Steam – Users and Games

Alexandra Figueiredo
83420
IST, Portugal
alexandra.figueiredo@tecnico.ulisboa.pt

Denis Voicu
83443
IST, Portugal
denisvoicu@tecnico.ulisboa.pt

ABSTRACT
The problem in our hands was to make a visualization of anything we wanted. Since we like games a lot and steam is the biggest digital platform we decided to do a visualization about Steams users and games. To do that, we looked online for data, treated the data to select only the relevant data for what we wanted to show and, in the end, we made the different idioms to show the data we treated.

This document describes how we made it work.

Author Keywords

Steam, users, games, visualization, data, d3, idioms.

INTRODUCTION
The chosen theme was about the players habits and the computer games in steam. This choice was motivated by our passion for computer games and by the fact that every member is a steam user. Besides that, it would be interesting to check whether the computers games industry is growing considering the new competitors (Smartphones, tablets, etc). And it would also be great to verify if there is some sort of connection between some social demographical data like obesity, depression, homicides and the people that have access to the internet and the number of users and the number of friends. (Basically, to confirm whether that old sentence proffered repeatedly by our parents is true: "You're sad/fat/more violent because you spent your whole day in front of the computer, playing your games"). So, we thought that the best way to verify that would be by collecting data from the largest digital distribution platform for PC gaming - the **Steam**.

We didn't find anything similar that could address our questions. We only encountered a few bar charts with the most played games and we found a web site that had tables with statistics about the games and the users. There was nothing where you could analyse and visualize the data. So, the goal of this project was to make a visualization where we could analyse different metrics of the games genres from 2003 to 2013. To better know the distribution of users around the world and their friends from 2008 to 2014 and to be able to compare all this with demographic metrics.

So, the tasks proposed were:

- Task 1 – **Explore** the distribution of users. More specifically the friendships around the world throughout the years.
- Task 2 – **Analyse** the evolution of the games genres throughout the years. This is to understand how the number of developers, number of publishers, the average price and the average rating changed in each game genre throughout the years.
- Task 3 – **Explore** if the number of players or users in a country is somehow related with the depression rate, obesity rate, homicide rate and percentage of the population using the internet.
- Task 4 – **Compare** all metrics of 2 different (or not) games genres in the same year or in different years.

With these tasks in mind, we can come out with the following questions:

- **Question 1:** Do the Portuguese users have friendships with the Russian users between 2009 and 2011?
- **Question 2:** Which is the game genre with the most expensive games throughout the years?
- **Question 3:** How does the number of users and the homicide rate varies between 2008 and 2013 for Italy? Is there any relation?
- **Question 4:** Was there more developers and publishers in the action genre in 2008 or 2009?
- **Question 5:** Are there more players in countries with higher depression rate?
- **Question 6:** The Spanish players have more friendships with the Portuguese players or with the USA players?
- **Question 7:** Which is the game genre with more developers? And in which year did that happened?

RELATED WORK
No work similar to the theme or visualization chosen for our project was found. So we based our project on multiple idioms found in various places. One of the sources were the slides given in the lectures. Other inspiring source was the videos available in the hall of fame that gave us some insight of what was expected from us and gave us some ideas, such as the heatmap. We also browsed the web with the hope of finding a less used idiom, but as informative as the others and that could help us represent some of our data. And we found the idiom represented in Figure 1, which represents the cyber-attacks that are taking place in real time.

Figure 1Cyber-attacks taking place in real time

We thought this idiom would be great to represent the number of friendships between countries.

Unfortunately, the only scientific paper found was this: https://steam.internet.byu.edu/oneill-condensing-steam.pdf. This scientific paper was found while searching for the data and it has some relevant information about the games and the users of the steam and their habits like for instance the number of games owned per genre, the evolution of steam friendships, etc. However, even though the paper was interesting, the represented visualizations were very simple. Therefore, it didn't add anything relevant to add to our project.

DATASET

Most of the data used is from a dataset collected to be analyzed for the 2016 ACM Internet Measurement Conference[1]. This data set has over 160 Gb of SQL inserts instructions. The description of this data set can be found in the website.

Since the data set was huge, we faced some troubles treating the data. First, we selected the tables that we thought that would be interesting to visualize just by looking at the description. And then we tried to select the instructions matching those tables. However, neither the text editors (like Sublime, Visual Studio Code, etc) neither the MySQL workbench could open such a big file, so we weren't able to collect the instructions. After some research we found a script that could split big files in minor ones. Nonetheless, the script was too slow and sometimes would stop running in the middle of the separation. So, we built a python script to do the same task and it worked better since it was faster.

Another challenge was to populate our tables with the SQL instructions(about 161.000 instructions).Initially we tried to run each file manually in MySQL Workbench, but soon we realized that was inefficient. So, we built a php script to do the insertions. Even though it took a lot of time to make all the insertions, it was more effective since the only work we had was to keep our computer on. After all that work we had to derive the data that really mattered from the original tables. The tables that we found interesting for our work were the following:

- **APP_ID_INFO:** which contained the selected information for each product offered on steam.
- **FRIENDS:** It contains a list of the friendships of steam users.
- **GAMES_DEVELOPERS:** It contains the names of the developers for each product on Steam.
- **GAMES_PUBLISHERS:** It contains the names of the publishers for each product on Steam.
- **GAMES_GENRES:** It contains the names of the genre for each product on Steam.
- **PLAYER_SUMMARIES:** it contains a profile summary for each Steam user.

From that tables we were looking to derive the following measures: the number of developers per game genre per year, the number of publishers per game genre per year, the average rating per game genre per year, the average price per game genre per year, the number of friendships between countries, the number of friendships per country and the number of users per country. With the purpose of doing that we had to do a few MySQL queries. Again, due to the excessive amount of data, we also had some issues running the queries, since sometimes the queries would abruptly stop running because it was taking too long. While doing those queries, we still noticed there were some missing data, so we searched for other data set that could complete the missing values and we found **Steam Spy**[2]. Steam Spy is a website that uses an API to estimate statistics about steam with reasonable accuracy. We downloaded the tables necessary to complete our data set. Imported the files in MySQL and treated the data using SQL queries.

Besides the measures already mentioned we also were looking for the number of hours played per game genre, the amount of money spent in each genre per year and the number of hours played per country per year. That data was interesting to better understand the behavior of the steam players. Unfortunately, despite heavy research we didn't find any of those measures.

As mentioned before we were also planning to relate the data about the users with social demographic data. Most of the social demographic data was collected from **Our World in Data**[3]. The data was also downloaded in the csv format and then transferred to MySQL.

In the end, after treating all data, we ended up with four tables that were extracted to the csv format.

When we were programming the visualizations, we had to do some small changes to our data set. One of the changes made was to add records with zeros to game genres where there were no recorded games. Other issue were the

[1] https://steam.internet.byu.edu/

[2] https://steamspy.com/

[3] https://ourworldindata.org/

countries. The countries were in ISO format and we thought that wasn't user friendly, so we add a column with the regular country name.

VISUALIZATION

Overall Description

The visualization was created to fit a screen without having to scroll up and down, so it could be easy to visualize what is happening and what is changing while exploring the visualization.

We use a dark color for the background of our visualization with the purpose of emphasizing the colors used for the idioms. The colors were carefully chosen to be harmonious between them.

Since we want to represent data from steam users and games, we divided our visualization in two parts. The two idioms on top is where you can see information about the users and the two on the bottom is where you can see information about the games. The information about the games is composed by a heatmap and a radar chart. The metrics seen in the heatmap can be chosen in the dropdown selection located to the right of the idiom. At first the radar chart doesn't show any information, because the data represented in the radar chart should be chosen from the heat map by clicking on a box in the heat map corresponding to the genre and the year pretended. The information about the users is composed by path edging bundling map and a scatter plot. You can choose the countries you want to see in the map and in the scatter plot by clicking on the circles in the map or by selecting the wanted countries in the multiselect dropdown menu situated above the map to the left. You can also choose the represented range of years by using the slide bar. The final representation can be seen in the figure bellow-Figure 2.

In the game section, the heat map represents different metrics for various games genres over several years. The different metrics can be chosen in the dropdown select menu on the left and the metrics available are the number of games, the number of publishers, the number of developers, the average rating and the average price. Each time a new metric is selected the heat map changes softly the color. You can hover each box to know the precise value of the metric chosen of each game genre in each year.

Radar chart

The radar chart represents all the metrics available on the heat map for a game genre and year. The game genre and year can be chosen by clicking in the heatmap on the box corresponding to the desired game genre and year. The user can select a maximum of two boxes. To unselect one of the boxes the user can select another box and the older one will be overridden or it can double click the box.

You can hover each polygon to highlight it from the other polygon if 2 boxes were selected. At the end of each vertex of the polygon is a circle which can also be hover to know the precise value of the metric represented by that edge.

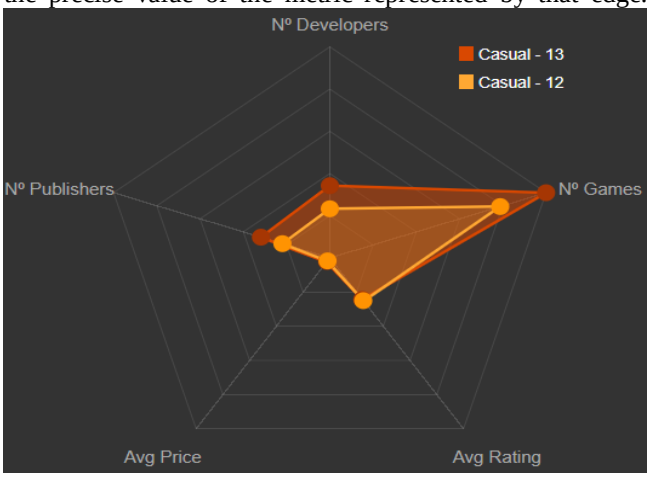

Figure 3. Radar chart with the boxes Casual, 2013 and Casual,

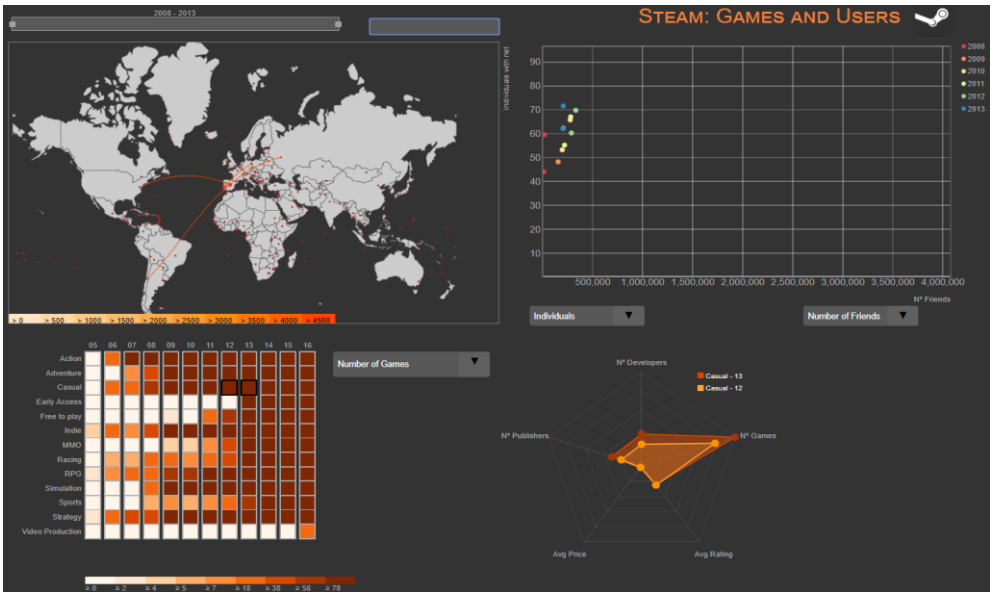

Figure 2.Final visualization with Portugal and Spains selected in the map. Casual, 2012 and Casual 2013 selected in the heat map.

2014 selected in the heat map.

Map

The map represents the friendships between countries. Each country is represented on the map by a circle positioned over the capital´s country location. Each circle can be hovered to know the exact country name referred by the circle. You can click on a circle which will make it grow in size, highlighting that it was selected. After selecting a country, it will appear, in most cases, lines to other circles representing that at least one user from the selected country has a friend in the country where the line ends. Each line has a color representing the total number of friendships. The amount corresponded by each color can be found in the color legend below the map. For a precise knowledge of the amount of friendships represented by a color, the user can hover the line and know the precise value and to which countries does the line belong to. Also, while hovering a line it will be highlighted so it would make it easier to see it on the map in case various countries are selected - Figure 6.

For better visualization and exploration of the map we also implemented a zoom for the map. Allowing to zoom in and zoom out at any point. Also, you can drag the map around to fit your needs.

You can find above the map a slider -Figure 4- representing the range of years selected that will influence the map and a dropdown selection -Figure 5- with all the countries. In the dropdown selection you can select as many countries as you want and also unselect since the dropdown and the map are synchronized. There is also an option in the dropdown to select all the countries. Also, the countries in the dropdown where ordered alphabetically to allow the user a much quicker search and it is easy to understand.

Figure 4. Slide Bar.

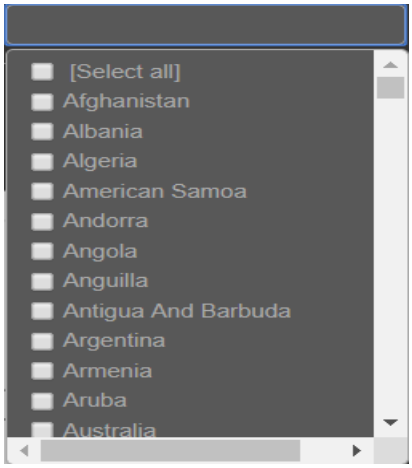

Figure 5. Dropdown multiselect.

Once selected a country in the map, it will appear on the scatter plot.

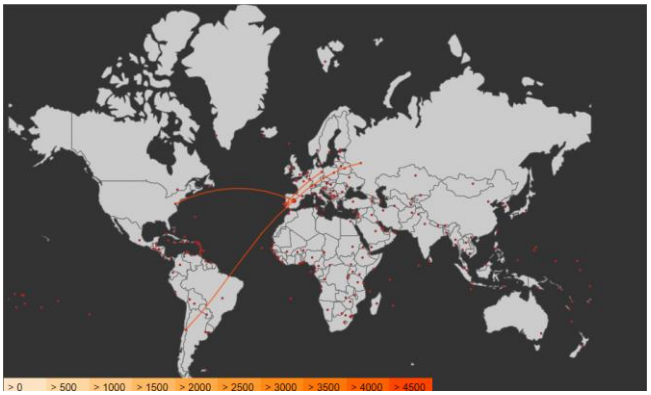

Figure 6. Map with Portugal and Spain selected.

Scatter Plot

The main goal of the scatter plot is to analyse and compare the selected countries in the map according to the metrics chosen for the y axis and x axis.

The y axis has two metrics: number of friends and number of users. By default, the y axis starts with number of friends selected. The dropdown selection for the y axis is located below the scatter plot to the left.

The x axis has four metrics: obesity, depression, homicide and individuals with access to internet. By default, the x axis starts with obesity. The dropdown selection for the x axis is located below the scatter plot to the right.

After changing any metric in the x axis or y axis, the scale in the axis will adjust smoothly and accordingly to the metric chosen.

It is possible to hover the dots to obtain a more precise information about the metrics and to witch country it belongs. It is also possible to zoom in and drag the dots to better fit your needs.

Figure 7. Scatter plot with Portugal and Spain selected in the map. Axis x refers to number of friends and axis y refers to individuals with internet.

Heat Map

The heat map represents how the games genres have changed throughout the years according to a certain metric selected in the dropdown selection on the left side of the idiom.

On a mouse over event on a box, you can obtain the concrete value of the metric. It is possible only to select 2 boxes at the same time and each box when selected is highlighted by a border that appears around the box.

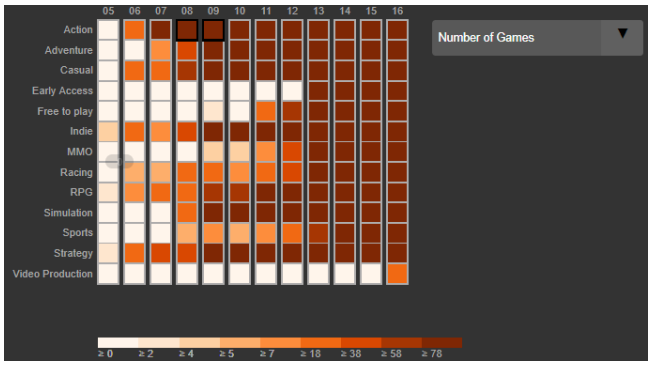

Figure 8. Heat map with two boxes selected.

Rationale

We used the visual encodings depicted in table 1, as we thought that, in conjunction would result on a clear visualization of the data.

Table 1. Visual Encoding for each item

Item	Visual Encoding
Year	Label, Position
Genre	Label, Color
Metrics	Label, Color, Position
Country	Color, Label, Position,

We choose a heat map because it was a good idiom to get an overview of the evolution of a certain metric throughout the years for the various games' genres. Allows you to quickly compare for example 2 different years for the same genre or 2 different genres for the same year. Although, if one desires to compare all the metrics of a specific year and game genre, we choose a radar chart.

The radar chart was chosen because it allows to compare all the metrics. You can easily compare all the metrics for a specific year since all of them are in front of you, so you don't have to change metrics in the heat map to see the value you are looking for. Basically, radar chart was very appealing because we have 5 metrics that we want to compare at the same time, and it is easier to compare 2 vectors based on shape than to remember values. Also, we needed something were the scale was different for each metric and radar charts allows us to do that.

We wanted to show the relations between countries based on steam friendships. It was obvious for us that we would choose a map to show that.

We also were interested in compare the number of users or the number of friends with some social demographic metrics. So, we are talking about 2Q variables where we want to express their value with horizontal and

vertical spatial position and were we want to compare, explore and find trends, outliners. So we picked a scatter plot.

The visualization did not change much from the initial sketches since the feedback was positive. We changed the final layout since we thought this way was better.

Initially, the radar chart was supposed to be a star glyph or a circular line/area graph but as soon as we saw the radar chart we knew it was the one.

Also, the scatter plot initially was supposed to be a multiple scatter plot were each scatter plot would represent a year, but we soon found out that would work. Not because it was wrong, but we were advised to use only one scatter plot and multiple would occupy too much space on our visualization.

We had another idiom that represented relations between developers and publisher. We wanted to put theirs name on the layout and when selected it would show the collaborations that name had, and the average rating of their games developed. Since, we had about 8000 entries it was impossible to put it all on the layout, so we decided to eliminate that since we didn't find an efficient way to represent the data. Also, our initial objective was to fit all on one screen without scroll, so that was also a factor to have in mind.

Demonstrate the Potential

To demonstrate the potential, we are going to answer to 2 questions from the Introduction section.

Question 1: Do the Portuguese users have friendships with the Russian users between 2009 and 2011?

To answer this question we will have to use the map. On the map we select Portugal or we can choose Portugal from the dropdown selection. Instantly it will appear the lines corresponding to the Portuguese friendships with other countries. Since we what to see only between 2009 and 2011, we have to change our time line on the slider.

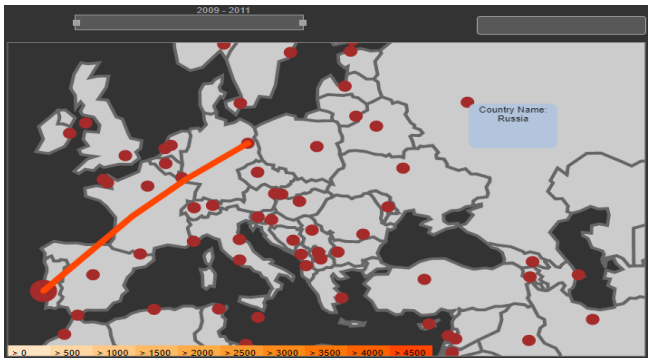

Figure 9. Portugal selected and there is no line to Russia. Time line is between 2009 and 2011.

Soon, we can conclude that the Portuguese users don't have any Russian friends between 2009 and 2011.

Question 4: Was there more developers and publishers in the action genre in 2008 or 2009?

To answer this question, we select on the heat map the box corresponding to the year 2008 and the games genre Action and the box corresponding to the year 2009 and the games genre Action.

Then, in the radar chart we can easily tell that for Action, 2009 there were more developers and publishers than for Action,2008.

Figure 10. Action,2008 and Action 2009 selected in the heat map. In the radar chart we can identify that there are more developers and publishers in 2009 than in 2008 just by looking at the position of the dots.

Question 6: The Spanish players have more friendships with the Portuguese players or with the USA players?

On the map we select Spain. Then we hover the lines corresponding to USA and Portugal, we can see in the tooltip that Spanish players have more friendships with USA than Portugal.

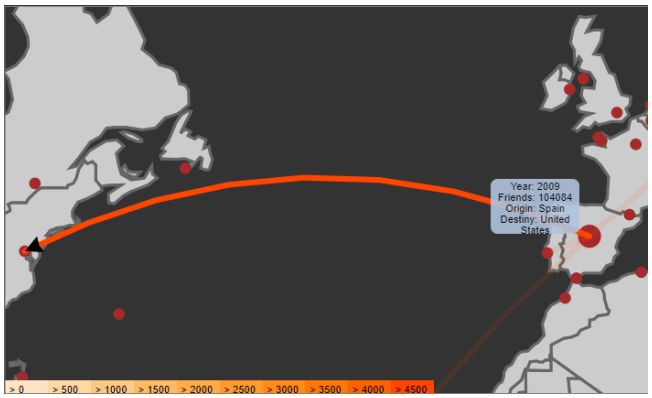

Figure 11. Hover on the line that connects Spain and USA to see the number of friends.

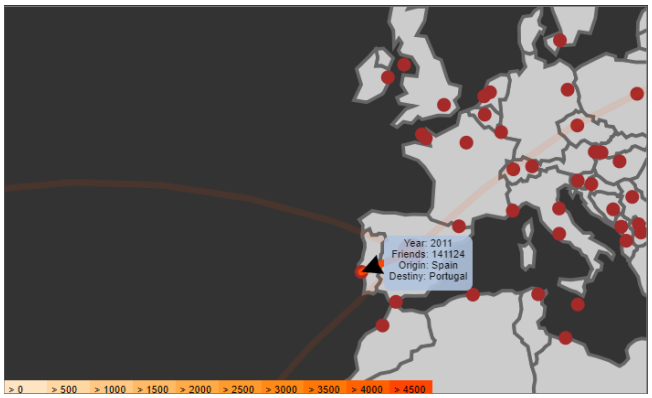

Figure 12.Hover on the line that connects Spain and Portugal to see the number of friends.

Implementation Details

To implement our visualization we used: d3.js, script for the slider since we didn't manage to do it from scratch and jQuery for the dropdown multiselect since it was impossible to do it efficiently only with CSS, HTML and D3. All the other idioms, heat map, radar chart, scatter plot and the map were made from scratch. Of course we went online to find inspiration, documentation and analyze some examples, but all of the idioms mentioned before even if they look similar they were all heavily adapted for our needs.

The interactions between the different idioms was all made based on d3 events and global variables

To help the user keep track of what he is doing and don't lose focus we used highlighting mechanisms. In all idioms we have listeners to detect mouse click or mouse over events.

In the heat map we highlight the boxes selected by creating a border around the box selected, something that is not don't in the other idioms since we use the opacity of the colors to simulate highlight. The reason for that is that the heat map was the first idiom that we implemented so by the time we discovered how to do a proper highlight we were already running out of time.

Conclusion

We learned a lot about data visualizations and what works and what doesn't. We learned how to use new tools and how to use them to achieve a desired state.

We were able to address all the tasks, and every question can be answered using our visualization.

If we were to start over we would be more careful regardless the dataset size, since laptops can not handle 160 GB of data.

Future Work

If we had more time to spend on the project, probably it would be very interesting to make a collaboration with Steam and have access to data like what games does each user has on steam, how much time did each user spend on different games, the amount of time spend playing games, etc… And with that data enrich our visualization by creating new idioms to make a more complete visualization.