	60	35.922	10.068
Dance	72	13.341	5.326
Rock	85	20.342	9.074
Pop	209	15.145	7.100

Table 3: The mean and median values for mean infectious period $(1/\gamma)$ by genre for songs in our data set.

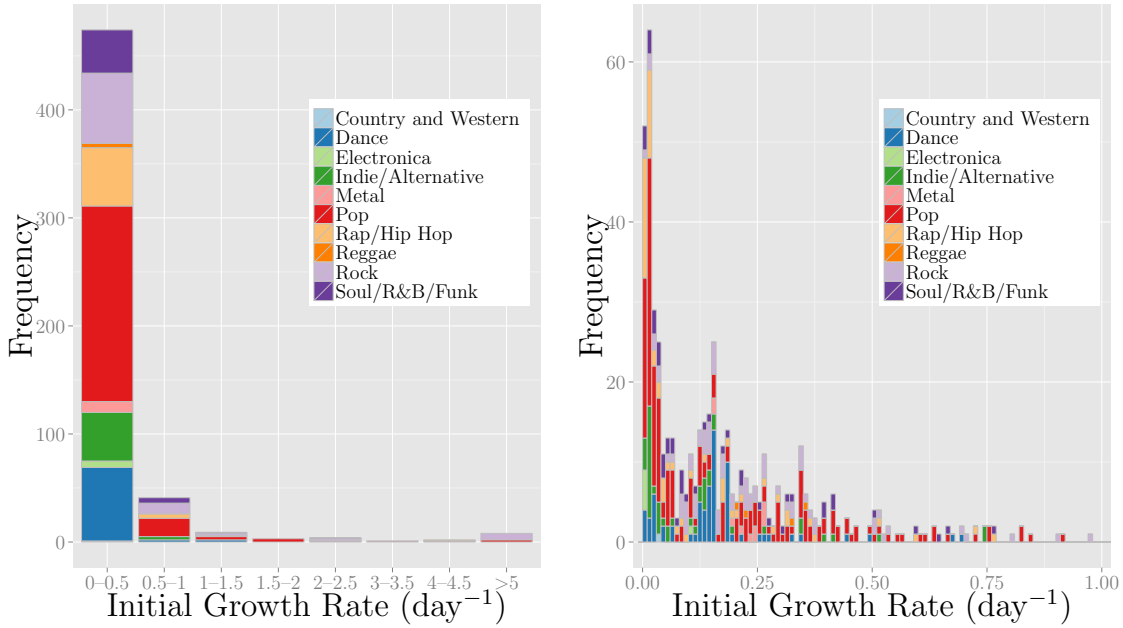


Figure 15: Distribution of the initial growth rate extracted from the SIR model fitted to the 542 songs in our sample set.

Genre	No. Songs	Mean r	Median r
Country and Western	1	0.035	0.035
Reggae	4	0.278	0.273
Electronica	6	0.019	0.001
Metal	10	0.224	0.241
Soul/R&B/Funk	47	0.251	0.094
Indie/Alternative	48	0.128	0.048
Rap/Hip Hop	60	0.218	0.040
Dance	72	0.190	0.151
Rock	85	1.125	0.215
Pop	209	0.331	0.126

Table 4: The mean and median initial growth rate (r) by genre for songs in our data set.

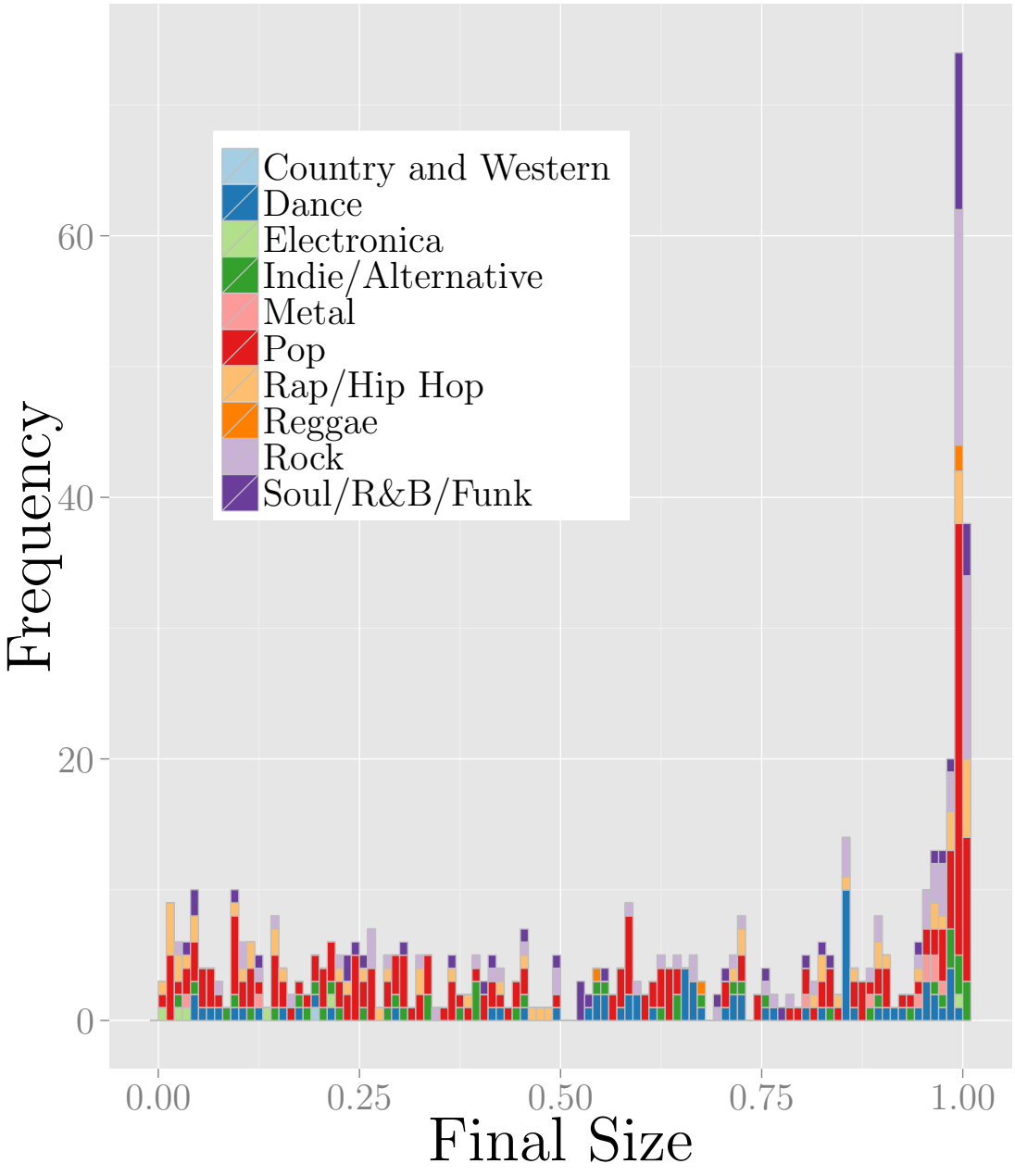


Figure 16: Distribution of final size for the 542 songs in our sample set, calculated based on parameters extracted from the SIR fit for each song.

genre	No. Songs	Mean Z	Median Z
Country and Western	1	0.192	0.192
Reggae	4	0.801	0.832
Electronica	6	0.238	0.090
Metal	10	0.761	0.948
Soul/R&B/Funk	47	0.673	0.780
Indie/Alternative	48	0.548	0.497
Rap/Hip Hop	60	0.546	0.479
Dance	72	0.645	0.688
Rock	85	0.767	0.959
Pop	209	0.568	0.589

Table 5: The mean and median final size Z by genre for songs in our data set.

recovery rate γ (see Figure 18). When looking at the same scatter plot on a smaller scale, it becomes apparent that many of the points are clustered at very small values of β and γ. The minimum $\mathcal{R}_0$ used to exclude songs from analysis was 1.003563. This is very close to $\mathcal{R}_0 = 1$ and since it is greater than 1 there are no points below the line $\beta = \gamma$ in the scatter plot for γ vs. β. The histograms of $\mathcal{R}_0$ in Figure 13 show that many of the songs in our data set have basic reproduction numbers quite close to 1. This combined with the minimum $\mathcal{R}_0$ criterion we have set may have resulted in many points falling close to the line $\beta = \gamma$ in Figure 18. In other words, it is possible that the strong correlation between β and γ is simply a result of the fact that many songs have an $\mathcal{R}_0$ close to 1, but no songs have an $\mathcal{R}_0$ less than 1.

There is a strong linear correlation for each of β and γ with each of extracted initially susceptible population and extracted population size N. However, in each case the correlation looks much less convincing when examining the plot on a smaller scale (see Figure 19 for the γ vs. N example).

A mild linear correlation also exists between the mean infectious period $1/\gamma$ and the initially infectious population I_0 extracted from the SIR fits (see Figure 20). It is possible that a different trend from that displayed exists for points in the region where $I_0 < 0.2$ and $1/\gamma < 25$.

Since we used extracted $\mathcal{R}_0$ values to calculate a final size for each song, the final size relationship given in Equation (2) above is displayed in the scatter plot of $\mathcal{R}_0$ vs. final size Z (see Figure 21a). It is slightly surprising to see the similarly shaped curve that results from plotting calculated S_0 against $\mathcal{R}_0$ in Figure 21b. The curve appears to resemble $1/Z(\mathcal{R}_0)$. As previously discussed, we calculated an estimate

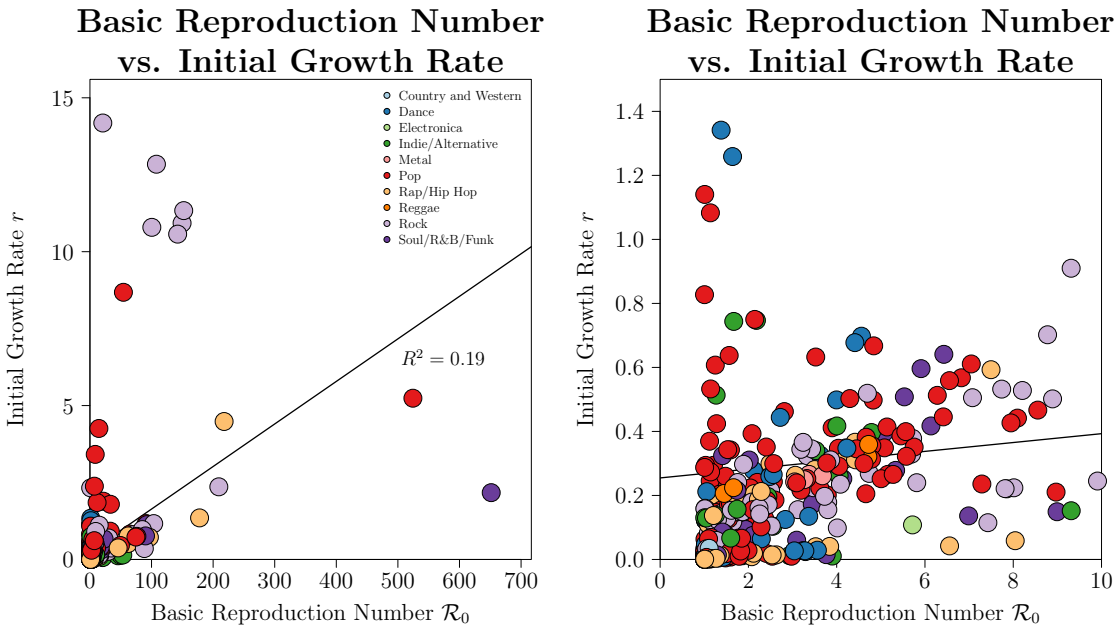


Figure 17: Scatter plots demonstrating the weak linear correlation between R_0 and initial growth rate r. The black line in the second plot is the same linear fit as that shown in the first plot.

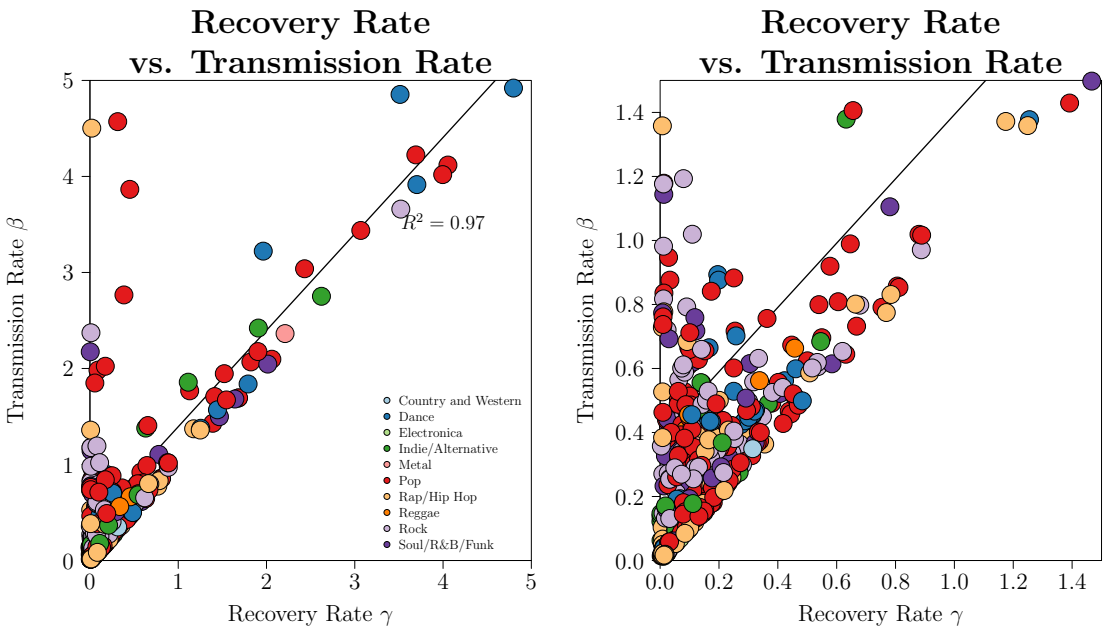


Figure 18: Scatter plots demonstrating the linear correlation between recovery rate γ and transmission rate β. 27 songs with larger transmission or recovery rates are not shown in the first plot.

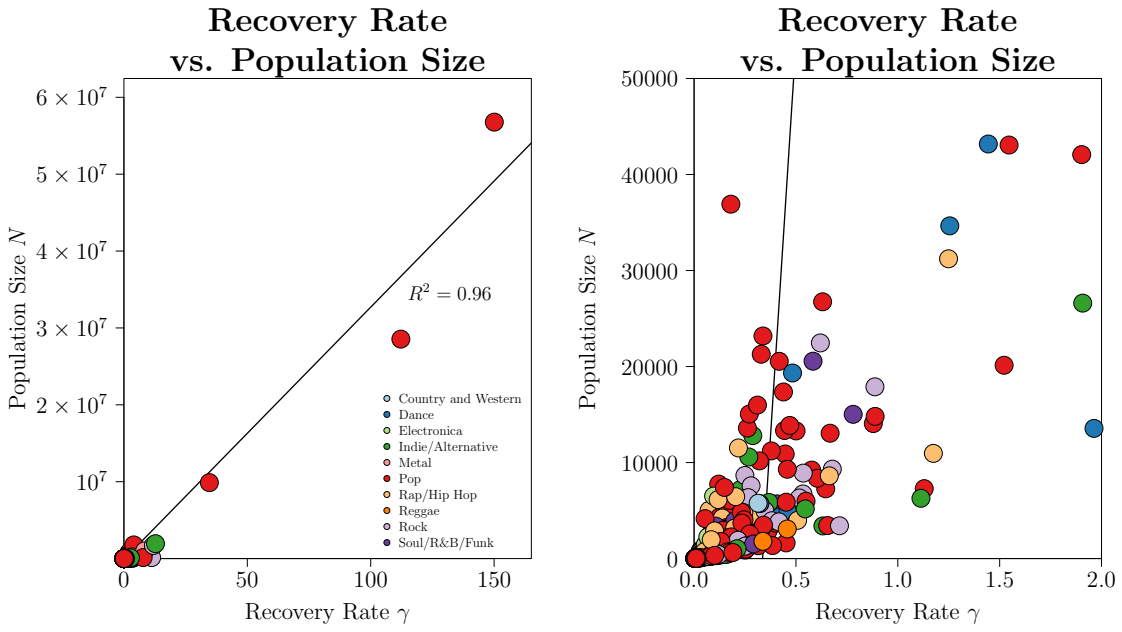


Figure 19: There appears to be a strong linear correlation between recovery rate γ and extracted population size N, however when looking at small values of γ and N the correlation looks less accurate. It is possible that different relationship might exist for the set of points that lie in the region where $N < 50000$ and $\gamma < 2$.

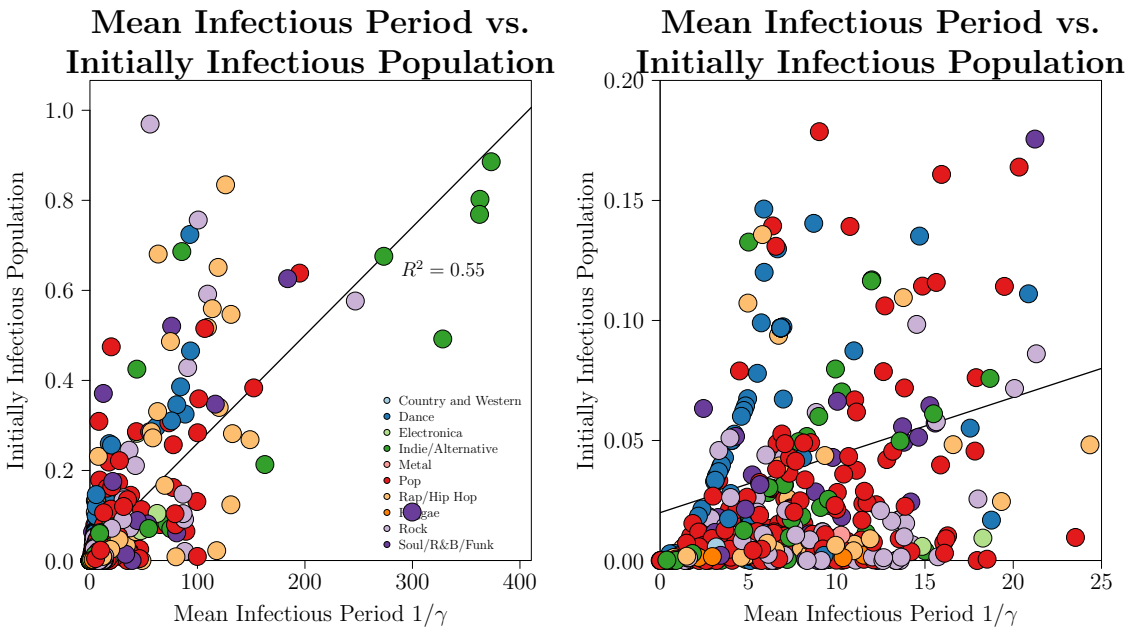
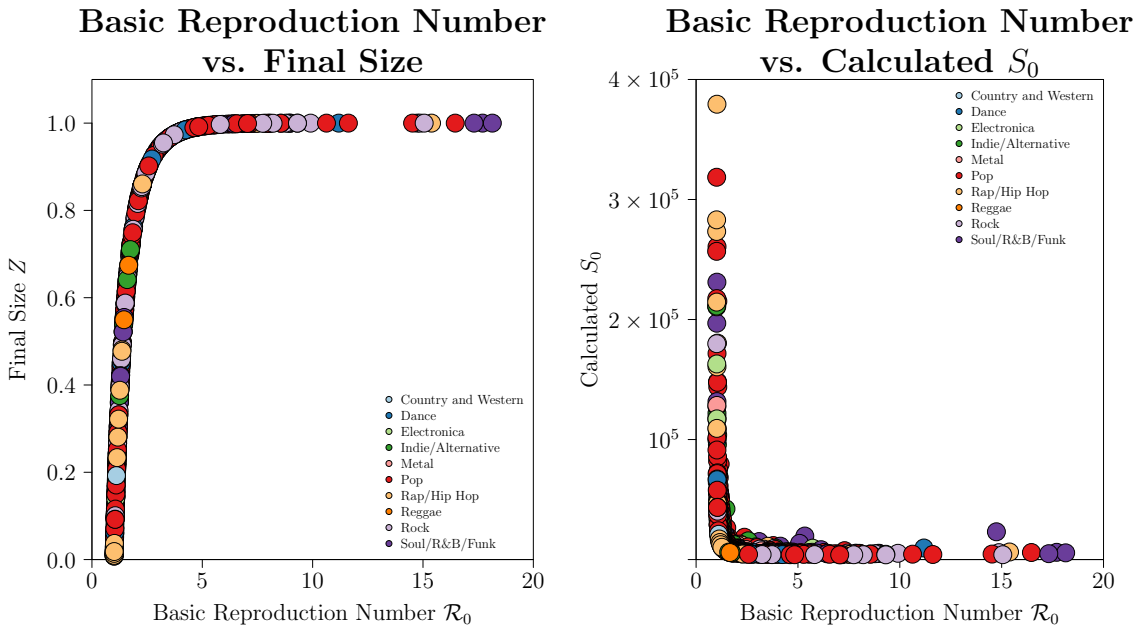


Figure 20: Scatter plots demonstrating the mild linear correlation between mean infectious period $1/\gamma$ and initially infectious population I_0. There may exist a different trend for points that lie in the region where each of these parameters is small.



(a) The final size relationship is demonstrated by the scatter plot of $\mathcal{R}_0$ vs. final size Z. 40 songs with $\mathcal{R}_0 > 20$ do not appear in the plotting area.

(b) Scatter plot showing $\mathcal{R}_0$ vs. calculated S_0. 50 songs with $\mathcal{R}_0 > 20$ and/or calculated $S_0 > 400000$ do not appear in the plotting area.

Figure 21

for S_0 for each song by taking

$$S_0 = \frac{\text{Total Downloads}}{Z(\mathcal{R}_0)} \tag{5}$$

There is indeed a $1/Z(\mathcal{R}_0)$ factor in the calculation of S_0, however it is multiplied by the total number of downloads for the song in question. Unlike $\mathcal{R}_0$ the total number of downloads is not an extracted parameter. It varies greatly and there is no correlation between the total number of downloads and $\mathcal{R}_0$. It is therefore somewhat surprising that the shape of the curve $1/Z(\mathcal{R}_0)$ is still apparent in Figure 21b. Interestingly, when the S_0 extracted from SIR fits is plotted against $\mathcal{R}_0$ there is also a trace of the $1/Z(\mathcal{R}_0)$ curve (see Figure 22), yet extracted S_0 and calculated S_0 are not correlated.

It might be logical to expect a correlation between number of XHeads in each genre and the average S_0 for each genre – indeed, part of our motivation to calculate an initially susceptible population was so that we could compare these results with

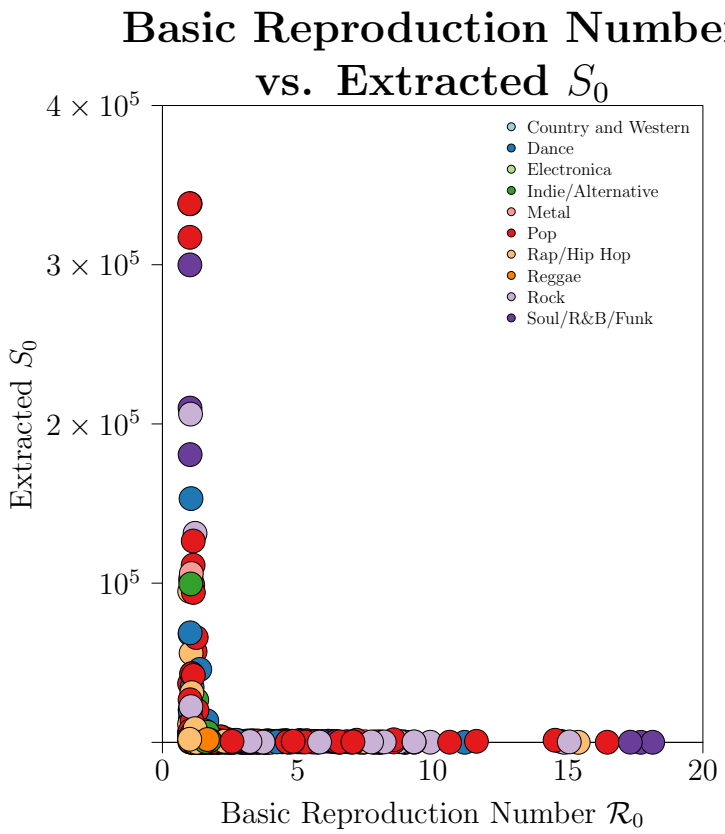


Figure 22: Scatter plot showing $\mathcal{R}_0$ vs. extracted S_0. 52 songs with $\mathcal{R}_0 > 20$ and/or extracted $S_0 > 400000$ do not appear in the plotting area.

the size of XHead populations. As can be seen in Figure 23, there is no correlation between calculated initially susceptible population S_0 and the number of XHeads in each genre, and the range of calculated S_0 is quite large with some significant outliers. The lack of correlation could be an indication that the basic SIR model is not capturing the process of song transmission well enough, or perhaps there is another reason that it is unreasonable to expect a correlation here. There is also no correlation between S_0 extracted from the SIR fits and number of XHeads in each genre.

The number of XHeads was plotted against the average values for all other extracted and calculated parameters for songs. Correlations were found for the mean values of transmission rate, recovery rate, N and extracted S_0 by genre, but these correlations disappeared when XHeads was plotted against genre medians. It is therefore likely that they were driven by outliers. Due to the strong correlations between XHeads and number of songs, total downloads and filtered total downloads (see §3.2), similar patterns appeared when each of these was plotted against mean and median genre values for each of the extracted and calculated song parameters.

It is unfortunate that no relationships were found between any parameters that can be measured early in an epidemic and a parameter that is indicative of overall song popularity, such as total downloads. It would be useful to be able to predict cumulative song popularity from the initial pattern of downloads for that song.

Previous work has shown that the pattern of downloads is different for songs in different genres [43]. However, none of our comparisons of extracted and calculated parameters yielded trends that appeared to rely on genre. This could imply that if a 'song transmission' mechanism is at play then it functions similarly across genres, or simply that further investigation is warranted.

4.2.2 Users Who Made Between 5 and 5000 Downloads

Download time series for the top 1000 songs in GB when only considering downloads made by users with between 5 and 5000 downloads were also analyzed. The same fitting and extraction methods were used as those described above for download time series that included downloads by all users. In this case, 544 songs met our criteria for yielding a reasonable fit (compared to the 542 songs that met the criteria when all downloads were included). For a given song, the curve fitted to the time series that only included downloads by users who had made 5–5000 downloads was generally similar in shape to the curve fitted to the time series that included all downloads, but had different epidemiological parameters. Figure 24 shows examples of two songs with epidemiological curves fitted to their download time series that includes all

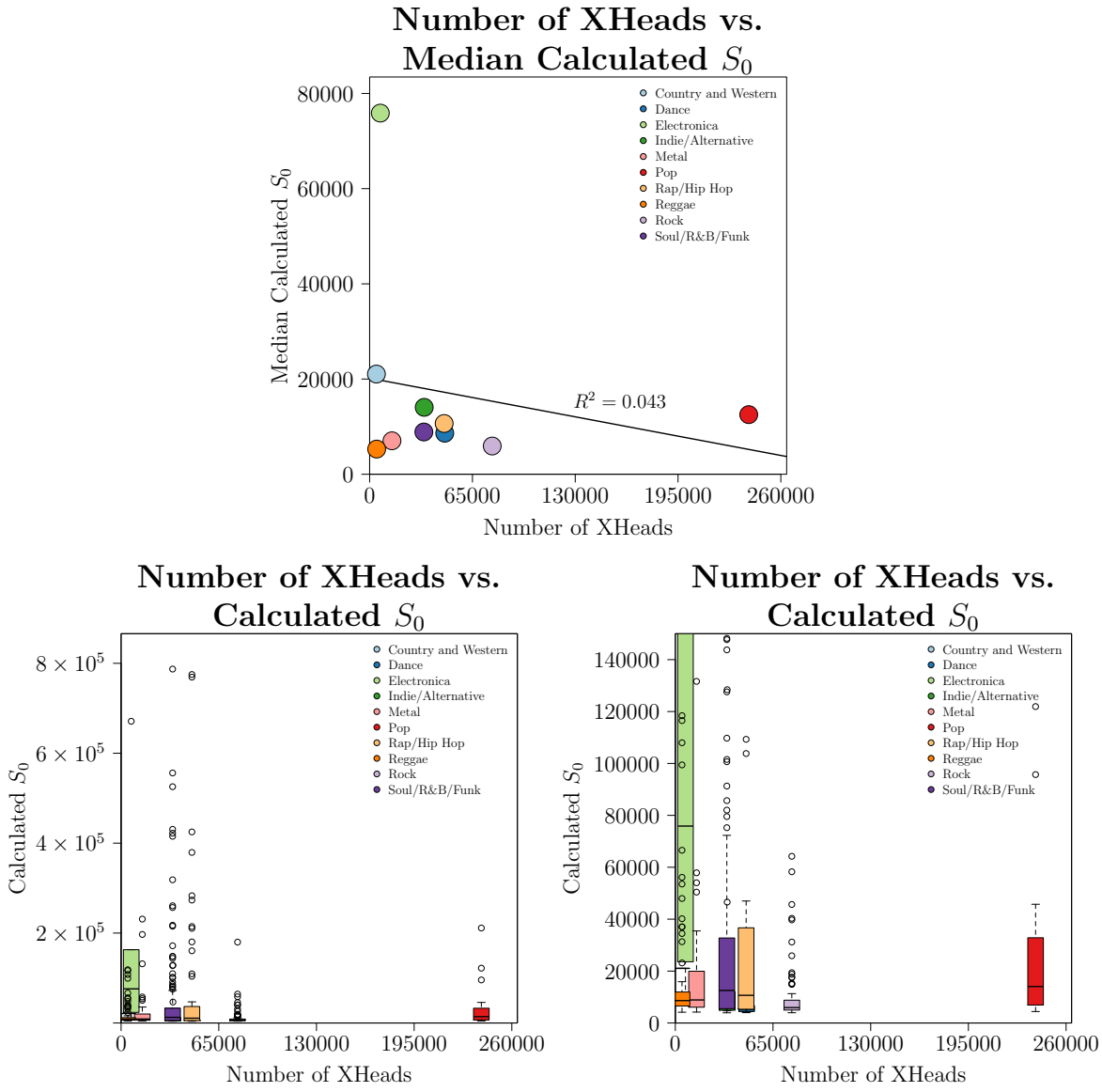


Figure 23: The scatter plot in the first figure shows that there is no correlation between median calculated S_0 and number of XHeads for a given genre. The subsequent two boxplots show how calculated S_0 is distributed in each genre.

Genre	No. Songs	Mean $\mathcal{R}_0$	Median $\mathcal{R}_0$
Country and Western	4	58.96	37.81
Reggae	4	3.02	3.24
Electronica	7	1.33	1.16
Metal	10	2.16	1.85
Soul/R&B/Funk	41	3.32	1.50
Rap/Hip Hop	47	11.70	1.50
Indie/Alternative	52	1.75	1.20
Dance	63	2.89	2.02
Rock	69	4.70	2.44
Pop	246	16.10	1.49

Table 6: The mean and median basic reproduction number ($\mathcal{R}_0$) by genre when only considering downloads by users who have made 5–5000 downloads.

downloads and to their download time series that only includes downloads by users with 5–5000 downloads. The average values by genre for $\mathcal{R}_0$ and mean infectious period $1/\gamma$ can be seen in Tables 6 and 7, and the average values for initial growth rate r and final size Z can be seen in Tables 8 and 9. These are often similar to those found in the previous section that considered downloads by all users, but sometimes quite different, in particular when the number of songs in the genre in question is quite small.

Despite individual song parameter differences in the 5–5000 downloads population, the overall data trends when comparing parameters appeared to be similar to those found when examining song time series that included all users. All parameter correlations and relationships discussed in §4.2.1 were the same when comparing extracted and calculated parameters in the 5–5000 downloads population, except for the correlation between number of XHeads and mean recovery rate γ, which is no longer present. Some linear correlations were slightly weaker or stronger, but the basic patterns were consistent.

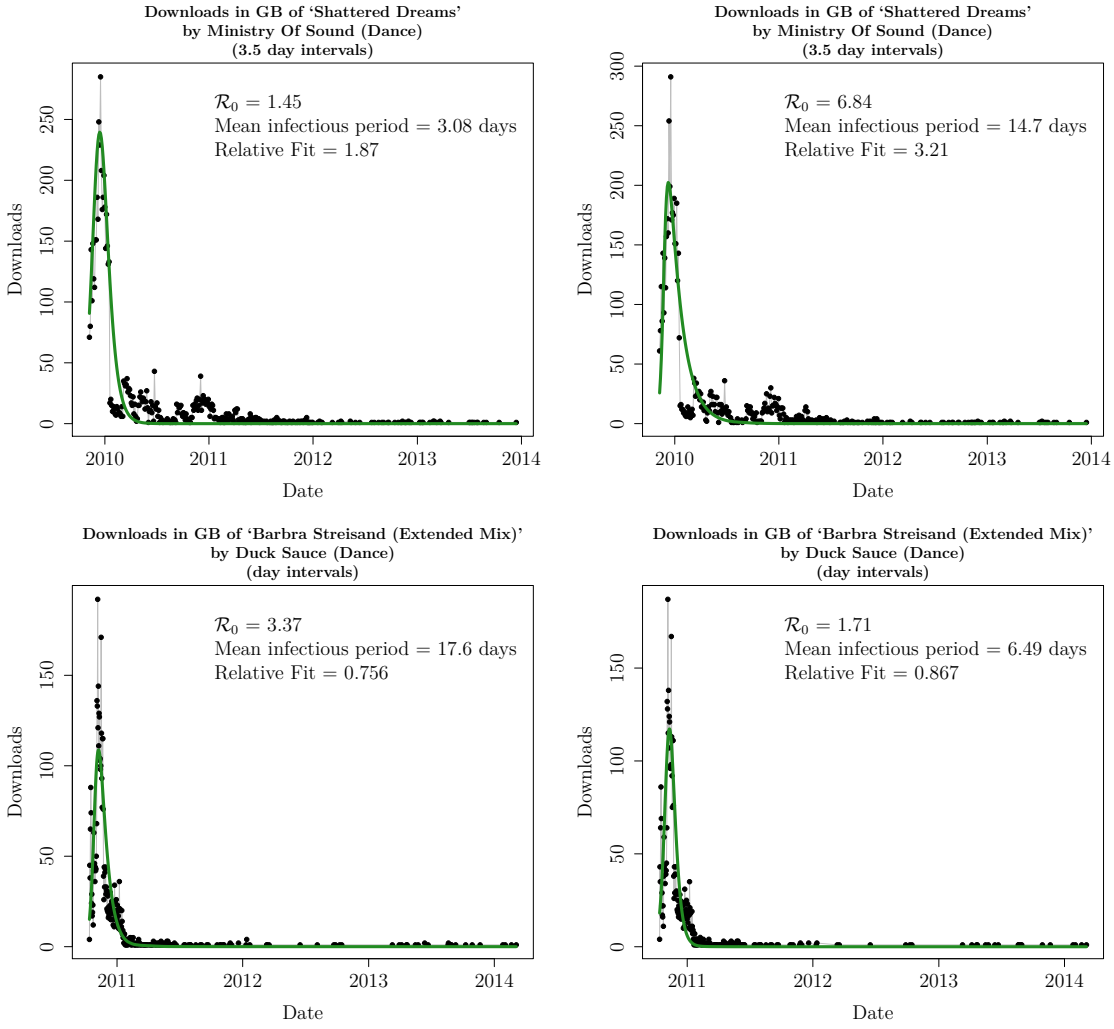


Figure 24: The download time series and fitted curves are shown for these two songs, first when considering all users and then only considering users who have made 5–5000 downloads.

Genre	No. Songs	Mean $1/\gamma$	Median $1/\gamma$
Country and Western	4	112.741	55.382
Reggae	4	9.884	12.344
Electronica	7	70.440	56.786
Metal	10	5.513	4.211
Soul/R&B/Funk	41	19.976	7.462
Rap/Hip Hop	47	26.624	6.865
Indie/Alternative	52	13.327	6.533
Dance	63	10.309	6.893
Rock	69	14.215	9.033
Pop	246	13.560	7.418

Table 7: The mean and median values for infectious period ($1/\gamma$) by genre when only considering downloads by users who have made 5–5000 downloads.

Genre	No. Songs	Mean r	Median r
Country and Western	4	0.690	0.246
Reggae	4	0.193	0.186
Electronica	7	0.006	0.003
Metal	10	0.195	0.189
Soul/R&B/Funk	41	0.167	0.082
Rap/Hip Hop	47	0.244	0.041
Indie/Alternative	52	0.071	0.033
Dance	63	0.314	0.191
Rock	69	0.374	0.191
Pop	246	0.904	0.099

Table 8: The mean and median initial growth rate (r) by genre when only considering downloads by users who have made 5–5000 downloads.

Genre	No. Songs	Mean Z	Median Z
Country and Western	4	0.591	0.624
Reggae	4	0.749	0.944
Electronica	7	0.364	0.268
Metal	10	0.631	0.750
Soul/R&B/Funk	41	0.593	0.580
Rap/Hip Hop	47	0.550	0.580
Indie/Alternative	52	0.397	0.312
Dance	63	0.754	0.802
Rock	69	0.747	0.885
Pop	246	0.575	0.575

Table 9: The mean and median final size Z by genre when only considering downloads by users who have made 5–5000 downloads.

5 Possible Extensions of the Basic SIR Model

Based on our analysis in the previous sections, the SIR model appears to capture some elements of 'song transmission', but it is likely not the best model to describe this process. The SIR model is the one of the most basic epidemiological models and, as such, does not take into account every aspect of disease transmission. The spreading of song preference is also inherently different from the spreading of a disease. While diseases are spread almost exclusively through human contact, songs can also be spread via the radio, streaming services or media services. Transmission of a song is also likely influenced by the specific people that an individual comes into contact with. For example, it is likely that a friend's opinion of a song will carry more weight than the opinion of a stranger, but in the simple SIR framework contact with an infectious stranger or an infectious friend has the same probability of resulting in transmission of a disease. There are many details of song spread (and disease spread) that the basic SIR model does not account for. What other models might we consider applying to our data set to study song transmission?

5.1 Adding Vital Dynamics

The basic SIR model that we applied to our song data does not include vital dynamics. In the context of this system, 'birth' would be interpreted as signing up for the MixRadio service and 'death' would be interpreted as quitting the MixRa-

dio service. Including vital dynamics would make the SIR model only slightly more complicated (see Equation (6) below) and might allow us to better capture certain trends in the song download data. For example, the basic SIR model cannot capture damped oscillations or multiple peaks [2, 21]. Holiday songs display seasonal peaks in their download time series (see Figure 25) and some other songs in our song data set appear to display damped oscillations (see Figure 26). It is possible that the SIR model with vital dynamics would be able to capture these behaviours. Adding vital dynamics would also allow us to take into account the fact that the size of the population of MixRadio users changes over time. With vital dynamics, the SIR model becomes:

$$\frac{dS}{dt} = \mu - \beta SI - \mu S \tag{6a}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I \tag{6b}$$

$$\frac{dR}{dt} = \gamma I - \mu R, \tag{6c}$$

where μ is both the birth and death rate, making $1/\mu$ the average lifespan (*i.e.,* it is assumed that the birth rate is equal to the death rate) [2, 21].

Incorporating vital dynamics might be one way of addressing the fact that some downloads are paid for and some are not. If users who pay and do not pay for downloads were treated as two separate populations with different average lifespans, an estimate for each of these average lifespans could be found and used to fit an SIR model that included vital dynamics to song data for each group of users. The fitted curves and parameters extracted for songs from each group could then be compared to determine whether there are differences in download behaviours and song spread.

5.2 Nonlinear Incidence Term

There are certain aspects of song transmission that are fundamentally different from disease transmission. As such, it makes sense to consider possible modifications that could be made to the transmission term in the SIR model. The current transmission term is simply βSI. Perhaps the correct transmission term for songs is not linear in I, *i.e.,* maybe it should involve a nonlinear incidence term. Given the nature of song transmission, perhaps it should be some function of I instead, such as a logistic type function or I^q, where $q > 0$.

A logistic function of I could be justified if the first few people who recommend a song have a strong influence, but eventually that influence tapers off as more people continue to recommend it. A function of the style I^q, $q > 0$ has been used

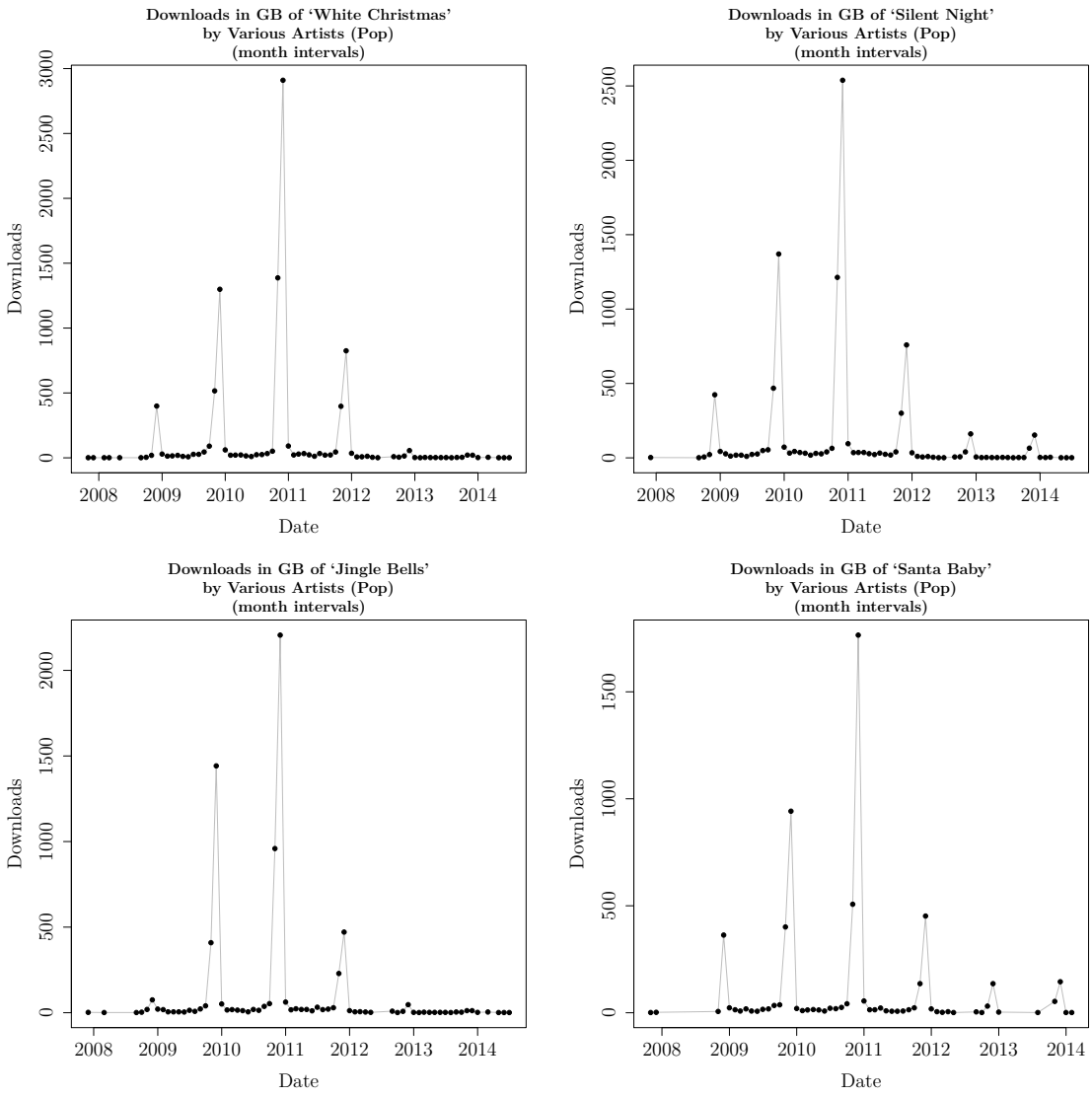


Figure 25: Much like seasonal epidemics, the download time series for holiday songs display seasonal peaks.

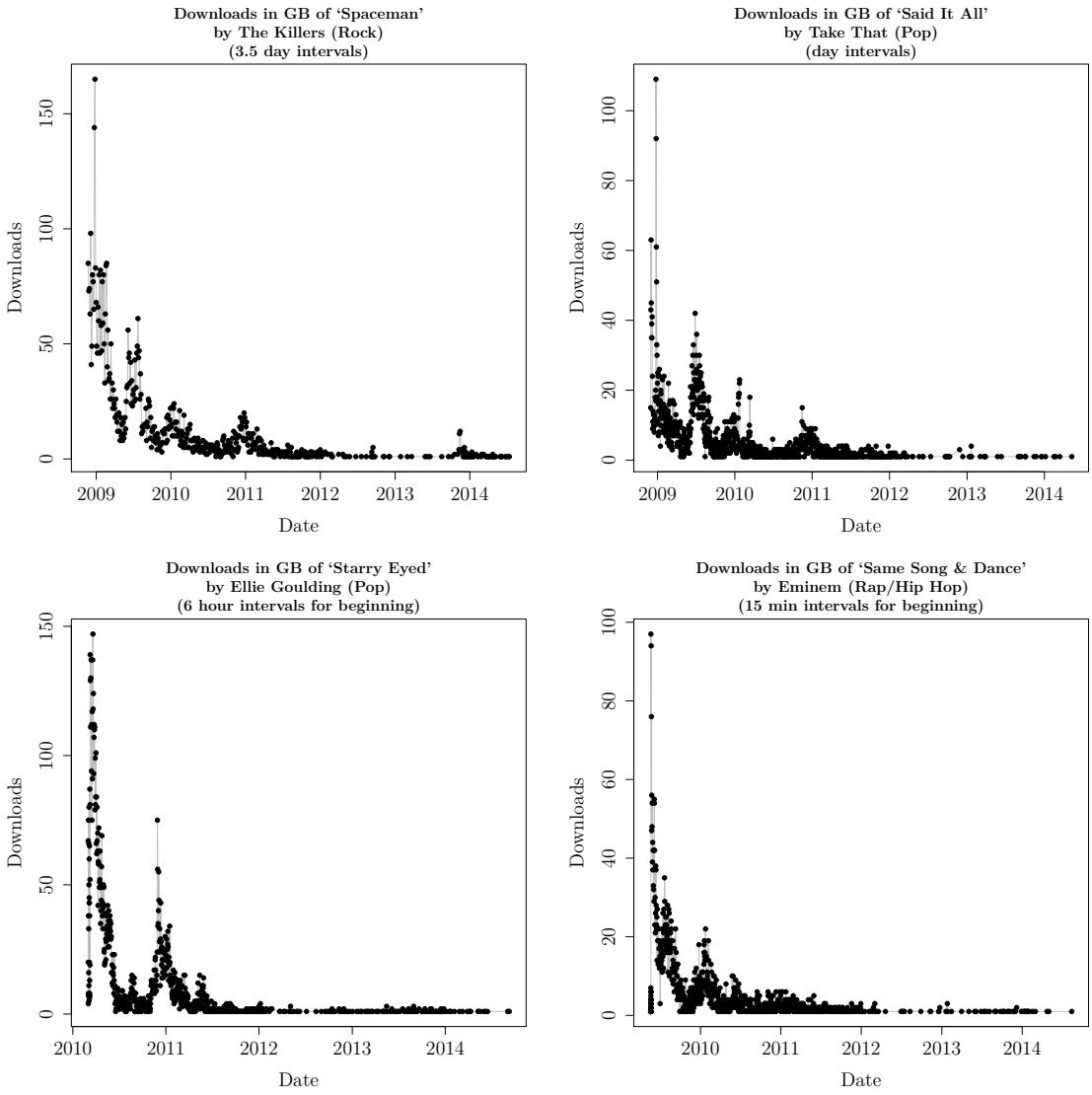


Figure 26: The download time series for some songs appear to display damped oscillations.

to mimic the effects of spatial structure or other aspects of inhomogeneous mixing in other models. Thorough investigation of the effects of including a function of I in this style has been conducted elsewhere in the literature [27, 28]. These authors looked at various epidemiological models (including the SIR model) and found that a transmission term of the form βSI^q resulted in a much greater range of possible behaviours for solutions, including the possibility of periodic solutions. They also found that for some ranges of parameters, solutions behaved differently depending on the initial conditions. In the context of song popularity, it sometimes seems that the conditions under which a song is released play a large role in its success or failure – work by Liu *et al.* [27] suggests that this could be represented using a model that has a transmission term of the form βSI^q, making such models good candidates to apply to this data set in future. Other authors have implemented functions $g(I)$ that obey specific conditions in the transmission term of the SIR model and thoroughly examined the effect of this on model dynamics as well [12].

5.3 Models With Social Structure

There has been a vast amount of work done on constructing models of the effect that social structure and human behaviour have on disease dynamics (for review see Funk *et al.* [19]). Areas of mathematics such as network and game theory are instrumental in this pursuit.

The SIR model assumes homogeneous mixing of the population, *i.e.,* each individual is just as likely to interact with each other individual. In real life, humans form social groups with which most of their interactions tend to occur. There has been a great amount of work done to build and analyze network models that take this into account in disease transmission [33, 34]. This type of network modelling would also be useful in the context of song spread. Social networks directly influence how songs spread through a population. As mentioned above, it is likely that music recommendations from friends will carry more weight than those from strangers and it is also not unlikely that individuals who form part of the same group of friends listen to similar types of music. Factors like these could be better accounted for with network models.

Game theory is also very useful in modelling the effects of human behaviour. For example, it is often used in infectious disease research to analyze vaccination scenarios [4, 5, 18, 32]. While it is difficult to imagine an analogue to 'choosing to be vaccinated' against a song, there are other human choices involved in song popularity that could be modelled using game theoretical models. These include deciding whether to download a single track or an album by an artist, and choosing to share

your opinion of a song through social media.

5.4 Multiple Infectious Classes

It may be worth reinterpreting the meaning of 'infectious' in the context of songs. After downloading a song, one could become infectious by forming a positive opinion of the song and actively sharing this positive opinion with others. However, one could also become infectious by forming a negative opinion of the song and actively sharing this opinion with others. The first individual described is spreading a preference for the song, while the second is spreading an aversion or repulsion to the song. Each can still be thought of as 'infectious' because they are 'carrying' the song and spreading an opinion about it, however the difference in opinion suggests that they should not be treated the same way in a modelling scenario. One way of dealing with this would be to implement two infectious classes, one for those who are transmitting song preference and are therefore 'positively infectious' and one for those who are transmitting song aversion and are therefore 'negatively infectious'.

Other authors have investigated the use of models with multiple infectious classes to study antibiotic resistance in tuberculosis [13] and the interactions between vector and host strains for Dengue fever [17]. Multiple infectious classes are also sometimes used to model diseases with progressive stages of infection, such as HIV [23], to create a model with a more realistic distribution for the infectious period [1, 26, 29], or to create different structure within classes, such as age structure [10, 39] or spatial structure [20].

In §6 we present a new model that includes multiple infectious classes.

6 A New Epidemic Model of Song Spread

The following model examines the idea of having individuals in the population who have downloaded a song, formed a negative opinion of this song and are actively voicing this negative opinion. These individuals are referred to as 'negatively infectious' and spread an aversion to the song, rather than a preference for the song.

6.1 Description

The model is given by the following set of equations:

$$\frac{dS}{dt} = -\beta_I SI - \beta_J SJ \tag{7a}$$

$$\frac{dI}{dt} = \beta_I SI(1-J) - \gamma_I I \tag{7b}$$

$$\frac{dJ}{dt} = \beta_I SIJ - \gamma_J J \tag{7c}$$

$$\frac{dR}{dt} = \beta_J SJ + \gamma_I I + \gamma_J J \tag{7d}$$

where:

S is the proportion of the population that is susceptible to the song in question,

I is the proportion of the population that has downloaded the song, likes the song and is therefore talking about it in a positive way ('positively infectious'),

J is the proportion of the population that has downloaded the song, doesn't like the song and is therefore talking about the song in a negative way ('negatively infectious'),

R is the proportion of the population who has either downloaded the song and is no longer interested in it (recovered), or has developed a negative opinion about the song as a result of hearing poor reviews from negatively infectious individuals and will therefore never download the song (*i.e.,* they are 'vaccinated' against the song),

$S + I + J + R = 1$ and $0 \leq S, I, J, R \leq 1$,

β_I is the transmission rate of for song preference,

β_J is the transmission rate of for song aversion,

γ_I is the recovery rate for being positively infectious,

γ_J is the recovery rate for being negatively infectious,

$\beta_I, \beta_J, \gamma_I, \gamma_J > 0$.

Susceptible individuals will interact with positively infectious individuals and negatively infectious individuals. Interactions with negatively infectious individuals

have some probability of influencing the susceptible to never download the song (or to become 'vaccinated' against the song). In this case, the individual will move directly to the recovered class R. This happens at a rate β_J. Interactions between susceptible individuals and positively infectious individuals may result in the susceptible individual downloading the song, which happens at a rate of β_I. After downloading the song, an individual will decide whether they like or dislike the song, thus becoming either positively or negatively infectious. This process is influenced by the general opinion of the population on the song; an individual's opinion could be formed based on subsequent interactions with other infectious individuals around them or through exposure to media relating to the song. Therefore, the probability that the individual will then become positively infectious is proportional to the number of individuals in the population who are not negatively infectious (and hence proportional to $(1-J)$). The probability that the individual will become negatively infectious is proportional to the number of individuals in the population who are already negatively infectious (and hence proportional to J). This means that there is an inherent assumption in the model that individuals who download the song are slightly more likely to develop a positive opinion of that song – we assume that they are inclined to like the song unless they hear a sufficient number negative reviews from negatively infectious individuals, are exposed to enough negative media relating to the song, etc. We also assume (as in the standard SIR model) that the population is well-mixed so that individuals have an equal probability of interacting with every other individual, making it reasonable to assume that the probability of becoming positively or negatively infectious depends, in effect, on the overall opinion of the population on a given song. The work mentioned previously by Berns *et al.* [6] supports this assumption, at least in adolescents. Positively infectious individuals recover (meaning they stop talking about how much they like the song) at a rate of γ_I and negatively infectious individuals recover at a rate of γ_J.

This model seeks chiefly to take into account the effect of individuals who do not like a song and are actively sharing this opinion. It does not directly address anything to do with radio, streaming services or media coverage. It will be referred to as the SIJR model.

6.2 Analysis

We can now conduct some basic analysis of our new model for song spread. We want to learn as much as possible about this model so as to determine whether it might help to explain aspects of song transmission that the SIR model failed to capture. Comparing the dynamics of this new model with the basic SIR model will help to

determine the effect of adding a second infectious class.

We will look at the basic reproduction number $\mathcal{R}_0$ and initial growth rate for this model, find equilibria and attempt to analyze their stability, and think about the influence of any new parameters that appear in the framework of this model.

6.2.1 Biologically Well-Defined

Since this is a new model, we start be checking that it is biologically well-defined. Note that since S', I' and J' do not involve R, the dynamical system is completely specified by the first three equations in Equation (7) above. It will therefore suffice to analyze the three-dimensional system given by:

$$\frac{dS}{dt} = -\beta_{_I} SI - \beta_{_J} SJ \tag{8a}$$

$$\frac{dI}{dt} = \beta_{_I} SI(1-J) - \gamma_{_I} I \tag{8b}$$

$$\frac{dJ}{dt} = \beta_{_I} SIJ - \gamma_{_J} J \tag{8c}$$

In order to show that the model is biologically well-defined we must show that solutions that start in the biologically relevant region will remain there, *i.e.,* the region where $S + I + J + R = 1$ (and $S, I, J, R \geq 0$) or in the three-dimensional case where $0 \leq S + I + J \leq 1$ (and $S, I, J \geq 0$).

Notice that if $S = 0$ then $S' = 0$. This means that solutions in the biologically relevant region will not cross the $S = 0$ plane. Similarly, solutions in the biologically relevant region will not cross the $I = 0$ plane or the $J = 0$ plane since if $I = 0$ then $I' = 0$ and if $J = 0$ then $J' = 0$. Lastly, we can also show that if $S + I + J \leq 1$ (and individually $0 \leq S, I, J \leq 1$) this remains true. Notice that:

$$(S + I + J)' = S' + I' + J' = -R' \tag{9a}$$

$$= -(\beta_{_J} SJ + \gamma_{_I} I + \gamma_{_J} J) \tag{9b}$$

$$= -J(\beta_{_J} S + \gamma_{_J}) - \gamma_{_I} I \tag{9c}$$

If $S + I + J \leq 1$ and $0 \leq S, I, J \leq 1$ this will always be non-positive, meaning that solutions will stay inside the biologically relevant region.

6.2.2 The Basic Reproduction Number $\mathcal{R}_0$

Next, we determine what the basic reproduction number $\mathcal{R}_0$ is in this model. van den Driessche and Watmough define $\mathcal{R}_0$ as 'the number of new infections produced by a

typical infective individual in a population at a disease free equilibrium (DFE)' [41]. Note that this does not have to be a unique DFE. We will determine the basic reproduction number for our model using the method presented in van den Driessche and Watmough's 2002 paper (for full description see [41]). The equations of our new model (see Equation (7)) are first arranged so that equations representing the rate of change of infected compartments appear first, followed by those that represent rate of change of other compartments:

$$\frac{dI}{dt} = \beta_{_I} SI(1-J) - \gamma_{_I} I \tag{10a}$$

$$\frac{dJ}{dt} = \beta_{_I} SIJ - \gamma_{_J} J \tag{10b}$$

$$\frac{dS}{dt} = -\beta_{_I} SI - \beta_{_J} SJ \tag{10c}$$

$$\frac{dR}{dt} = \beta_{_J} SJ + \gamma_{_I} I + \gamma_{_J} J \tag{10d}$$

From Lemma 6.1 below, the disease free equilibria in this system occur in the line described by $I = 0$, $J = 0$, $S = S_{_1}^*$ and $R = 1 - S_{_1}^*$. Thus, one disease free equilibrium (not unique) of this system occurs when $I = 0$, $J = 0$, $S = 1$ and $R = 0$. This equilibrium corresponds to the case of a wholly susceptible population, which is the scenario that is considered in the definition of $\mathcal{R}_0$.

Let $\mathbf{X}$ be the vector (I, J, S, R). The function $\mathscr{F}(\mathbf{X})$ is constructed using the rates of appearance of new infections in each compartment:

$$\mathscr{F}(\mathbf{X}) = \begin{bmatrix} \beta_{_I} SI(1-J) \\ \beta_{_I} SIJ \\ 0 \\ 0 \end{bmatrix} \tag{11}$$

The function $\mathscr{V}(\mathbf{X})$ is constructed using the rates at which individuals move into and out of compartments by means other than infection. For each compartment, the rate at which individuals move into the compartment is subtracted from the rate at which they move out of the compartment to give:

$$\mathscr{V}(\mathbf{X}) = \begin{bmatrix} \gamma_{_I} I \\ \gamma_{_J} J \\ \beta_{_I} SI + \beta_{_J} SJ \\ \beta_{_J} SJ - \gamma_{_I} I - \gamma_{_J} J \end{bmatrix} \tag{12}$$

Diekmann *et al.* define $\mathcal{R}_0$ as the spectral radius of the next generation matrix [15]. van den Driessche and Watmough [41] prove that for the class of models they consider

(which includes this one), the next generation matrix can be written as FV^{-1} where:

$$F = \left[\begin{matrix} \frac{\partial \mathscr{F}}{\partial I} & \frac{\partial \mathscr{F}}{\partial J} \\ \frac{\partial \mathscr{F}}{\partial I} & \frac{\partial \mathscr{F}}{\partial J} \end{matrix}\right]\Bigg|_{\mathbf{X}=\mathbf{X}^*} \tag{13}$$

and:

$$V = \left[\begin{matrix} \frac{\partial \mathscr{V}}{\partial I} & \frac{\partial \mathscr{V}}{\partial J} \\ \frac{\partial \mathscr{V}}{\partial I} & \frac{\partial \mathscr{V}}{\partial J} \end{matrix}\right]\Bigg|_{\mathbf{X}=\mathbf{X}^*} \tag{14}$$

$\mathbf{X}^*$ is a disease free equilibrium point. Evaluating F at $\mathbf{X} = (0, 0, 1, 0)$ gives:

$$F = \begin{bmatrix} \beta_I S(1-J) & -\beta_I SI \\ \beta_I SJ & \beta_I SI \end{bmatrix}\Big|_{\mathbf{X}=(0,0,1,0)} \tag{15a}$$

$$= \begin{bmatrix} \beta_I & 0 \\ 0 & 0 \end{bmatrix} \tag{15b}$$

Similarly, evaluating V at $\mathbf{X} = (0, 0, 1, 0)$ gives:

$$V = \begin{bmatrix} \gamma_I & 0 \\ 0 & \gamma_J \end{bmatrix}\Big|_{\mathbf{X}=(0,0,1,0)} \tag{16a}$$

$$= \begin{bmatrix} \gamma_I & 0 \\ 0 & \gamma_J \end{bmatrix} \tag{16b}$$

Then V^{-1} is:

$$V^{-1} = \frac{1}{\gamma_I \gamma_J} \begin{bmatrix} \gamma_J & 0 \\ 0 & \gamma_I \end{bmatrix} \tag{17a}$$

$$= \begin{bmatrix} 1/\gamma_I & 0 \\ 0 & 1/\gamma_J \end{bmatrix} \tag{17b}$$

Taking FV^{-1} yields:

$$FV^{-1} = \begin{bmatrix} \beta_I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\gamma_I & 0 \\ 0 & 1/\gamma_J \end{bmatrix} \tag{18a}$$

$$= \begin{bmatrix} \beta_I/\gamma_I & 0 \\ 0 & 0 \end{bmatrix} \tag{18b}$$

The eigenvalues of FV^{-1} are β_I/γ_I and 0, hence the spectral radius of FV^{-1} is:

$$\rho(FV^{-1}) = \frac{\beta_I}{\gamma_I} \tag{19}$$

meaning that for our model $\mathcal{R}_0 = \beta_I/\gamma_I$.

Therefore, the basic reproduction number for our new model is $\mathcal{R}_0 = \beta_I/\gamma_I$. This result is rather intuitive. Since only positively infectious individuals influence susceptibles to download a song (and thus become either positively *or* negatively infectious), it is logical that the basic reproduction number for this system would be the transmission rate for individuals in the I class multiplied by the average length of time that individuals spend in this class. Another way of thinking about this that relates to the definition of $\mathcal{R}_0$ given by van den Driessche and Watmough is that in a wholly susceptible population the average negatively infectious individual produces an average of 0 new infections, while the average positively infectious individual produces an average of β_I/γ_I new infections.

6.2.3 Equilibria

Lemma 6.1. *Let* $\mathbf{Y}$ *be the vector* $\mathbf{Y} = (S, I, J)$. *There are two equilibria in the system described by Equation* (8) *above. The first is a line of disease free equilibria of the form* $\mathbf{Y}_1^* = (S_1^*, 0, 0)$, *where* $0 \leq S_1^* \leq 1$. *The second equilibrium is of the form* $\mathbf{Y}_2^* = \left(\frac{1}{\mathcal{R}_0} - \frac{1}{\mathcal{R}_J}, \frac{\beta_J/\gamma_I}{\mathcal{R}_J - \mathcal{R}_0}, \frac{\mathcal{R}_0}{\mathcal{R}_0 - \mathcal{R}_J}\right)$, *where* $\mathcal{R}_0 = \beta_I/\gamma_I$, $\mathcal{R}_J = \beta_J/\gamma_J$ *and* $\mathcal{R}_0 \neq \mathcal{R}_J$. *The second equilibrium is never biologically relevant and if* $\mathcal{R}_0 = \mathcal{R}_J > 0$ *the second equilibrium* $\mathbf{Y}_2^*$ *disappears.*

Proof. It continues to be sufficient to analyze the three-dimensional system given above in Equation (8). We begin by finding the nullclines for S, I and J. S-nullclines occur when:

$$0 = -\beta_I SI - \beta_J SJ \tag{20a}$$
$$0 = -S(\beta_I I + \beta_J J) \tag{20b}$$

This is true if:

$$S = 0 \tag{21}$$

or if:

$$\beta_I I + \beta_J J = 0 \tag{22a}$$
$$I = -\frac{\beta_J}{\beta_I} J \tag{22b}$$

making these our two S-nullclines. I-nullclines occur when:

$$0 = \beta_I SI(1 - J) - \gamma_I I \tag{23a}$$
$$0 = I(\beta_I S(1 - J) - \gamma_I) \tag{23b}$$

This is true if:

$$I = 0 \tag{24}$$

or if:

$$\beta_{_I} S(1-J) - \gamma_{_I} = 0 \tag{25a}$$

$$S = \frac{\gamma_{_I}}{\beta_{_I}} \frac{1}{1-J} \tag{25b}$$

making these our two I-nullclines. Note that Equation (25) could also be written as $S = (1/\mathcal{R}_0)(1/[1-J])$ and that for this nullcline $J \neq 1$. Finally, J-nullclines occur when:

$$0 = \beta_{_I} SIJ - \gamma_{_J} J \tag{26a}$$

$$0 = J(\beta SI - \gamma_{_J}) \tag{26b}$$

This is true if:

$$J = 0 \tag{27}$$

or if:

$$\beta_{_I} SI - \gamma_{_J} = 0 \tag{28a}$$

$$S = \frac{\gamma_{_J}}{\beta_{_I} I} \tag{28b}$$

making these our two J-nullclines. Note that for the second nullcline, $I \neq 0$.

In order to find equilibria, we start by letting $S = 0$ ($\implies$ $S' = 0$). If this is the case, then the I-nullclines (Equations (24) and (25)) tell us that in order to have $I' = 0$, either $I = 0$ or $\gamma_{_I} = 0$. According to the definitions of our model, $\gamma_{_I} > 0$, so it must be true that $I = 0$. Similarly, in order to have $J' = 0$, Equations (26) and (28) tell us that either $J = 0$ or $\gamma_{_J} = 0$. Since $\gamma_{_J} > 0$ by definition, it must be true that $J = 0$. Therefore, our first equilibrium point is $\mathbf{Y}_{_1}^* = (0, 0, 0)$.

Next let $I = (-\beta_{_J}/\beta_{_I})J$ ($\implies$ $S' = 0$). In order to have $I' = 0$ we must have either $I = 0$ or $S = \frac{\gamma_I}{\beta_I}\frac{1}{1-J}$. If $I = 0$, then:

$$I = (-\beta_{_J}/\beta_{_I})J \implies J = 0, \tag{29}$$

which means that $J' = 0$. In this case, S can take on any value in the interval $[0, 1]$, thus creating a line of equilibria. Notice that $\mathbf{Y}_{_1}^*$ is included in this line. We therefore amend our definition of $\mathbf{Y}_{_1}^*$ to be the line of disease free equilibria defined by $\mathbf{Y}_{_1}^* = (S_{_1}^*, 0, 0)$, where $0 \leq S_{_1}^* \leq 1$.

Next we examine the case where $I = (-\beta_J/\beta_I)J$ ($\implies S' = 0$ and $J \neq 1$) and $S = \frac{\gamma_I}{\beta_I}\frac{1}{1-J}$ ($\implies I' = 0$). In order to have $J' = 0$ we must have either $J = 0$ or $S = \gamma_J/(\beta_I I)$. If $J = 0$, then:

$$I = -\frac{\beta_J}{\beta_I}J \implies I = 0, \tag{30}$$

which would mean that we were looking at points on the line of DFEs again. Instead, let us assume that $J \neq 0$ so that we must have $S = \gamma_J/(\beta_I I)$ ($\implies I \neq 0$). It is interesting to note that by assuming at this point that $J \neq 0$, we are effectively differentiating from the basic SIR setup. Since it must also be true that $S = \frac{\gamma_I}{\beta_I}\frac{1}{1-J}$, we have:

$$\frac{\gamma_I}{\beta_I}\frac{1}{1-J} = \frac{\gamma_J}{\beta_I I} \tag{31a}$$

$$\frac{\gamma_I}{1-J} = \frac{\gamma_J}{I} \tag{31b}$$

$$I = \frac{\gamma_J}{\gamma_I}(1-J) \tag{31c}$$

It must also be true that $I = (-\beta_J/\beta_I)J$. Combining this with the above result gives:

$$-\frac{\beta_J}{\beta_I}J = \frac{\gamma_J}{\gamma_I}(1-J) \tag{32a}$$

$$\left(\frac{\gamma_J}{\gamma_I} - \frac{\beta_J}{\beta_I}\right)J = \frac{\gamma_J}{\gamma_I} \tag{32b}$$

$$\left(\frac{1}{\gamma_I} - \frac{\beta_J}{\gamma_J\beta_I}\right)J = \frac{1}{\gamma_I} \tag{32c}$$

$$\left(\frac{\beta_I}{\gamma_I} - \frac{\beta_J}{\gamma_J}\right)J = \frac{\beta_I}{\gamma_I} \tag{32d}$$

$$J = \frac{\frac{\beta_I}{\gamma_I}}{\left(\frac{\beta_I}{\gamma_I} - \frac{\beta_J}{\gamma_J}\right)} \tag{32e}$$

Let $\mathcal{R}_J = \beta_J/\gamma_J$. Notice that J can also be expressed as:

$$J = \frac{\mathcal{R}_0}{\mathcal{R}_0 - \mathcal{R}_J} \tag{33}$$

We find I by substituting this expression into the relationship $I = (-\beta_J/\beta_I)J$:

$$I = -\frac{\beta_J}{\beta_I} J \tag{34a}$$

$$= -\frac{\beta_J}{\beta_I} \frac{\frac{\beta_I}{\gamma_I}}{\left(\frac{\beta_I}{\gamma_I} - \frac{\beta_J}{\gamma_J}\right)} \tag{34b}$$

$$= -\frac{\beta_J}{\gamma_I} \frac{1}{\left(\frac{\beta_I}{\gamma_I} - \frac{\beta_J}{\gamma_J}\right)} \tag{34c}$$

$$= \frac{\frac{\beta_J}{\gamma_I}}{\left(\frac{\beta_J}{\gamma_J} - \frac{\beta_I}{\gamma_I}\right)} \tag{34d}$$

This can be verified by substituting J into $I = (\gamma_J/\gamma_I)(1 - J)$. Notice that I can also be written as:

$$I = \frac{\beta_J/\gamma_I}{\mathcal{R}_J - \mathcal{R}_0} \tag{35}$$

or as:

$$I = \frac{1}{\left(\frac{\gamma_I}{\gamma_J} - \frac{\beta_I}{\beta_J}\right)} \tag{36}$$

Next we find S by substituting I into the relationship $S = \gamma_J/(\beta_I I)$:

$$S = \frac{\gamma_J}{\beta_I I} \tag{37a}$$

$$= \frac{\gamma_J}{\beta_I} \left(\frac{\gamma_I}{\gamma_J} - \frac{\beta_I}{\beta_J}\right) \tag{37b}$$

$$= \frac{\gamma_I}{\beta_I} - \frac{\gamma_J}{\beta_J} \tag{37c}$$

Notice that S could also be written as:

$$S = \frac{1}{\mathcal{R}_0} - \frac{1}{\mathcal{R}_J} \tag{38}$$

The second equilibrium for our model $\mathbf{Y}_2^* = (S_2^*, I_2^*, J_2^*)$ is therefore:

$$\mathbf{Y}_2^* = \left(\frac{\gamma_I}{\beta_I} - \frac{\gamma_J}{\beta_J}, \frac{1}{\frac{\gamma_I}{\gamma_J} - \frac{\beta_I}{\beta_J}}, \frac{\frac{\beta_I}{\gamma_I}}{\left(\frac{\beta_I}{\gamma_I} - \frac{\beta_J}{\gamma_J}\right)}\right) \tag{39}$$

It can also be expressed as:

$$\mathbf{Y}^*_2 = \left(\frac{1}{\mathcal{R}_0} - \frac{1}{\mathcal{R}_J}, \frac{\beta_J/\gamma_I}{\mathcal{R}_J - \mathcal{R}_0}, \frac{\mathcal{R}_0}{\mathcal{R}_0 - \mathcal{R}_J}\right) \tag{40}$$

What if $\mathcal{R}_0 = \mathcal{R}_J$? In this case, Equation (32d) would give us:

$$0 \cdot J = \frac{\beta_I}{\gamma_I} = \mathcal{R}_0 \tag{41}$$

The definitions of β_I and γ_I make this impossible. Since there is no reason why it should not be true that $\mathcal{R}_0 = \mathcal{R}_J$, we can conclude that if $\mathcal{R}_0 = \mathcal{R}_J$ then the second equilibrium $\mathbf{Y}^*_2$ does not exist. By the definitions of $\beta_I, \beta_J, \gamma_I, \gamma_J$, $\mathcal{R}_0, \mathcal{R}_J > 0$.

$\mathbf{Y}^*_2$ and the line of disease free equilibria represented by $\mathbf{Y}^*_1$ are the only possible equilibria for our new model. When is each equilibrium biologically relevant? The disease free equilibria are relevant so long as $0 \leq S^*_1 \leq 1$. However, $\mathbf{Y}^*_2$ is actually never biologically relevant. We assume $\mathcal{R}_0 \neq \mathcal{R}_J$. If $\mathcal{R}_0 > \mathcal{R}_J$ then:

$$\mathcal{R}_0 > \mathcal{R}_0 - \mathcal{R}_J \implies \frac{\mathcal{R}_0}{\mathcal{R}_0 - \mathcal{R}_J} > 1 \tag{42}$$

i.e., $J^*_2 > 1$ and the second equilibrium is therefore outside the biologically relevant region. If $\mathcal{R}_0 < \mathcal{R}_J$ then:

$$\mathcal{R}_0 - \mathcal{R}_J < 0 \implies \frac{\mathcal{R}_0}{\mathcal{R}_0 - \mathcal{R}_J} < 0 \tag{43}$$

i.e., $J^*_2 < 0$ and the second equilibrium is therefore still outside the biologically relevant region.

□

6.2.4 Stability of Equilibria

The next step is to analyze the stability of our equilibria. We first find the Jacobian matrix of our three-dimensional system (see Equation (8)), which will be notated as $Df_{\mathbf{Y}^*}$:

$$Df_{\mathbf{Y}^*} = \begin{bmatrix} -\beta_I I - \beta_J J & -\beta_I S & -\beta_J S \\ \beta_I I(1-J) & \beta_I S(1-J) - \gamma_I & -\beta_I SI \\ \beta_I IJ & \beta_I SJ & \beta_I SI - \gamma_J \end{bmatrix} \tag{44}$$

This can be used to examine the linearized system $\mathbf{Y}' = Df_{\mathbf{Y}^*}(\mathbf{Y} - \mathbf{Y}^*)$. We will start with the line of disease free equilibria $\mathbf{Y}^*_1 = (S^*_1, 0, 0)$. In this case, the Jacobian

is:

$$Df_{\mathbf{Y}_1^*} = \begin{bmatrix} 0 & \beta_I S_1^* & -\beta_J S_1^* \\ 0 & \beta_I S_1^* - \gamma_I & 0 \\ 0 & 0 & -\gamma_J \end{bmatrix} \tag{45}$$

This is an upper triangular matrix, meaning that the eigenvalues can be read off of the diagonal. This means that the equilibria in the line of disease free equilibria are all non-hyperbolic and that the linearization cannot tell us anything about their stability. Note that an equilibrium point that forms part of a line of equilibria cannot be asymptotically stable. In a linear system, a line of equilibria can be a repelling or attracting line, or part of the degenerate case where all points in the system are equilibria. In any of these three cases, at least one of the corresponding eigenvalues for the system would be 0. It is therefore not surprising that the linearized system cannot tell us anything about the line of disease free equilibria in our nonlinear system. Is there some other way that we can analyze the stability of this line of equilibria?

Consider the following generalized version of Lyapunov's stability theorem, which can be proved in a similar manner to how the standard theorem is proved (as is done, for example, in Hirsch *et al.* [22, pg. 196]):

Theorem 6.2. *Consider a closed invariant set $\mathcal{C}$ of $\mathbf{Y}' = f(\mathbf{Y})$ and an open set $\mathcal{O}$ that contains $\mathcal{C}$. If $\exists$ a differentiable function $L : \mathcal{O} \to \mathbb{R}$ such that:*

(a) $L(\mathbf{Y}) = 0 \quad \forall \mathbf{Y} \in \mathcal{C}$ *and* $L(\mathbf{Y}) > 0 \quad \forall \mathbf{Y} \in \mathcal{O} \setminus \mathcal{C}$

(b) $\dot{L}(\mathbf{Y}) \leq 0 \quad \forall \mathbf{Y} \in \mathcal{O} \setminus \mathcal{C}$

then $\mathcal{C}$ is stable and L is a Lyapunov function. If in addition,

(c) $\dot{L}(\mathbf{Y}) < 0 \quad \forall \mathbf{Y} \in \mathcal{O} \setminus \mathcal{C}$

then $\mathcal{C}$ is asymptotically stable and L is a strict Lyapunov function.

We are considering the biologically relevant region:

$$\Delta = \{(S, I, J) : S, I, J \geq 0 \,\text{and}\, S + I + J \leq 1\} \tag{46}$$

Note that Δ is an open set when considering the relative topology on Δ, *i.e.,* it is open relative to itself. Let Δ be our open set $\mathcal{O}$ and the line of disease free equilibria $\mathbf{Y}_1^* = (S_1^*, 0, 0)$ be our closed invariant set $\mathcal{C}$. Note that $\mathcal{C}$ is indeed invariant since all points in this set are equilibria.

Consider the function $L(\mathbf{Y}) = I+J$. $L(\mathbf{Y}) = 0$ for all points $\mathbf{Y} \in \mathcal{C}$ and $L(\mathbf{Y}) > 0$ for all points $\mathbf{Y} \in \mathcal{O} \setminus \mathcal{C}$ as required by condition (a). $\dot{L}(\mathbf{Y})$ is given by:

$$\begin{aligned}
\dot{L}(\mathbf{Y}) &= \nabla L \cdot f(\mathbf{Y}) && \text{(47a)}\\
&= I' + J' && \text{(47b)}\\
&= \beta_{_I} SI(1-J)\gamma_{_I} I + \beta_{_I} SIJ - \gamma_{_J} J && \text{(47c)}\\
&= \beta_{_I} SI - \gamma_{_I} I - \gamma_{_J} J && \text{(47d)}\\
&= I(\beta_{_I} S - \gamma_{_I}) - \gamma_{_J} J && \text{(47e)}
\end{aligned}$$

We are interested in the case where $\dot{L} < 0$ for all points $\mathbf{Y} \in \mathcal{O} \setminus \mathcal{C}$. In order for this to be true:

$$\begin{aligned}
I(\beta S - \gamma_{_I}) - \gamma_{_J} J < 0 && \text{(48a)}\\
\frac{I}{\gamma_{_I}}(\mathcal{R}_0 S - 1) - \frac{\gamma_{_J}}{\gamma_{_I}} J < 0 && \text{(48b)}\\
I(\mathcal{R}_0 S - 1) - \gamma_{_J} J < 0 && \text{(48c)}
\end{aligned}$$

Since $\gamma_{_J} > 0$ and $J > 0$ in the set $\mathcal{O} \setminus \mathcal{C}$, it will always be true that $-\gamma_{_J} J < 0$ on the set $\mathcal{O} \setminus \mathcal{C}$. I is also greater than 0 on the set $\mathcal{O} \setminus \mathcal{C}$ so we must determine when $\mathcal{R}_0 S - 1 < 0$, *i.e.,* when $\mathcal{R}_0 S < 1$. $S = 1 - I - J$ and since $I > 0$, $J > 0$ on the set $\mathcal{O} \setminus \mathcal{C}$, it must be true that $S < 1$ on this set. Therefore $\mathcal{R}_0 S < 1$ if $\mathcal{R}_0 \leq 1$, meaning $\mathcal{C}$ is asymptotically stable on Δ if $\mathcal{R}_0 \leq 1$. In other words, the line of disease free equilibria is a globally attracting line if the basic reproduction number is less than or equal to 1, which is a biologically sensible conclusion.

Next we look at the system linearized about $\mathbf{Y}^*_{_2}$. Although we have already determined that this equilibrium is not biologically relevant, it is still interesting to examine its stability. In this case, the Jacobian is given by:

$$Df_{\mathbf{Y}^*_{_2}} = \begin{bmatrix} 0 & \frac{\beta_{_I}\gamma_{_J}}{\beta_{_J}} - \gamma_{_I} & \gamma_{_J} - \frac{\beta_{_J}\gamma_{_I}}{\beta_{_I}} \\ \frac{\beta_{_J}^2 \beta_{_I} \gamma_{_J} \gamma_{_I}}{\left(\beta_{_I}\gamma_{_J} - \beta_{_J}\gamma_{_I}\right)^2} & 0 & -\gamma_{_J} \\ -\frac{\beta_{_J}\beta_{_I}^2\gamma_{_J}^2}{\left(\beta_{_I}\gamma_{_J} - \beta_{_J}\gamma_{_I}\right)^2} & -\frac{\beta_{_I}\gamma_{_J}}{\beta_{_J}} & 0 \end{bmatrix} \qquad (49)$$

The Routh-Hurwitz condition tells us that the eigenvalues of $Df_{\mathbf{Y}^*_{_2}}$ all have negative real parts if and only if:

1. $\text{trace}(Df_{\mathbf{Y}^*_{_2}}) < 0$

2. $\det(Df_{\mathbf{Y}_2^*}) < 0$

3. $\text{trace}(Df_{\mathbf{Y}_2^*}) \times M - \det(Df_{\mathbf{Y}_2^*}) < 0$

where M is the sum of the second-order principal minors of $Df_{\mathbf{Y}_2^*}$ [11]. Since $\text{trace}(Df_{\mathbf{Y}_2^*}) = 0$, the eigenvalues of $Df_{\mathbf{Y}_2^*}$ cannot all have negative parts. This tells us that if $\mathbf{Y}_2^*$ is a hyperbolic equilibrium, it is not stable. Unfortunately, the characteristic equation for $Df_{\mathbf{Y}_2^*}$ is too complicated to deduce whether this is a hyperbolic equilibrium, so we are unable to glean any more information from the linearization about $\mathbf{Y}_2^*$. More advanced methods are required to further analyze the stability of this equilibrium, however seeing as it is not biologically relevant we have yet to pursue these methods.

6.2.5 Initial Growth Rate r

The initial growth rate for this model can be determined by examining the linearized model about the disease free equilibria (see Equation (45) above). The real part of the eigenvalue with the largest positive real part for this system is the growth rate for our model. The eigenvalues of Equation (45) are 0, $\beta_I S_1^* - \gamma_I$ and $-\gamma_J$. Since $\gamma_J > 0$ by definition, $-\gamma_J$ is always negative. We therefore turn our attention to $\beta_I S_1^* - \gamma_I$. $\beta_I S_1^* - \gamma_I > 0$ if:

$$\frac{\beta_I}{\gamma_I} S_1^* - 1 > 0 \tag{50a}$$

$$\mathcal{R}_0 S_1^* > 1 \tag{50b}$$

In other words, the initial growth rate will be positive if the effective basic reproduction number $\mathcal{R}_0 S_1^*$ is greater than 1. If $\mathcal{R}_0 S_1^* \leq 1$, the initial growth rate will be non-positive, *i.e.,* there will not be any initial growth.

6.3 What Next?

Further analysis of the SIJR model will help to better understand its dynamics, how they differ from the SIR model and how they are similar. It would be helpful to derive formulae for the final size and epidemic peak height. Looking at the typical time series that can be produced by this model would also be useful in helping to determine whether the SIJR model would be able to capture aspects of song transmission that the SIR model was unable to, or whether we need to include more biological information in the model before it is suited to the scenario we are

attempting to model. For example, it is unclear whether the SIJR model is able to capture seasonal epidemics or damped oscillations either.

The basic SIR model should be a limiting case of the SIJR model when $J = 0$. It would therefore be informative to look more closely at the dynamics of the SIJR model as this limit is approached. The relationship between $\mathcal{R}_0$ and $\mathcal{R}_J$ in the SIJR model also warrants further exploration, in particular the transition from $\mathcal{R}_0 > \mathcal{R}_J$ (*i.e.,* when on average more people are influenced to download a song by a positively infectious individual than to never download it by a negatively infectious individual) to $\mathcal{R}_0 < \mathcal{R}_J$ (when the opposite scenario is true). It might also be useful to examine the effects of this transition on the second equilibrium and to think more about the interpretation of this second equilibrium, which does not exist in the basic SIR model.

7 Conclusions and Future Directions

This thesis has presented exploratory work in applying a standard epidemiological model to song download time series in order to study song popularity. We believe these preliminary results show that epidemic models offer a powerful tool for analyzing music downloading trends and studying the mechanisms that drive music popularity. Download time series for popular songs are often similar in shape to epidemic curves, making models like the SIR model well-suited for fitting to these curves. Epidemiological parameters extracted from these fits can be interpreted in the context of song preference transmission, and analyzed and compared to draw new conclusions about how songs become popular. Because of the exploratory nature of this work, there are many possible directions for future exploration.

The basic SIR model is able to capture some aspects of 'song transmission', however it does not appear to be the best model to describe and analyze this process. Each of the models discussed in §5 covers a different aspect of biology and/or song spread that is not accounted for by the basic SIR model. Fitting models of these kinds to our data set and comparing the results of each might help to determine which biological aspects of song transmission it is most important to include in a model of this process. These could include vital dynamics, details of transmission that would be better portrayed by a nonlinear incidence term, or a social network that is not homogeneously mixed. Currently, we only have access to song *download* data. If we were to obtain access to live-streaming data, it would also be worth considering applying a model that accounts for decay in immunity to the data set, such as the SIS or SIRS model [16]. Various formats of these have been thoroughly analyzed by Liu *et al.* [27,28]. Another concept from epidemiological research which

might be interesting to pursue in the context of song download epidemics is the idea of super spreading, *i.e.,* the idea that $\mathcal{R}_0$ can vary substantially in a population so that certain individuals have a much higher degree of infectiousness than others [30]. In the context of songs, a super spreader might be someone who expresses their opinion of a song much more often and readily and/or more strongly/passionately, possibly through social media. Another potential avenue of future research would be to try to identify characteristics of song super spreaders that would be visible in the database.

As mentioned in §5.1, applying a model with vital dynamics to our data set is also one approach to examining the difference between users who pay to download a track and users who download a track for free. Every user starts the MixRadio service with a one year period during which they can download music for free, and after that must pay for their music. In this work, we assumed that whether or not a user paid to download a track would not have any significant implications for our results. However, it may be important to further investigate this issue. When examining the download time series for a song, it might be informative to treat the users who paid for the song as a separate population from the users who downloaded the song for free. Users who downloaded the song for free would have a one year lifespan and the average lifespan of users who paid for the song could be estimated. At present, we do not have a method for calculating the lifespan of users – this is an area of future research as it will help in the application of any model that includes vital dynamics and particularly in the application of a model that includes vital dynamics to the issue of looking at paid vs. free downloads.

A new model was developed in this thesis specifically to describe the transmission of songs. This model included a second 'negatively infectious' class in an effort to account for the idea that some individuals will openly dislike a song after downloading it, thus spreading an aversion to that song. Future work will include analyzing this model in more detail, as described in §6.3. It will also include fitting this model to our song download data and comparing our results with those from fitting the basic SIR model to determine the effect of including a negatively infectious class.

In the work described here, we focused on downloads in Great Britain. In future, the study could be expanded to other countries or to look at worldwide trends. It would also be interesting to look at song sets other than the top 1000 downloaded songs. For example, it might also be informative to analyze songs that do not become hits to see if their download data time series possess any characteristics similar to time series for infectious disease outbreaks that fizzle before they become epidemics. There are many questions that could be applied to artist case studies. Do songs from the same artists follow similar 'epidemic patterns'? Do they have similar results for

calculated initially susceptible populations? When artists record a song together, do their subsequent individual tracks share download epidemic trends? When an artist releases a new song, do their older songs experience 'recurrent epidemics'?

We only looked for linear relationships among pairs of parameters in this study. In future, more sophisticated statistical comparisons could be considered. No genre patterns emerged in our parameter comparisons. It is possible that patterns might emerge when considering audio characteristics of songs, such as tempo or key.

The work done in this thesis is clearly just the beginning of how epidemic models might be used to study song popularity. There are a multitude of directions to pursue in the future that will help us to learn more about how songs become popular and how the mechanisms that drive song popularity relate to those that drive disease epidemics.

References

[1] D. Anderson and R. Watson. On the spread of a disease with gamma distributed latent and infectious periods. *Biometrika*, 67(1):191–198, 1980.

[2] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1991.

[3] J. Bansal. Unlocking the Predictive Power of Personality on Music-Genre Exclusivity. Unpublished BSc. thesis, McMaster University, Canada, 2015.

[4] C. T. Bauch and D. J. D. Earn. Vaccination and the theory of games. *PNAS – Proceedings of the National Academy of Sciences of the U.S.A.*, 101(36):13391–13394, 2004.

[5] C. T. Bauch, M. Li, G. Chapman, and A. P. Galvani. Adherence to cervical screening in the era of human papillomavirus vaccination: how low is too low? *The Lancet Infectious Diseases*, 10(2):133–137, 2010.

[6] G. S. Berns, C. M. Capra, S. Moore, and C. Noussair. Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage*, 49(3):2687–2696, 2010.

[7] G. S. Berns and S. E. Moore. A neural predictor of cultural popularity. *Journal of Consumer Psychology*, 22:154–160, 2011.

[8] K. Bischoff, C.S. Firan, M. Georgescu, W. Nejdl, and R. Paiu. Social knowledge-driven music hit prediction. In R. Huang, Q. Yang, J. Pei, J. Gama, X. Meng, and X. Li, editors, *Advanced Data Mining and Applications*, pages 43–54. Springer-Verlag Berlin Heidelberg, 2009.

[9] B. Bolker, J. Dushoff, D. J. D. Earn, I. Papst, and D.P. Rosati. *fitsir: SIR fitting tools*. `R` package version 0.1.0.

[10] B. M. Bolker and B. T. Grenfell. Chaos and biological complexity in measles dynamics. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, 251:75–81, 1993.

[11] F. Brauer and C. Castillo-Chavez. *Mathematical models in population biology and epidemiology*, volume 40 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2001.

[12] V. Capasso and G. Serio. A Generalization of the Kermack-McKendrick Deterministic Model. *Mathematical Biosciences*, 42(1):43–61, 1978.

[13] C. Castillo-Chavez and Feng Z. To treat or not to treat: the case of tuberculosis. *Journal of Mathematical Biology*, 35(6):629–656, 1997.

[14] R. Dhanaraj and B. Logan. Automatic Prediction of Hit Songs. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 488–491, 2005.

[15] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious-diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):18, 1990.

[16] J. Dushoff, J. B. Plotkin, S. A. Levin, and D. J. D. Earn. Dynamical resonance can account for seasonality of influenza epidemics. *PNAS – Proceedings of the National Academy of Sciences of the U.S.A.*, 101(48):16915–16916, 2004.

[17] Z. Feng and J. X. Velasco-Hernández. Competitive exclusion in a vector-host model for the dengue fever. *Journal of Mathematical Biology*, 35(5):523–544, 1997.

[18] F. Fu, D. I. Rosembloom, L. Wang, and M. A. Nowak. Imitation dynamics of vaccination behaviour on social networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1702):42–49, 2011.

[19] S. Funk, M. Salathe, and V. A. A. Jansen. Modelling the influence of human behaviour on the spread of infectious diseases: A review. *Journal of the Royal Society Interface*, 7:1247–1256, 2010.

[20] B. T. Grenfell, O. N. Bjornstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[21] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[22] M. W. Hirsch, S. Smale, and R. L. Devaney. *Differential equations, dynamical systems, and an introduction to chaos.* Academic Press, Waltham, MA, 3rd edition, 2013.

[23] J. M. Hyman, J. Li, and E. A. Stanley. The differential infectivity and staged progression models for the transmission of HIV. *Mathematical Biosciences*, 155(2):77–109, 1999.

[24] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A*, 115:700–721, 1927.

[25] Y. Kim, B. Suh, and K. Lee. #nowplaying the Future Billboard: Mining Music Listening Behaviors of Twitter Users for Hit Song Prediction. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, pages 51–56, 2014.

[26] O. Krylova and D. J. D. Earn. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of the Royal Society Interface*, 10:20130098, 2013.

[27] W. M. Liu, H. W. Hethcote, and S. A. Levin. Dynamical behavior of epidemiological models with nonlinear incidence rates. *Journal of Mathematical Biology*, 25(4):359–380, 1987.

[28] W. M. Liu, S. A. Levin, and Y. Iwasa. Influence of nonlinear incidence rates upon the behavior of SIRS epidemiological models. *Journal of Mathematical Biology*, 23(2):187–204, 1986.

[29] A. L. Lloyd. Spatio-temporal dynamics of childhood diseases in the us, 2001.

[30] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438:355–359, 2005.

[31] J. Ma and D. J. D. Earn. Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bulletin of Mathematical Biology*, 68(3):679–702, 2006.

[32] C. Molina and D. J. D. Earn. Game theory of pre-emptive vaccination before bioterrorism or accidential release of smallpox. *Journal of the Royal Society Interface*, 12(107):20141387, 2015.

[33] M. Newman, A.-L. Barabasi, and D. J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, NJ, 2006.

[34] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, New York, 2010.

[35] J. C Nunes and A. Ordanini. I like the way it sounds: The influence of instrumentation on a pop song's place in the charts. *Musicae Scientiae*, 18(4):392–409, 2014.

[36] F. Pachet and P. Roy. Hit Song Science Is Not Yet A Science. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 355–360, 2008.

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[38] M. Schedl, T. Pohle, N. Koenigstein, and P. Knees. What's Hot? Examining Country-specific Artist Popularity. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 117–122, 2010.

[39] D. Schenzle. An age-structured model of pre- and post-vaccination measles transmission. *IMA Journal of Mathematics Applied in Medicine and Biology*, 1:169–191, 1984.

[40] V. Tweedle and R. Smith? A mathematical model of Bieber Fever: The most infectious disease of our time? In S. Mushayabasa and C.B. Bhunu, editors, *Understanding the Dynamics of Emerging and Re-Emerging Infectious Diseases Using Mathematical Models*, pages 157–177. Transworld Research Network, Kerala, India, 2012.

[41] P. van den Driessche and J. Watmough. Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(Sp. Iss.):29–48, 2002.

[42] P. Weinberg, J. Groff, A. Oppel, and A. Davenport. *SQL, the Complete Reference*. McGraw-Hill, New York, NY, 2010.

[43] M. Woolhouse, J. Renwick, and D. Tidhar. Every Track You Take: Analysing the Dynamics of Song and Genre Reception Through Music Downloading. *Digital Studies/Le champ numérique*, 4, 2014.