Information Visualization miniproject

Severi Rissanen

Introduction

I chose a data set with information about the concerts of Helsinki Philharmonic Orchestra since 1882. It contains the information about each composition, the concert that they were played in, its date, location, composer and many others. I explored a few ideas of what to do with the data, and ended up with two main angles: First, I tried to visualize the amount of work that the orchestra has produced, and how has it changed over time. Second, I wanted to also show the diversity of composers that the orchestra has represented in their playing. The idea was that maybe they could be good metrics of how well the orchestra is doing artistically, since having the chance to play lots of different kinds of music and thus represent different kinds of art seems like something that a classical music orchestra should aim for.

Amount of concerts and in-concert diversity

The data

Figure 1 shows the first two graphs, which visualize the amount of work that the orchestra has produced over the decades as well as the diversity of composers inside individual concerts. The plot on the left shows that the amount of concerts got off to a slow start in the 1880s (although the count isn't completely fair since the first concert was towards the end of 1882) and quickly jumped up at the beginning of the 20th century. 1920s to 1970s saw a clear decline in the amount of concerts, however, reaching its lowest value in the 1960s. After this, the amount of concerts rose back again.

I wasn't entirely satisfied with this from the point of view of the amount of work produced, since it could have been the case that the orchestra simply changed the way they work by doing more large concerts with lots of compositions in 1920-1980. The plot on the right shows that this is not the case, and instead the amount of compositions per concert dropped down as well. Noticeably, it only rose back to the levels it was during the early 20th century in the decade starting in 2010. Now one could ask further: Were the compositions played during different periods of different lengths as well on average? The data doesn't answer this question, and in the absence of that information, it seems reasonable to conclude that the amount of work done by the orchestra dropped down quite much from the 1920s, and recovered only quite a short time ago. This may have been due to a lack of funding, maybe in part because of the civil war in 1918 as well as the second world war in the 1940s in Finland.

I also wanted to take the chance to visualize the average amount of composers represented per concert throughout the decades in the second plot, as the ratio of this with the compositions per concert gives an interesting metric of *in-concert diversity*. If the amount of different composers was the same as the amount compositions, the diversity would be at maximum since no composer was repeated twice in a concert. In contrast to the total amount of work, this shows a clear decreasing trend from the 1880s until today, which means that the nature of concerts has slowly shifted in the direction that multiple compositions by the same composer are played in one concert. Note that this is not a measure of *global diversity*, since a high percentage score could still mean that the same few composers are repeated in every concert.

Comments about the visualization

One thing to note about all of the plots in general is that I've tried to maximize the data-ink ratio by removing the boxes around the graphs and around the legend in the second graph, for instance. I also tried to keep a consistent visual style by not changing the font sizes or the style of the x-axis ticks.

The plot on the left is fairly straightforward, and the main issue I had with this graph in special was the fact that the orchestra didn't start playing until 1882 and the data for the last decade is incomplete. I didn't want to fool the viewer with the visual expectation of equal size intervals, and decided to add short descriptions about this in the graph. I also decided to use a line plot because the absolute values

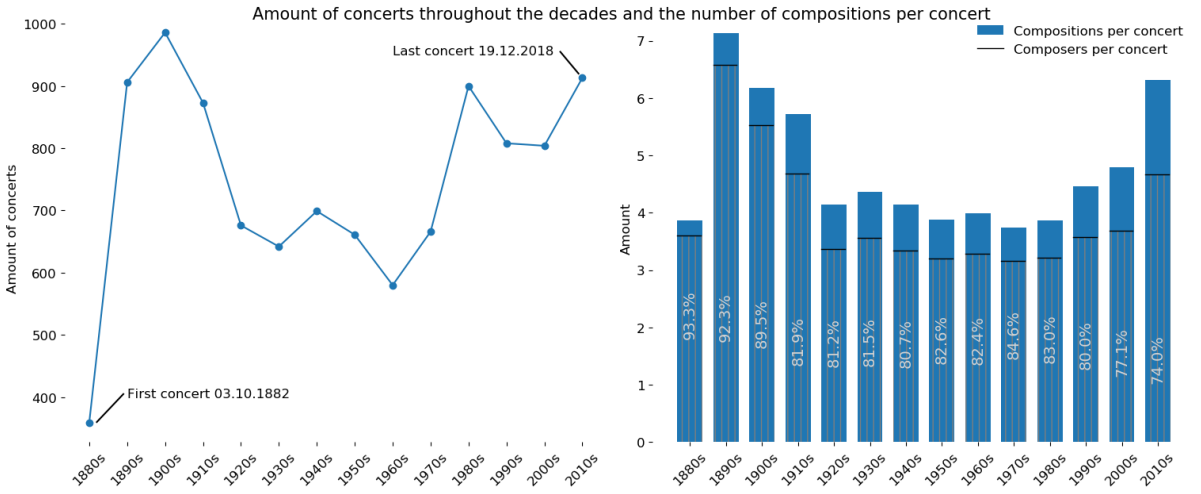

Figure 1: Left: The absolute amount of concerts that the orchestra held for each decade. Right: The average amount of compositions per concert and the average amount of composers represented per concert. The percentages show the ratio of composers and compositions per concert: 100% would mean that a composer was never repeated twice during a concert.

of the amount of concerts aren't necessarily the most important thing here, and instead the trend and changes are the interesting part.

For the second graph, I decided on using a bar plot, because the actual values are interesting (it's the number of compositions you hear when you go to a concert). The main feature of compositions per concert in the bar heights is quite straightforward, but I also wanted to add information about the composers per concert. I stumbled in to the problem that the actual value is only really interesting when compared to the compositions per concert, and decided to highlight that by presenting it as a fraction of the bar height. I decided to use distinct black lines at the top and colored lines in the middles of the bars, because a different color entirely could have lead the viewer to believe that this is a stacked bar plot. I was worried about Moirè effects and chartjunk because of the lines, and I tried to account for these by making the lines not pop out too much by using a color which has a low color contrast with the underlying blue as well as a similar luminance.

The next issue was that the heights of the bars change, and thus it is difficult to see how does the ratio change over time. First I decided to add the percentages with a light grey colour to be visually distinctive enough and easy to read but also to not be quite the first thing that attention of the viewer is directed to. To let people associate the numbers with the striped area, I turned the stripes to be oriented vertically along with the text. The idea actually came from the Gestalt law of good continuation (or maybe similarity would apply here, I'm not sure) so that vertically oriented text can be associated with vertically oriented lines, and I think that the result is much better compared to other line orientations. The final feature of the plot is the position of the text along the y-axis. I figured that it's still difficult to see the trend of the ratio by just looking at each number (which are sideways), so I made the position depend on the ratio as well. I didn't add an extra explanation about this in the plot in the hopes that it's obvious that the height corresponds to the numbers on some linear scale. This feature is slightly experimental, but I think that it works quite well: The general trend of the ratio can be seen pre-attentively quite easily, and if the viewer is interested in some particular values, they are quite easy to read as well.

Also, while one could say that adding the data about the ratios is redundant data-ink, I think that it would be quite difficult to see it properly with just the "composers per concert" absolute values.

Global composer diversity

The data

Figure 2 shows the second set of two graphs, visualizing the distribution of different composers represented throughout the decades. The plot on the left shows the proportion of total compositions played by the five most popular composers (in the entire data set). We see that in the 1890s, Wagner was quite popular, but none of the composers had a very large proportion of all compositions. This started changing as time

went on, with Jean Sibelius accounting for almost 20% of all compositions played in the 1960s and the top 5 accounting for over 30%. This changed dramatically in the 1970s and after 2010 the proportion of compositions by the top 5 was less than 15%.

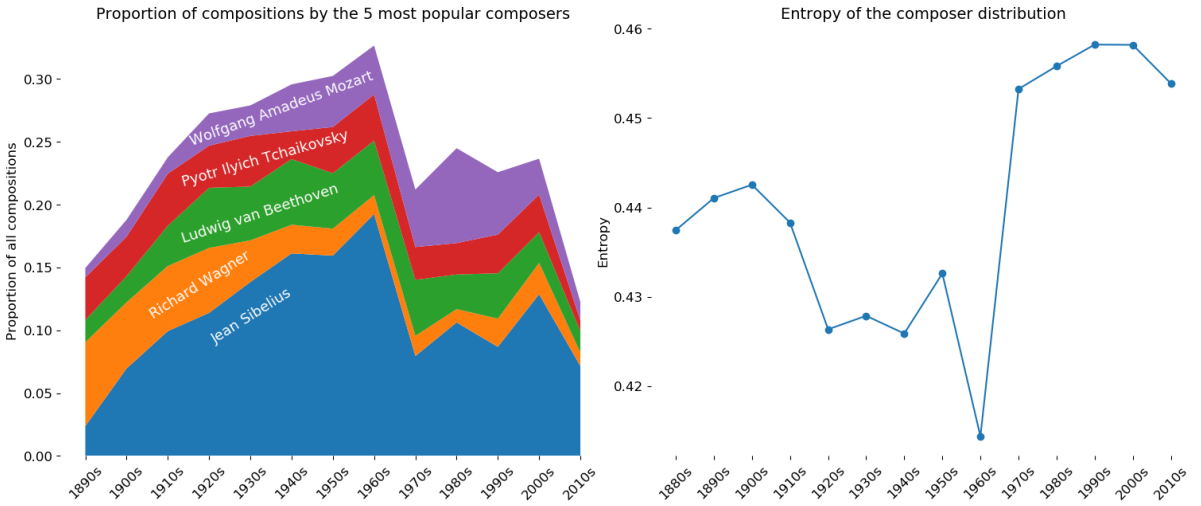

Figure 2: The distribution of the composers represented in the compositions played by the orchestra over time. The plot on the left shows the proportions of the composers with the largest total amount of played compositions in the entire data set (with Jean Sibelius having the most and Mozart having the least here), while the plot on the right shows the entropy of the distribution over time.

The prevalence of compositions by the top 5 does give an intuitive view of how the diversity evolved over time, but since it doesn't account for all of the other composers, I decided to take a more principled approach. For each decade, we can define the distribution of compositions by composers by $p(\text{composer}) = \frac{\#\text{compositions by composer}}{\#\text{total compositions}}$. The entropy,

$$H = -\sum_{\text{composer}} p(\text{composer}) \ln p(\text{composer}), \tag{1}$$

is then a measure of spread of the distribution and thus the diversity of different composers represented. The plot on the right shows the evolution of entropy over time, and it does confirm the general trend seen in the first plot: Diversity started decreasing after the 1910s, reached a peak low in the 1960s and improved very fast starting in the 1970s.

The two graphs show that in addition to problems with the amount of work being produced with the orchestra, the diversity suffered as well from the 1920s to the 1960s, with the worst decade being in the 1960s. The large amount of compositions by Sibelius does of course make sense as Finland was still creating its national identity after the second world war, and the orchestra was possibly affected by this by one way or another. It's also possible that in times of budget constraints, it's a good idea to resort to known favorites. It's notable that even before the rise in the amount of concerts in the 1980s, the diversity increased dramatically already in the 1970s, possibly reflecting that the orchestra wanted to move away from just performing works by Sibelius even before they got the chance to increase the actual amount of work being represented.

Comments about the visualization

The plot on the right is again a simple line plot, because something like a bar plot wouldn't make much sense here since the absolute values of entropy aren't interesting here. The left graph has some more interesting features, notably the absence of a legend and instead labeling the different composers directly in the colored areas in the stacked plot. This improves the data-ink ratio, but also the white color of the names gives a clear contrast with the underlying colored regions, and at least their locations are quite pre-attentive. It's very easy to see where to look if you want to find out what does the green color mean, for instance. I also tried to avoid visual stress by having a limited number of top composers presented in the plot, as it gets quite messy quickly. Jean Sibelius being represented with the color blue also should be intuitive to most Finnish viewers because of the shared cultural knowledge. This visualization might

not convey all of the information when printed black and white, but maybe that could be changed as well by having borders between the colored regions. On a computer, I think that not having a border looks somewhat better.