Visualizing Steam Game Information and Purchasing Records

Xuan Huang, Haihan Lin

Email, u1209767@utah.edu u12063262@utah.edu

uID, u1209767 u1206326

Github: https://github.com/xuanhuang1/vis-project

Proposal

Background and Motivation

Both of us are conducting research in Graphics & visualizations, and the project starts as a common hobby of us. After exchanging ideas we are interested in how games as media are connected and influence players' behavior, as well as creative ways to visualize the result.

Steam also has its own hardware & software survey to present some of its user information, but as the title suggested the focus is on devices and operating systems only. There are also many third-party websites designed to provide insights into steam games data, such as Steamspy and SteamDB. These, however, often just displays and sort the data and thus serves mostly for ranking purpose.

Using two different data files, the customer data and the game description data, we are aimed at giving a new insight of the game dataset by connecting games through both tags and user behaviors.

Project Objectives

Given the two datasets, we discover that they are connected through game name. Purchase data provides individual steam customer behavior, while Steam Games data provides detailed information on each game sold on Steam. By connecting these, we can pose the following question: what other games that a user is more likely to purchase if they have purchased one game? How does each game connect to other games, and how would this connection affect a user's behavior? With this visualization, we want to answer these questions, and offer advertisement tips for game developers and market sales based on our findings.

While developing this project, there are several questions we need to answer:

1. The given dataset has too many attributes. What should be visualized to best answer the questions above?
2. Two dataset have only one attribute in common, which is the game name. What visualization design can best associate the two datasets?

Data

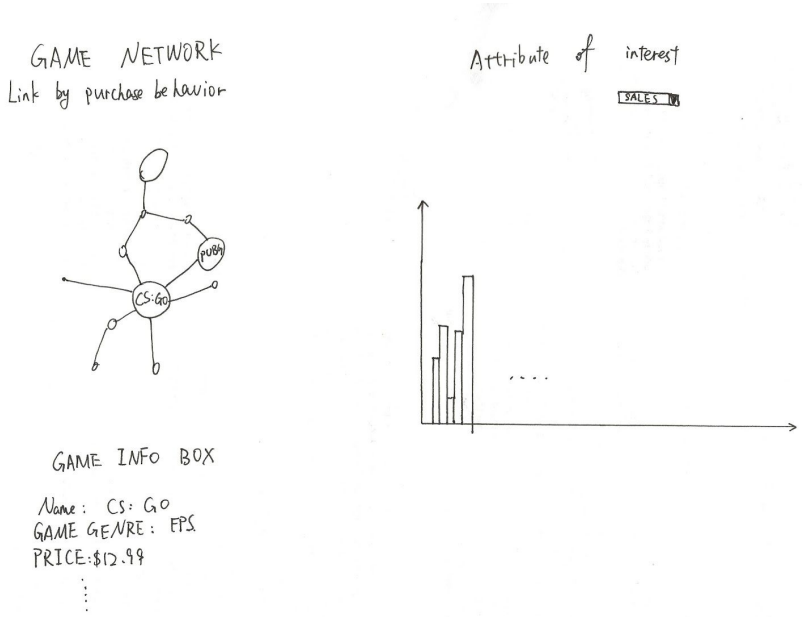
Both dataset comes from dataworld. The game-features.csv is adopted from a sample data project on github, and their original data is from public Steam datas on Steam's API and steamspy.com. The steam-200k-csv is from a kaggle page where they list 200k Steam user interaction.

Data Processing

The game-features dataset is a very large dataset and contains very extensive information about games on steam. It will slow the visualization significantly or even cause the browser not responding. Therefore we need to reduce columns in the dataset. Information such as minimal system requirement and supported languages is irrelevant to the question we are trying to answer. These columns will be deleted in data preprocessing. No extra quantities is expected to be derived from the data. The stream-200k-csv data only has four columns and are all essential to the visualization. So no cleanup will be done on this dataset.

Visualization Design

Design 1:

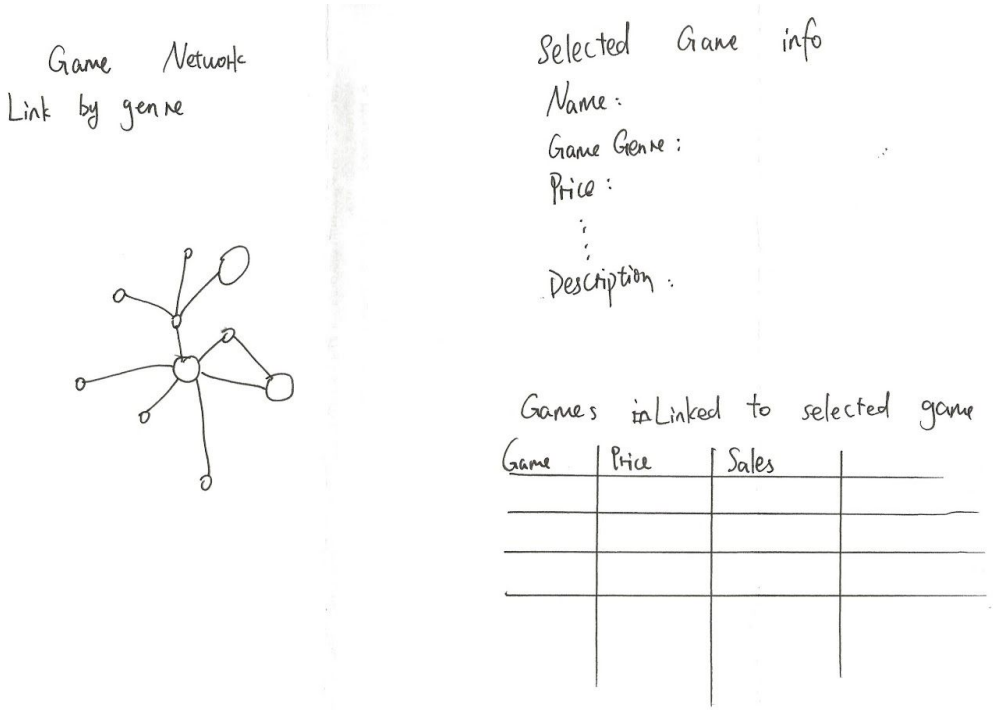


This is a multiview visualization. Left side is a network visualization. Nodes represent game titles, and there is a connection if the same user purchase the two games. The degree of node is the number of purchases on the same game. The network is interactive, and a user can select a game of interest. The info box below the network will show the detailed information of that game. On the right side, it is a visualization of a attribute selected by the user. If a quantitative attribute is selected, it is visualized using bar chart. If a qualitative attribute is selected, it is visualized as a table, column as each quality input, and the game title as a row under the column if it has the quality.

To connect the network and the attribute visualization on the right, we use highlighting to visualize the selected game title. If a node is selected in the network, the associated bar or title will be highlighted, and vice versa.

This visualization helps us to understand how a single attribute would affect the user's purchase behavior, and learn what has a bigger influence on user's decision.

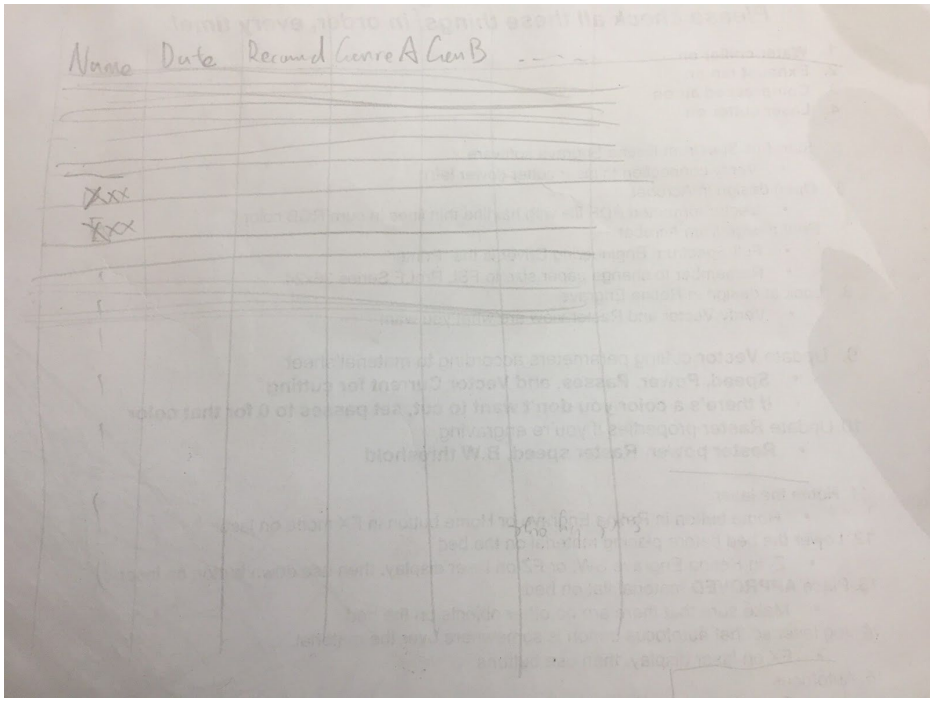
Design 2:



This design changes how the network is based on. Games are represented as nodes, and there is a link between two nodes if they share the same genre. And the degree of node is the number of purchases of a game. Similar to Design 1, it uses a info panel to show the detailed information if a game is selected. Instead of showing an attribute of interest, this shows all the attributes of the games being selected and its neighbors in the table. And the table can be sorted based on attribute selected.

This design is very easy to compare among all attributes, which makes it very easy to compare attributes in the table, and see if games that are connected have patterns among the attributes presented.

Design 3:



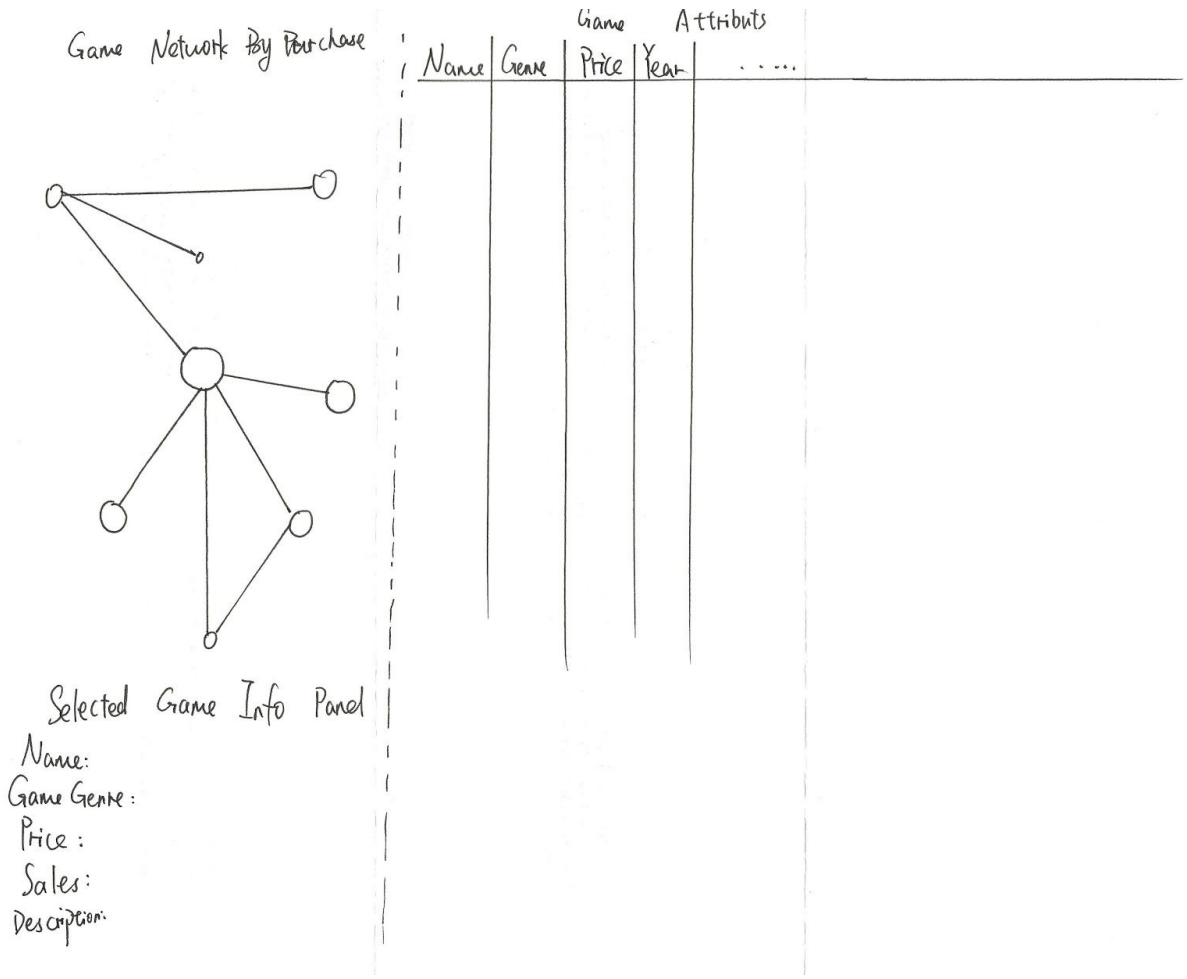
We can incorporate the idea of a magic lense into the game feature chart. The chart will be loaded in as the original, plus a column for the score from our relationship evaluation function. The function is generated by the games-200k-csv file according to user behaviors, and will be displayed and modifiable.

The chart will be resorted according to our function output. Only the selected row and its surrounding area are shown. And users can click to expand other areas of the chart to view details.

Others will be collapsed with only the genres values (true or false in the chart) shown as a black line for true or nothing for false.

Here we provide a quick 1D view of relationships, yet all information about the games are available when requested.

Final Design



The final design includes the network of games in design 1, and the attributes table in design 3. The network is based on users' purchasing behavior. The right table shows all the available data initially. Qualitative data is encoded with the the info written in the table directly, or simple dots showing categories, Quantitative data is encoded with a bar in each row. The table can be sorted according one's preference, and filtered based on attributes(only show rows with certain attributes). When selected in the network, the table will be updated to only show rows that are selected game and its neighbor. And when selected in the table, the network will only show nodes that contains the selected game and its neighbor in the network.

Must-Have Features

1. Displaying a list of the game dataset of size 10k
2. Select one game or a subset of games
3. Group (or connect in someway) games according to genre tags
4. Sort/rank the games according to user behavior data
5. An info panel to show detail attributes & descriptions
6. Multi-view visualizations

Optional Features

1. 3D view if appropriate
2. Display all games in the dataset (match all names and incorporate DLCs correctly)
3. Search for a particular game from user text input
4. System dealing with attributes other than specified above (prices, languages, number of packages etc)

Project Schedule

Week 1: clean data, decide data structure in js and view layout in html

Week 2: Populate data into webpage, has a working prototype

Week 3: Implement must have features

Week 4: Implement optional features if the progress allows

Process Book

- ### Overview and Motivation

The project will be a visualization tool of how one steam game is related to others through common buyers. Using two different data files, the customer data and the game description data, we are aimed at giving a new insight of the game dataset by connecting games through user behaviors.

There will be a zoomable view of connected graph, and a list of related games with their attributes shown in detail. The user can see, select, filter result and examine the list of games given as potential games to purchase.

- ### Related Work

Steam has its own hardware & software survey to present some of its user information, but as the title suggested the focus is on devices and operating systems only. There are also many third-party websites designed to provide insights into steam games data, such as Steamspy and SteamDB. These, however, often just displays and sort the data and thus serves mostly for ranking purpose.

There is also an embedded “Recommended > similar items” section in Steam, but the evaluation method is unclear and it is designed for commercial use, and thus might be biased.

- Questions

Given the two datasets, we discover that they are connected through game name. Purchase data provides individual steam customer behavior, while Steam Games data provides detailed information on each game sold on Steam. By connecting these, we can pose the following question: what other games that a user is more likely to purchase if they have purchased one game? How does each game connect to other games, and how would this connection affect a user’s behavior? With this visualization, we want to answer these questions, and offer advertisement tips for game developers and market sales based on our findings.

While developing this project, there are several questions we need to answer:

1. The given dataset has too many attributes. What should be visualized to best answer the questions above?
2. Two dataset have only one attribute in common, which is the game name. What visualization design can best associate the two datasets?

- Data

Data Source

Both dataset comes from dataworld. The game-features.csv is adopted from a sample data project on github, and their original data is from public Steam datas on Steam’s API and steamspy.com. The steam-200k-csv is from a kaggle page where they list 200k Steam user interaction.

Data Processing

The game-features dataset is a very large dataset and contains very extensive information about games on steam. It will slow the visualization significantly or even cause the browser not responding. Therefore we need to reduce columns in the dataset. Information such as minimal system requirement and supported languages is irrelevant to the question we are trying to answer. These columns will be deleted in data preprocessing. No extra quantities is expected to be derived from the data. The

stream-200k-csv data only has four columns and are all essential to the visualization. So no cleanup will be done on this dataset.

Data Clean Up

We combined the two dataset into one. Each entry in the game feature contains an edge list of all the games that this game is connected to. The form is in a map, with the key as the game, and the value as the link degree. The degree is number of links that these two games have, which is the total number people in the steam-200k-csv buying these two games together. Then we filtered the game-features dataset to have only game entries that have purchase data from steam-200k-csv.

- ## Exploratory Data Analysis

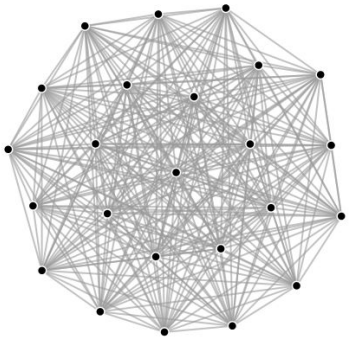
We use Excel spreadsheet to explore the data and remove some unnecessary columns. We discovered the steam-200k-csv contains game titles are all included in the game-features dataset. So we think making corresponding highlight between views will make it much easier to navigate. We also realize that the DLC content is not specifically labeled in game-features dataset, so there is no simple method to remove the DLC titles in the visualization for now.

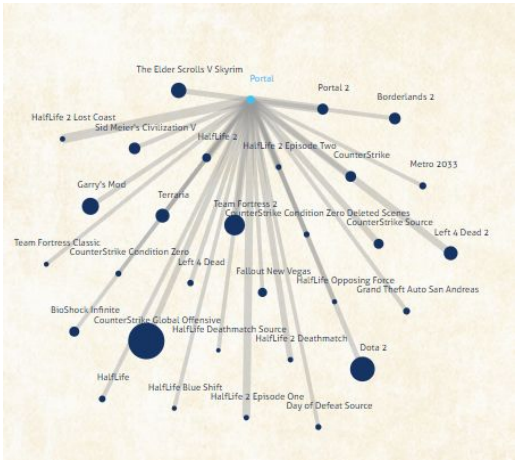
- ## Design Evolution

We received good feedback from the peer review session. We received several suggestions as add-on features to implement in the visualization. After considering all the visualization from our proposal, including different types of table visualizations and networks, we decide to have a network of games connected by player purchasing behaviors and a table with game information. This makes good use of the data we have, and is easy to navigate for the user.

After reviewing TA's feedback we decided to include filters in our table to present our large data set better. We include several reasonable choices when filtering games: release year, price, required age, platform, controller support and language.

Initially the network contains all the edges available in the network. We found out that it becomes almost a complete network every time, because several big games are almost connected to every other games. Therefore, we decided to only make a network with game being selected and its top 30 linked neighbors. This way, the network is much easier to read.

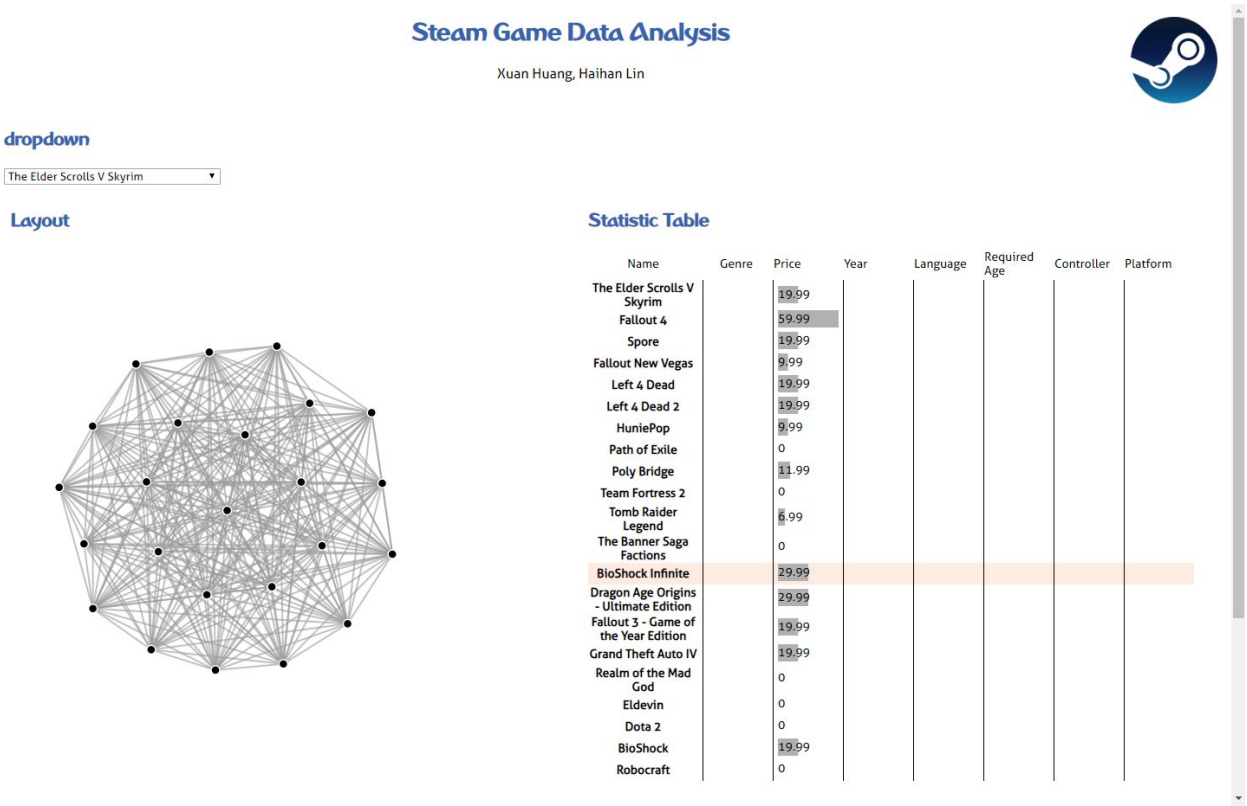


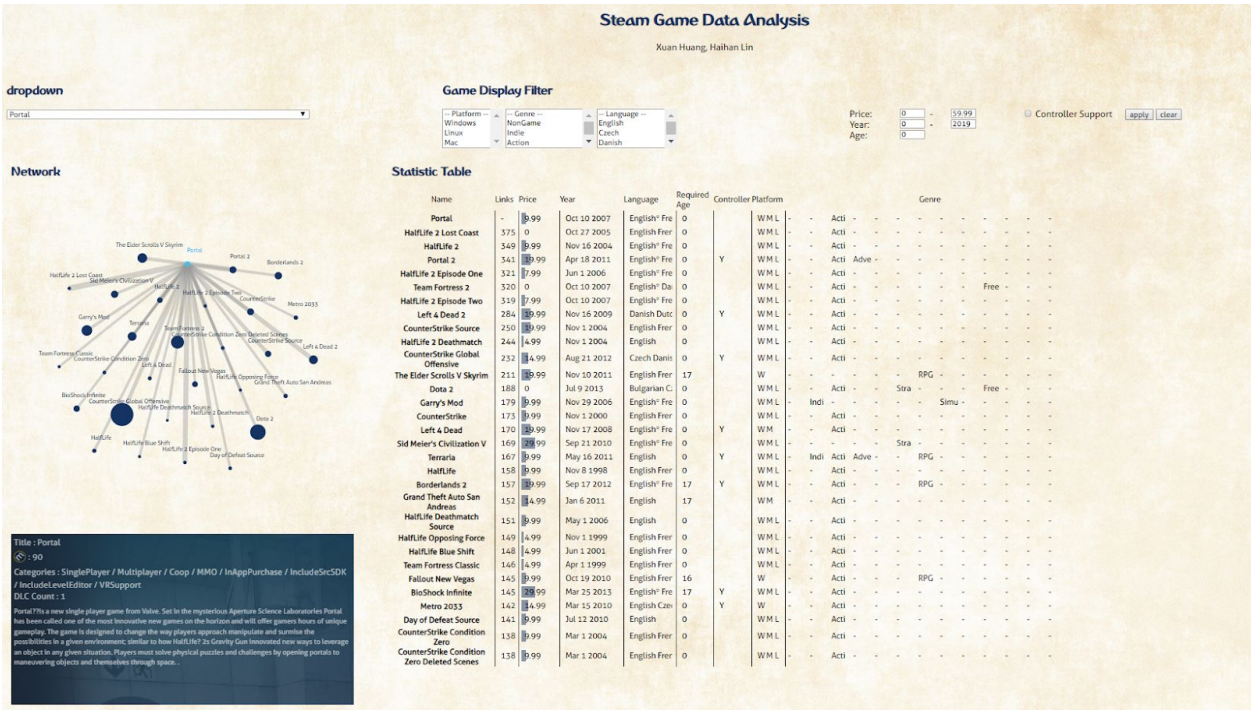


- Implementation

As in the final design we first visualized the table. As a first step the price bar is drawn as an attribute visualization example.

We use an experimental datasets of 23 games and 50 rows of user respectively to generate this initial view.





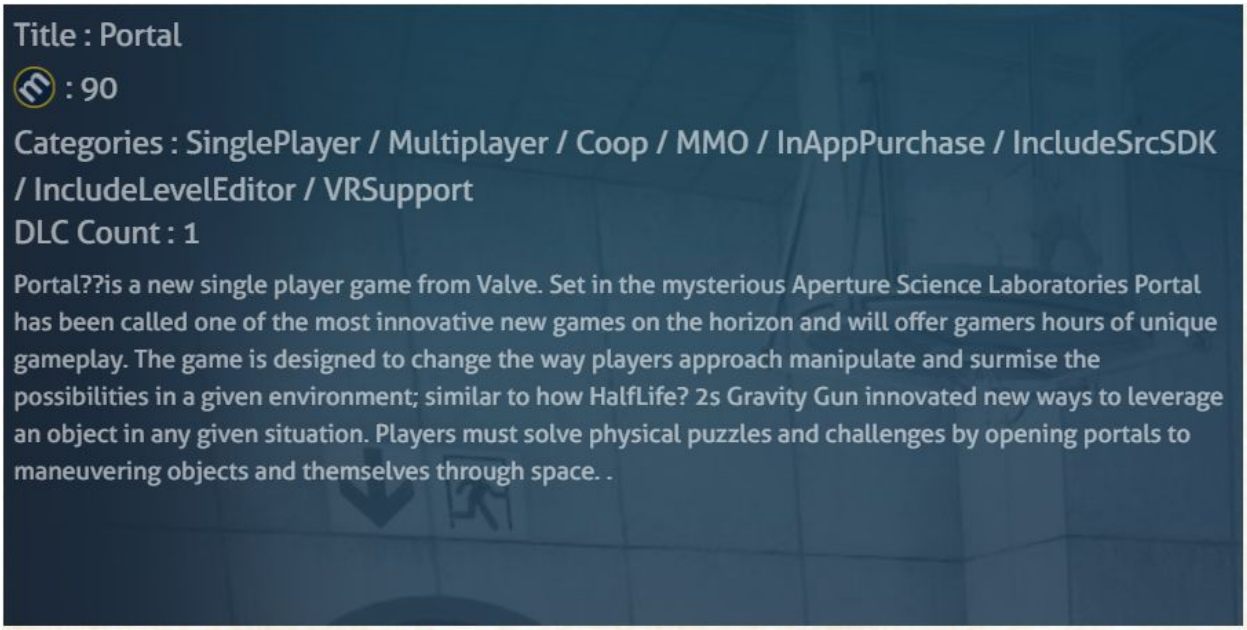
The relationship between games are defined as the number of users who purchase both. The larger the number is the more related two games are. User can select game by selecting from the dropdown menu, clicking on the name of game in table, or clicking the node in the network.

The static table contains 30 most related games of the selected one. The table will display basic informations on each game for a quick comparison, while the info panel gives more details.

Statistic Table

Name	Links	Price	Year	Language	Required Age	Controller	Platform						Genre					
The Elder Scrolls V Skyrim	-	19.99	Nov 10 2011	English Fren	17		W	-	-	-	-	-	-	RPG	-	-	-	-	-	-
Team Fortress 2	318	0	Oct 10 2007	English* Dai	0		W M L	-	-	Acti	-	-	-	-	-	-	Free	-	-	-
Left 4 Dead 2	293	19.99	Nov 16 2009	Danish Dutc	0	Y	W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
CounterStrike Global Offensive	291	14.99	Aug 21 2012	Czech Danis	0	Y	W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
Dota 2	266	0	Jul 9 2013	Bulgarian C;	0		W M L	-	-	Acti	-	-	Stra	-	-	-	Free	-	-	-
Portal 2	248	19.99	Apr 18 2011	English* Fre	0	Y	W M L	-	-	Acti	Adve	-	-	-	-	-	-	-	-	-
Borderlands 2	244	19.99	Sep 17 2012	English* Fre	17	Y	W M L	-	-	Acti	-	-	-	RPG	-	-	-	-	-	-
Sid Meier's Civilization V	230	29.99	Sep 21 2010	English* Fre	0		W M L	-	-	-	-	-	Stra	-	-	-	-	-	-	-
Garry's Mod	227	9.99	Nov 29 2006	English* Fre	0		W M L	-	Indi	-	-	-	-	-	Simu	-	-	-	-	-
HalfLife 2 Lost Coast	216	0	Oct 27 2005	English Fren	0		W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
Fallout New Vegas	213	9.99	Oct 19 2010	English Fren	16		W	-	-	Acti	-	-	-	RPG	-	-	-	-	-	-
Portal	211	9.99	Oct 10 2007	English* Fre	0		W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
Terraria	211	9.99	May 16 2011	English	0	Y	W M L	-	Indi	Acti	Adve	-	-	RPG	-	-	-	-	-	-
HalfLife 2	200	9.99	Nov 16 2004	English* Fre	0		W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
CounterStrike Source	194	19.99	Nov 1 2004	English Fren	0		W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
BioShock Infinite	188	29.99	Mar 25 2013	English* Fre	17	Y	W M L	-	-	Acti	-	-	-	-	-	-	-	-	-	-
Grand Theft Auto San Andreas	173	14.99	Jan 6 2011	English	17		W M	-	-	Acti	-	-	-	-	-	-	-	-	-	-
The Witcher 2 Assassins of	171	FALSE	Apr 16 2012	TRUE	17			NonG	Indi	Acti	Adve	Casu	Stra	-	-	-	-	-	-	-

The info panel contains the title, Metacritic score (if available), the category that the game is in, the DLC count till the date this data was collected and a short description of the game. The background of this info panel is the steam background image that the game has.



The network shows what the selected game's neighbors are. The node size is scaled by the number of recommendations that the game has, and the edge thickness is scaled by the link degree. (The number of people purchasing these two games together). When hovering on a node or a table entry, the corresponding entry or the corresponding network node will be highlighted. To change the selected game, one can use the dropdown box, click the table entry, or click the network node, and the network layout, the table elements, the info panel will change when another game is selected. (screenshot shown above in the Design Evolution section.)

- Evaluation

We have found that the game are highly connected in the network. Currently we only show the neighbors of the selected games. It will be interesting to implement some level of hiding and display other edges. Zoom-in/out did not get implemented due to time limitation. We would love to have that as future work.

We find out that games with similar tag are more likely to be purchased together, especially those with mainstream tags like Action or RPG. Free games also tend to have close relationships to each other. EA, however, doesn't seem to influence much on purchase behaviors.

The controller support doesn't seem to be a big factor of in purchase either, and neither is there a particular pattern of games with higher required age. PC is still the king

in terms of game release platform, and if a game compatible is on Mac/Linux, it is very likely that the most related games support Mac/Linux as well.

As we imagine, DLCs and games within same series appear together frequently. And the group of most related games usually fall in similar price range.