How did the mobile device market change?

Hengrui Liang, Victor Z. Nygaard, Odbayar Bold, Zecheng Li
COMP5048 Visual Analytics Assignment 2, 2022 S2
The University of Sydney,
Sydney, Australia.

Introduction to the project and data

This report constitutes the primary component of the University of Sydney COMP5048 assignment 2 submission for 2022. Throughout the project we will be exploring an amputed dataset of various specifications for mobile devices that have been released within the years $[1989, 2013]$, 1990 exclusive. Throughout the course of the project, several processing and visualisation technologies have been utilised including those of Python, and Tableau.

The year column of data is not integer but a float hence cannot be used directly. So a conversion has been made here. There are also other data processing and charts production codes implemented here:
The data sets given to us are not complete and many pre-processing steps can be made. Thus we did some data processing in python code such as change `Release Year` column to integers, add some extra columns to database which are `5years`, `Decades`, `CatTF`, and `Models`. Also we did data processing for different graphs that we would use like making a new data frame in pandas to produce nice plots in plotly package in python.

A - Discovery of types

For visual discernment (classification/assessment) of the different type of mobile phones that have been introduced throughout $1989 - 2013$, we may conduct exploratory data analysis. In aid of this, we may evaluate the relationship between the different features of the dataset in pairplots, made by `ggpairs`, as has been done in Figure 1, Figure 2, as well as is allowed via a interactive exploratory data visualisation tool, that has been made based on templated work. An introduction to (some (having since added additional)) of the app's functionality can be found at this link

Fig. 1: `ggpairs` pairplot of half of the available numeric features over pre-specified time periodes.

Fig. 2: `ggpairs` pairplot of other half of the numeric features, coloured by time periodes.

The splitting of the variables into two pairs of six has been done, as it would otherwise be additionally difficult to spot relationships amongst the features. We have split the release years of the devices into the periodes $[1989, 1996]$, $(1996, 2003]$, $(2003, 2007]$, $(2007, 2010]$, $(2010, 2013]$, in order to map time to colour without convoluting the visualisation unnecessarily. The mapping of the groupings to colour as a visual variable was the immediate choice made, given its ability for easy selectivity $\neq$ of multiple categories. Given the amount of information shown in each of the scatterplots, further work could be made to increase legibility. One may immediately spot a number of properties of the data based on Figure 1, Figure 2

1) The data is heavily skewed towards newer devices, as can be spotted by the `periode`-barplots in the lower righthand corner of the above figures. The point plots are thus often dominated by the data of the newer devices.
2) As expected, we see the general prevalence of a positive trend in the relationships of features. Growth at times looks linear, at times squarerooted, and sometimes even exponential.
3) The boxplots (supplied on the righthand side of the figures) at times reveal a large spread in the variables, and at times a high concentration, but with a seemingly significant number of outliers.
4) A (as could be expected) high degree of correlation between many of the features.
5) The pixel length as `pxLength`, `RAM`, pixel density as `pxDens`, CPU Clock as `CPU` along with the diagonal size of the screen as `inDispDiag` seem to be multimodal for several of the periodes.

One can also spot specific trends, that could allude to the type of mobile devices. The most obvious of these based on Figure 1 and Figure 2 would be the two-forked split in the relationship between `pxLength` and `pxWidth`. This split would suggest prevalence of two different paradigms in terms of devices. As resolution of a device, by definition, is `pxWidth`$\times$ `pxHeight`, the ends of the forks would be the most high-resolution devices. In terms of mobile devices, we thusly suspect that these devices could be high-end tablets or smartphones with devices along the lower fork would be those who are wider than they are tall, with the upper fork being comparatively taller tablets. Utilising we find do find this to approximately the case with the iPad4 with resolution 1536×2048 occupying the tip of the higher fork, and the Google Nexus 10 of resolution 2560×1600, the lower fork.

The question of uncovering the precise nature of the different types of mobile devices however still eludes. To get closer, one might attempt to carve up the feature space by employing decision tree -style linear decision boundries to categorise types. This however quickly becomes infeasible by the paradox of choice, for all but a severe restriction of the number of features involved.

Looking through the `Model` column of the data, we discover device types such as *smartphones*, *tablets*, *personal digital assistants*, *commercial GNSS devices* and *Nokia-style* phones. While one may quantify common and differing properties of and between such devices, converting the properties to absolute delineations of the feature space occurs several issues. Cheifly amongst these, interpretations of what combination of properties might constitute a PDA, Nokia-style phone, tablet or smartphone need not exist in isolation. These are subject to change over time, and will also depend on the other devices available on the marked.

In order to aid opportunities for further visual discernment of mobile device types, we introduce $2^4 = 16$ semi-arbitrary mobile device characteristics based on distributional properties of the dataset, thus letting the data *speak for itself*, and avoiding having to motivate choices for drawing particular decision boundries oneself. These distributional properties are restricted to only concern the `CPU`, `inDiagDisp`, `Mass` and `pxDens` features. In particular, for each year, we split the data into one of the 16 categories, which we will henceforth refer to either as the *category or type* $1-16$ *(C*$1-16$*)* or as *the* `CatTF`, based on whether it is true (`T`) or (`F`) false that the particular device is recorded as being $\geq$ the yearly median for each of the four features above. Making such bisectional cuts for the four features yields the 2^4 different categories. In addition to being recorded as $C1-16$, the categories are also recorded on the format $*_{j_1} *_{j_2} *_{j_3} *_{j_4}$-format with $*_{j_i} \in \{\mathtt{T}, \mathtt{F}\}$, $j_i = \mathtt{CPU}, \mathtt{inDiagDisp}, \mathtt{Mass}, \mathtt{pxDens}$, and counted up binarily. Thus, for example, devices that fall into *C14/TTFT* of a particular release-year will have higher $(\geq)$ than median CPU clock and diagonal display size, lower $(<)$ than median mass, and higher $(\geq)$ than median display density than the other devices with the same release year. We might thus imagine that typical entries that fall into *TTFT* would be large, high performance, slim (so as to keep the mass under the median for the year) devices with high-density screens, such as, for example, flagship phones.

The choice of the above mentioned four features for creating the `CatTF` categories was based on several considerations. The secondary reason was statistical power considerations based on the SVD of the features, the aforementioned multimodality and limiting the selection of the most highly correlated features. Chiefly however, the choice was made based on intuitive interpretability and discernability of different categories based on these features, with a mix of *performance*, *size* and *aesthetic* features, respectively, included. Consequently, we might then start to unfurl the different types of devices based intuitive delineations amongst the device types.

In also showcasing the interactive features of the app, that allow us to discover device types, we might take discovery of GNSS type mobile devices as an example. We might focus on these types of devices for all data within the years $2007-2009$, and thus apply the interactive sliders for subsetting number of data points and the range slider for years to select these data. In addition, we will colour the data based on `CatTF`, and thus make this selection. By general knowledge of the purpose-built nature of GNSS devices, we might expect them not to require too much compute power. Given that many such devices are to be mounted in automobiles, we might expect them to have large screens and be less sensitive to weight-considerations. And finally, they will often not have the most advanced display.

We might thus suspect *C7/FTTF* to be a fitting categorisation for such devices. We thus doubleclick the *FTTF* legend to isolate these points in the scatterplot. Given additional domain knowledge on, for example, comparatively low RAM and high depth of such devices, we might map one axis to depth, and the other to RAM by clicking through the dropdown menu. Finally, noting the amputed, degenerate and concentrated nature of the data, we might finally press the "jitter" button to perturb datapoints from lying on top of each other, and adjust the degree to which points are perturbed using the horisontal and vertical jitter sliders, as well as log transform the horisontal axis by pressing this button. The result is shown in Figure 3 and showcases that the user, subsequently to such interactivity choices, may easily find many GNSS navigational devices.

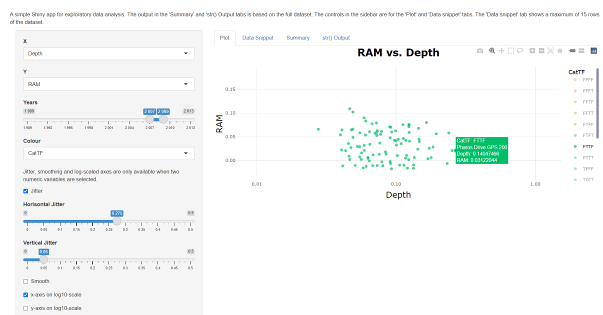

Fig. 3: Finding navigational devices amongst the $2007-2009$ data, by colouring based on `CatTF` and highlighting only *FTTF*. For example, here as hovered over, the *Pharos Drive GPS 250*. Also showcases design of app.

We might expect GNSS devices to be found amongst the low RAM, high depth segment. A manual sampling of entries however indicate that even amongst all entries of *FTTF*, more than 80% could be navigational devices.

This procedure of using `CatTF` to discover the *true* device types may be repeated for the other types of devices as well.

Finally, we may note that many of the design decisions of the interactive app, have been left unchanged as to their original template design, and could be subject for further improvement of the app. In particular, we might mention the horisontal real-estate of the sidebar panel, that might be better implemented if moved above or

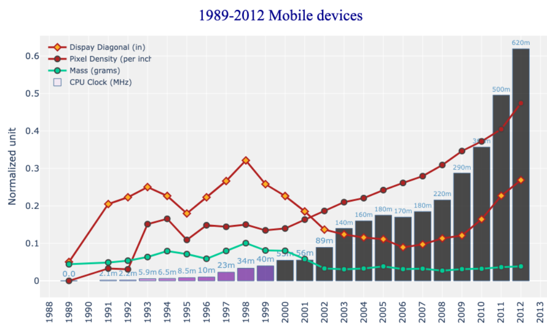

Fig. 4: Overall Features Trend

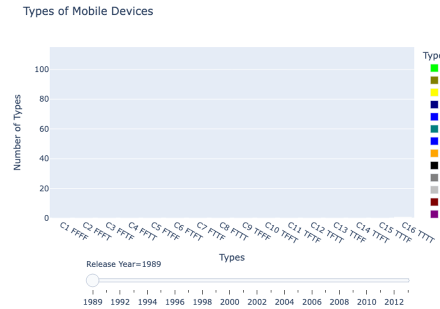

Fig. 5: Distribution of types introduced in 1989 (Recording only a single datapoint that falls into C16 (TTTT))

below the main plotting window. This plotting window, might also warrant an increase in screen-real estate to fit with its status as the magnus opus of the app.

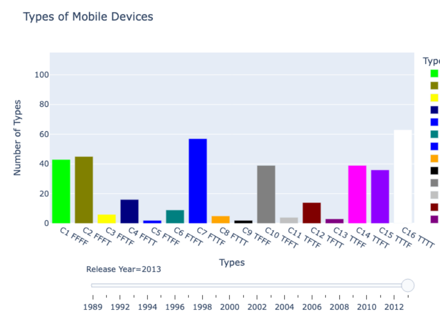

Fig. 6: Distribution of types introduced in 2013.

B - Type Trends

We have constructed five graphs to visualize the trend of the data sets. First graph in Question B shows trend of features over years. X-axis, y-axis and z-axis are designed perfectly for different reasons which x-axis is the Year which users can see how features changed over certain year, y-axis is the actual average data for features in certain years that help users to see the overall trend, and z-axis represents colours for different features which helps users to figure out data sets of different features in one graph.

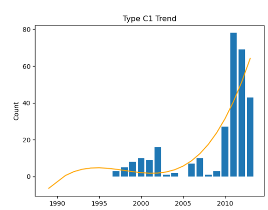

Fig. 7: Type 1 (FFFF) mobile devices.

The reason why average value of data sets for these features were used in Figure 4 is that it can helps users check exactly changes of features for the data sets in certain years which also users can benefit from it if they have questions about other plots that shown in Question B.

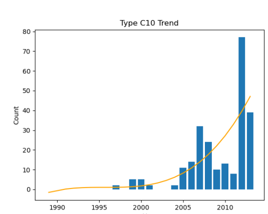

Fig. 8: Type 10 (TFFT) mobile devices.

We recall the true/false device type naming scheme presented in the previous section.

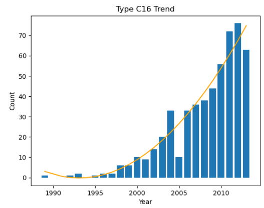

Fig. 9: Type 16 (TTTT) mobile devices.

Then we have another two graphs which we cannot show the graph completely due to the fact that this is an interactive graph. These two graph(Figure 5, Figure 6) can clearly show the distribution of each types of mobile devices. We will discuss the trend of types between 1989 and 2013 by five years due to the fact that if we analyze trend year by year that would be too complicated and inaccurate. The reason why we define x-axis as Type names of mobile devices, y-axis as number of different types of mobile devices, and z-axis(colours) as also different types of mobile devices are: 1. x-axis can ensure that every types of mobile devices would appear in every year in our data sets so users can compare or create any relationships among types of mobile devices. 2. y-axis can make it clear that how many each different types of mobile devices in each year so that we can see the trend by the length of each "bar" of each type of mobile device. 3. z-axis is used for making different types more obvious to observe so users can quickly find what kind of type they want to look or compare with.

And the reason why distribution of each types in each year is used here is simply because it can easily reflect what kinds of "feature trend" are popular for different years. And it also definitely benefit us to see the "trend of types" for mobile devices.

The interaction that used here is users can drag the line in the bottom of the graph which can change the graph by years. For instance, if user drag the line to year 2000, then data sets of 2000 would be presented in the graph. And it significantly helps with users to check the number of each type in different years or specific years.

(Users can watch our introduction video then check interactive graph for years that will be discussed following which we will talk about the trend we observed from our graph by roughly 5 years. Also we will select three different types of mobile devices to see their "own trend" through these five years. Overall, we will find four different trend: one for general trend of types, and three for trends of different types in each period)

From 1989 to 1996, this is the beginning of mobile devices evolution which only few types of mobile devices were published in the public, and most of features of mobile devices are not mature, people are exploring which kind of mobile devices is the best. For instance, we have one data point for Type 16 only in 1989, and two for Type 16, 9, one for Type 6, 8, 12 in 1996. There is no obvious trend in this five year period. If we can see graphs Figure 7, Figure 8, Figure 9, we can see that at the beginning they all at 0, and only Type 10 and 16 grew a little bit.

From 1996 to 2003, we can see that people were trying to develop mobile phones in every features(features that we chose in Question A) that it contains. It can be seen clearly that Type 16 grew rapidly, so do Type 14 and Type 15 which these two types are similar with Type 16 which only one of the features were not focused to develop. And we also have Type 2 which reject development many features and remain only "Pixel Density" feature to develop for mobile devices. If we see graphs Figure 6, Figure 7, and Figure 8, they all grew a little bit compare to the number of them in 1996 which means mobile device's producers are still exploring which combination of features is the best.

From 2003 to 2007, generally all types of mobile devices grew quickly these years especially for Type 2, 10, and 15. And Type 10 has the most significant growth where it got no data point in 2003 but its number of products increased to 32. Also Type 16 increased a little bit and stays high number of its type. Type 10 increased most likely because of mobile devices like Nokia N95-1 published with high CPU, low Display Diagonal, low Mass, and high Pixel Density(). And Type 15 is mostly link to the publication of first impression of Apple iPhone, also Type 2 is likely to be some kind of mobile devices with high resolutions published in 2007. Then if we look at Figure 6, Figure 7, and Figure 8, we can see that Type 1 is mostly disappeared which is obsolete in these years. However, Type 10 and 16 increased significantly even for Type 16 there is a drop in 2005, it increased again by year 2007.

From 2007 to 2010, generally speaking, Type 2, Type 10, and Type 15 were declining relatively. However, all other types of of mobile devices grew a bit which more types of mobile devices are produced such as: HTC (t-mobile) dream g1, iphone 3g, balckberry curve 8520. They all contains different combination of features, for example, blackberry is with High CPU, High Display Diagonal, Low Mass, and High Pixel Density. Moreover, Type 2, 10's declines can prove that in this period, people like mobile devices to be big with larger scales. If we check Figure 6 and Figure 8, both Type 1 and Type 16 are rapidly increased in this period even though Type 1 got a drop in year 2008. Type 10 expresses a completely different trend which it decreased, most likely in this period people like mobile devices to be large scale with bit higher weight.

From 2010 to 2013, Type 2, 7, 10 increased significantly which represents that on the one hand, people focus on only Pixel Density or combination of low CPU, high Display Diagonal, high Mass, low Pixel Density or combination of high CPU, low Display Diagonal, low Mass, high Pixel Density. Some typical mobile devices can be samsung galaxy note n7000 which belongs Type 10 in this case. The trend in this period is people more focused on high resolutions mobile devices can provide for them thus, Types like Type 10 with high Pixel Density increased very fast. Also if we can see Type 5 and 9 they dropped significantly which also proves that mobile devices with high Pixel Density is important for this period. Then if we look at Figure 6, Figure 7, and Figure 8, the trends for all of them are increasing rapidly. In this period all of these three types encountered a crush between 2012 and 2013, but the overall trend is still increasing as they all get popular in this period.

C - Anticipate Chart and Justification

Based on the existing device data from 1989 to 2010, We may compare Figure 5 with Figure 10

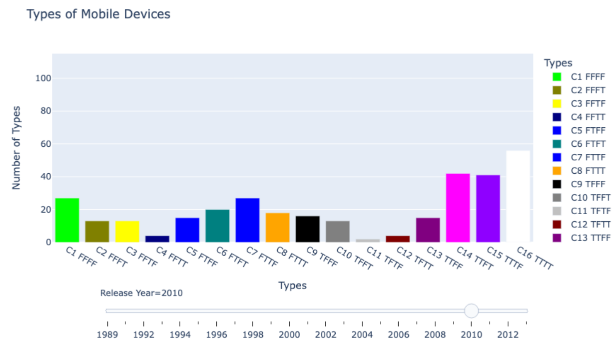

Fig. 10: 2010 type distributions

As the top three growth show at the most right part of the chart. C14, c15, C16 are three types show most growth overall. All of them featured 'TT' as the start of code. This means the high CPU and high Display diagonal devices are leading the market. Given CPU growth is highly correlated with RAM and Storage, we predict the new emerging device like the type C14, C15, C16 will be dominating the market. To summarize, the featuring attributes are as below:

- **High CPU, RAM, Storage** The RAM, CPU, Storage kept growing rapidly and made the phone smart. This could be double verified from the other chart.
- **High display diagonal** Phone screen pixels grew in width and length all the way. This made the phone screen bigger for richer content.
- **Low Mass** The mass of phone grew lighter each year.
- **High Pixel Density** Phone screen has grown higher density, better display fit of eyes.

CPU Clock Rocketing

There existed a gentle growth of CPU Clock from 1989 to 2005. However the growth became rocket rising from 2005 to 2010. The charts convinced us that there exists a strong force boosting up such rapid growth. We believe this trend will continue.

The RAM had shown an explosion ever since 2008, and the speed has accelerated all the way to 2010. We have good reason to believe such trend will continue in the future coming devices.

Bigger RAM will be introduced into newer devices consistently.
The Storage had been growing gradually until 2008. It grew faster and accelerated ever since then. Although not as not as CPU, but better CPU will demand bigger RAM to cooperate on for a better build device.

Diagonal

This display diagonal demonstrates the screen size. This is a good indicator of rich content are coming to phone.

Pixel Density Jump

The pixel density doesn't show much growth until 2008, and jump high ever since then. As better computing power will demand better display, we have reason to believe such trend momentum will continue in a long run.

Anticipate

Given above analysis, if we were to predict a newly emerging mobile device type, the faster computing(CPU, RAM, Storage) with big screen, and good display pixels yet still light device will be dominant during from 2011 to 2013.

Visualization justification

To see the 17 attributes trend in an easy way, we merged the highly correlated attributes like CPU RAM and Storage. We also selected the least correlated attributes to produce these four as a combination.
Once combined, we create type tag of the phone as the final visual variable on X and the count as Y to demonstrate the total quantity. This axes arrangement helps us to see a well delegated phone types development history data.
Further through a twenty years history data charts compare, we could easily observe the refined data trend.

I. Evaluation

First approach

We came up with two concepts to illustrate how common traits and current trends define different types of mobile devices, but each concept has advantages and disadvantages. We can recognize the strategies we build by their intended use. As

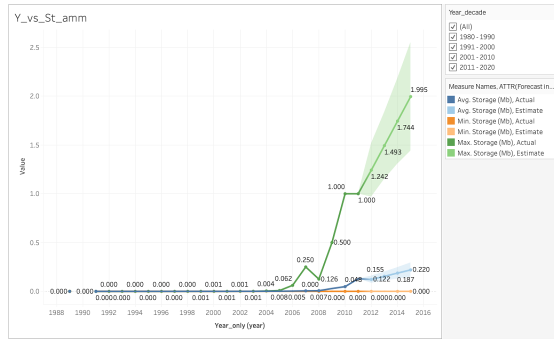

Fig. 11: Showing the Storage performances by Max, Min, Average with forecast.

an illustration, the first idea was to divide and categorize all devices into three groups: basic, pre-smart, and smart based on common characteristics, whereas the second idea was to identify more device types based on their individual characteristics. We used a variety of programmes and languages, including Python, R (shiny app), Tableau, and Microsoft Excel, to complete the task.
The dataset consists of 14 features, of which 8 are physical measurements and the remaining 10 are years and some very important specifications. Except for released years, dates, and model IDs, all of the data in the dataset were scaled and have values between 0 and 1. Values for the feature "Released year" in the dataset were a problem. The year has an extra digit after the comma, for instance, the Hewlett-Packard 95 LX (HP Yaguar) was released in 1991.2500. As a result, we developed a new feature that only includes years. We also included a feature called decades.
To identify mobile device types we illustrated basic graphs on some of the important features and understood that illustration.
Figure 11 shows three measurements that show how Storage performance has changed over the years. Years are represented on the X axis, and storage performance is shown on the Y axis. In this case, we used maximum, average, and minimum measurements and displayed mark labels to clearly see changes. In order to forecast the storage's performance over the next three years, we also enabled a forecasting tool. Since 2005, storage performance has improved while brands

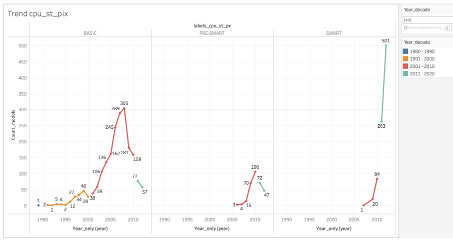

Fig. 12: Illustration of trends of 3 categories .

have continued to produce devices with very little storage. It also helps us recognize the three categories that we set out to achieve.

In the same manner, as described above, we used Tableau to illustrate all features so that we could choose which and how many features we needed to divide models into groups. But we also used R to implement a statistical technique that establishes a correlation between all features. Following the statistical analysis, we decided to stay with the three features of CPU, Storage, and Pixel density.

Despite the fact that there are three features, each chosen feature requires two scores. In essence, these points assist in dividing the dataset into three groups. Models below the first score (condition) tend to be more basic, and models above the second score tend to be smarter. Additionally, pre-smart mobile devices are those with a score between two measuring scores. As a result, there are three basic groups, three pre-smart groups, and three groups for smart devices. We overlapped groups according to their score points to create a final group. This approach is based on the idea that a smart device cannot simultaneously have a very low storage capacity and a high CPU score. According to the concept, it might be a pre-smart device.

The X-axis in figure 12 is divided into categories and represents years. The figure allows us to see the production trend of mobile devices because we assigned the number of models to the Y-axis. color can be used as a visual variable to show trends over time, and mark labels can be used to determine the precise size of particular dots. In order to clearly see trends by each decade, we also added an interactive filter tab to the visualization.

Result	Q1	Q2	Q3	Q4	Q5
Average	8.42	8.18	8.72	8.25	7.6

Fig. 13: Survey 2 response

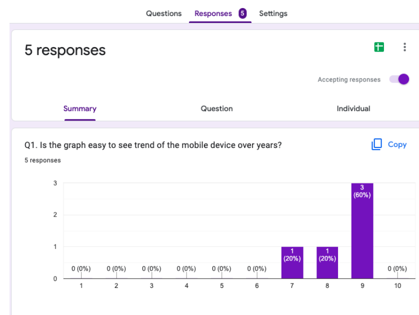

Fig. 14: Example of response

Although the method was effective and allowed us to see trends, we are still unable to distinguish between different device types. Each step in this method is mainly based on data visualization.

Survey 1

We conducted a survey and questioned the following groups: Groups CC-09-G03, CC-08-G03, CC-08-G01, and RE-09-G05. In the survey, it is asked how simple it is to see the trend of mobile devices from the graph. Basically, surveyors must respond to and grade questions ranging from 1 to 10. Our graph is difficult to understand if the response is 1, but it is simple to understand if the response is 10. The response, however, was poor, and the average score for each question was under five. We came to the conclusion that perhaps we need a different strategy that can reveal more about models.

Survey 2

To make sure our visualization is better and easier to understand, we conducted a second survey with 5 questions with the same participants. In addition, we gave surveyors access to a brief video that explains how to use and understand our interactive visualization. Unsurprisingly, the surveyors' responses progressively increased. This time, the survey's average score was much higher than five. As a result, we chose to stick with the current visualization.