Study on User Preferences of PC Games by Analyzing User-generated Tags

Steam Tag Analysis

An, Jaewoo [1], Kim, Keon Sik[2], Yi, Taeha[3].
Department of Bio and Brain Engineering, KAIST
Department of Mathematical Sciences, KAIST
Graduate School of Culture Technology, KAIST
Daejeon, Korea
[1]epaqu@kaist.ac.kr
[2]kca0987@kaist.ac.kr
[3]yitaeha@kaist.ac.kr

***Abstract*—This paper concerns community structure of game tags and its possible implications for various stakeholders in the game industry. All data have been collected from Steam, one of the largest digital distribution platforms, that has a game tagging system. Then the data is visualized as a network and the Louvain method is applied to reveal the community structure of the network. Further analysis is performed through the resulting cluster model and alluvial diagram.**

Keywords—Game Analytics, Tag, Commnunity Dectection, Netwrok Analysis, Alluvial Diagram

I. Introduction

The study of community structures in real-life networks has always been an important and interesting area of research. In a network of research articles, for instance, communities often correspond to different academic disciplines. So, understanding what the communities mean in a network is highly beneficial to understanding the entire network. This paper deals with a game tag network, in which the nodes represent user-defined game tags. A pair of nodes is connected if a game is tagged by both nodes. Hence, the resulting network is an undirected network with weighted edges.

As for the meaning of the communities in the game tag network, the first and more intuitive hypothesis was that they correspond to game genres. If similar tags such as cRPG and tRPG appear a lot in a cluster, that cluster may be characterized as an RPG cluster. However, there is a danger in understanding the communities as game genres because genre is not an easy concept to objectively define. Classification by genre is a consensual agreement between the audience and the producers affected by categorizing in the literature or film industry [1].

The next trial is to understand the communities as users' game preferences. The game tags are, after all, user-defined and therefore they generally reflect users' intentions, thoughts, and preferences. Users will create new tags if they find the current pool of the popular or global tags not enough to appropriately describe the games. Morever, intuitively, users are more likely to tag games they play themselves. In this light, this paper can be considered as a GUR (Game User Research) because it explores each community in the game tag network. Later in this paper, a user model just like a player model is provided, using the characterizations of the communities.

One virtue of this paper is that it provides a new objective context of doing game analytics. So far, game analytics have been focusing on individual games. This is expected, because the game is a product sold, serviced, and administered by the game studios. Also, some of the traditional game metrics make sense only in terms of the game. Thus, it is very uncommon for the industry to allow diverse research contexts other than the game-specific one. But this paper will provide the game developers and analysts with a new way to look at games. This new context is tag-oriented, which is interesting, because most game tags appear universally across games. Additionally, the tag network allows various methods of network science and social sciences to be used for research purpose. In many ways, this paper will be innovative.

II. Related Works

A. User-generated tags

Tags mean the user-generated keywords that have been suggested as a lightweight way of enhancing explanation. [3] Basic features of tags are first, 'folksonomy,' which means "Folk (people) + order + nomos(law)." Second, tagging is easy to understand and do, even without training and previous knowledge in classification or indexing. User makes their words through the tagging system easily and freely. To better comprehend the idea of folksonomy, look at the case of *Doom*. <Fig 1> shows the number of occurrences for the phrases "doom clone" and "first-person shooter" in Usenet posts following the release of *Doom*. [4]

Like the previous case, users have made new terms related to the game trend. When doom came out, a lot of shooting games had appeared. People called it as "doom clone" and "First Person Shooter." Over time, all game community called doom style game as "FPS." People make a game trend and even the words.

B. *Network analysis*

Even after 10 years of tagging, games that require new tags are being released. We believe this issue can be solved by the network analysis of user tagging, because a network of user related data usually shows a trend that can explain new and rising game styles. Rosvall, M., & Bergstrom, C. T. (2010) analyzed the community structures of a citation network of articles in the field of science. They found that each community in the network corresponds to a scientific discipline, like neuro science, geology [1]. For analyzing user-generated tags, we are looking for the meanings of the communities in the game tag network.

Steam, a digital game distribution platform (<Fig 2>), has a very sophisticated tagging system that covers a lot of games [5]. So, we collected the Steam tagging data. After collecting data, we transformed and visualized the data. Next, we applied a community detection algorithm. Finally, we supported the network analysis through network metrics, logistic regression, random subnetwork and alluvial diagram.

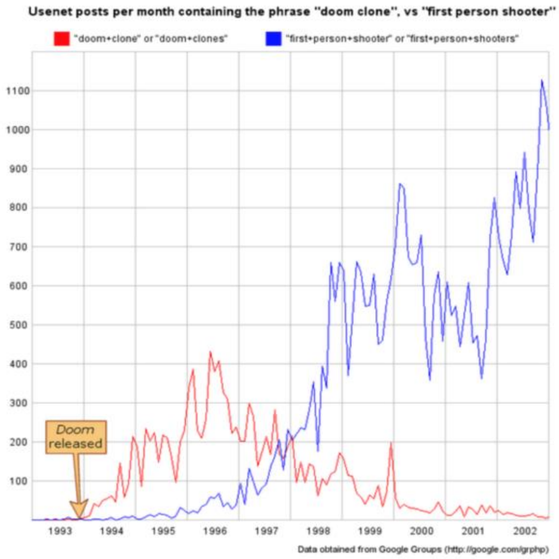


Figure 1. Appearance of FPS after the release of Doom. [4]

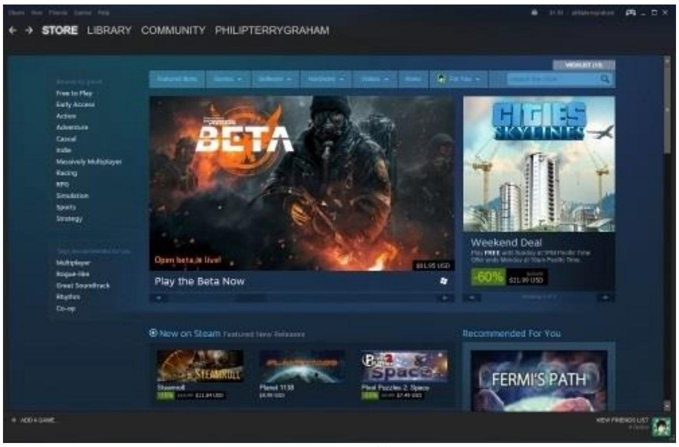


Figure 1. Steam official site

III. METHOD

A. *Data Collection*

The data necessary for this research include a complete list of all Steam tags, a complete list of all Steam games, the release dates of all Steam games, the user reviews of all Steam games and the lists of tags for each and every Steam game. The complete list of all Steam tags are fairly easy to obtain. Steam's global tags web page already lists them all. Simply copying and pasting the said page into a .csv file so the R program can read is enough. The list of all Steam games, on the other hand, has proven to be a time-consuming task. The release dates and review scores of all Steam games have proven to be just as difficult to obtain for technical issues. Both of them, however, have been collected while crawling the Steam website for the list of tags for each and every Steam game.

First, our R program reads the full list of Steam tags. For every tag, Steam has a page that lists all games that are tagged with that very tag. Our R program accesses this page for every tag, and crawl the entire list of games and their store pages. Next, the program goes to the store pages and crawl the release dates and reviews. By this point, the data in possession includes the full list of Steam tags, lists of all games for each tag, the list of all game store pages, and the list of all game release dates. Through the unique() function in R, all duplicates within the data are taken care of.

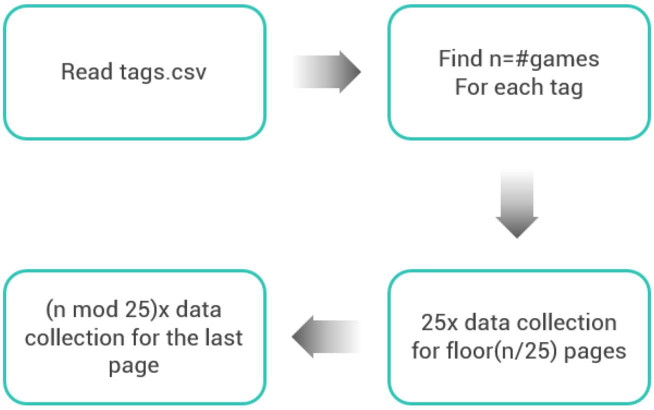


Figure 2. Process of data collection

B. *Data Transformation*

The next step is to transform the data into the desired format: the lists of tags for each and every Steam game. First, the list of all games is obtained. This is done by numerous appending of a list of all games for one tag to a list of all games for another tag. This long list is then reduced by removing duplicates with a unique() call. Once we have the list of all games, we run a for loop within a for loop in order to find all tags for each game. For every individual game in the list of all games, the program checks if the game is tagged for each tags.

When the complete list of all tags for each and every Steam game is obtained, we transform that data one step further for visualization. We use both R and Gephi for data and network visualization, and Gephi takes a list of edges for data. Thus, for each game, all possible pairs of its tags must be listed in a two column data frame (source-to-target). They are then column-bound by the release dates and the numbers of positive and negative reviews. This *edges.csv* table is further processed for Gephi. For clustering, *edgefilter20.csv* which filters out any edge with less than 20 occurrences.

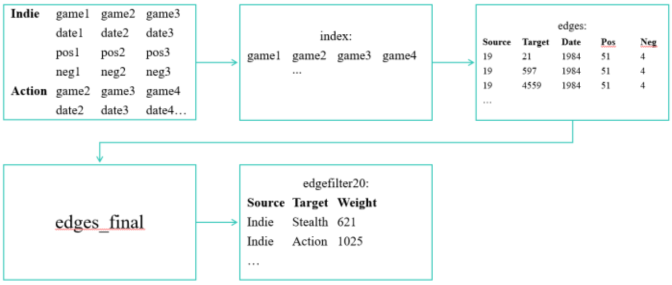


Figure 3. The process of data transformation

C. *Community Detection*

One important fact about real networks is that they tend to have community structures and each community has a meaning in the real world. For example, given a citation network of articles in physics, each community tends to represent a subfield of physics such as particle physics and statistical physics. So naturally, given a game tag network and its communities, it is reasonable to expect that each community will represent some type of games. Community detection algorithms help notice such meanings in the communities. The algorithms are based on ideas from mathematics, physics and computer science and they exploit network topologies to detect clusters in the networks. There are various algorithms:

1) Fast-greedy: As the name suggests, this is fast but is not as accurate as others in detecting communities. However since we expect a large sized data, we can apply this just to see how our network looks like, and try a different one for more accurate detection of clusters.

2) Walktrap: The basic idea behind this algorithm is that a trajectory of a random walk on a network is likely to be limited within dense subnetworks. And such subnetworks are the desired communities. This not only fast enough, but also accurate enough in detecting communities.

3) Girvan-Newman (also known as Edge Betweenness): This is a divisive algorithm, which means that the whole network is divided into communities if some given criteria are satisfied. For this one, an edge with the highest edge betweenness is removed and for every edge, edge betweeness of the new network is recalculated and so on.

4) Louvain Method: Louvain method is a optimization method that attempts to optimize the modularity of a partition of the netwrok. This method have strong point efficently maximizing the network modularity.[6]

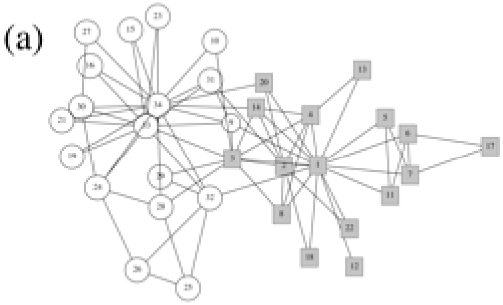


Figure 5. The friendship network for Zachary's karate club study after Girvan-Newman algorithm. [7]

One important point before applying any community detection algorithm is that our network may not be sparse. In other words, because there are too many links while the number of nodes is small. In this case there is no meaning in detecting communities. To overcome such case, we filter the links by link weight. This way we can make our network sparse enough and we can extract some useful insights from it. For this reason, we filter the edge 20. The result is 201 tags and 1,762 edges remained.

After filtering the links, we applied many community detection algorithms and found that the Louvain method gave us the best result. It gave us a modularity 0.19 and 5 communities. Below are the communities:

Table 1: Result of clustering

Cluster	Tag
Cluster 1	Strategy , Simulation, RTS, War, Tactical, Turn-Based_Strategy, Multiplayer, Historical, World_War_II, Singleplayer, Action_RPG, Turn-Based_Combat, Tower_Defense, Moddable, Hex_Grid, 4X, Military, Medieval, Realistic, Grand_Strategy, Wargame, City_Builder, Naval, Real-Time_with_Pause, Base-Building, Resource_Management, Tanks, Post-apocalyptic, Fantasy, MOBA, Strategy_RPG, Hack_and_Slash, Dark_Fantasy, Party-Based_RPG, Card_Game, Board_Game, Isometric, Turn-Based, Trading_Card_Game, Turn-Based_Tactics, Trains, Trading, Replay_Value, Flight, Driving, Third-Person_Shooter, Remake.
Cluster 2	Building, Survival, Open_World, Space, Sandbox, Sci-fi, Crafting, Space_Sim.
Cluster 3	Casual, Racing, Music, Platformer, Co-op, Comedy, Great_Soundtrack, Arcade, Shooter, Local_Co-Op, 2D, Pixel_Graphics, Retro, Difficult, Funny, Shoot_'Em_Up, Twin_Stick_Shooter, Top-Down, Bullet_Hell, Family_Friendly, Local_Multiplayer, Controller, Physics, Cute, Side_Scroller, Education, Fast-Paced, Rhythm, Minimalist, Stylized, Colorful, Abstract, 4_Player_Local, Puzzle-Platformer, Robots, Memes, Clicker, Metroidvania, Top-Down_Shooter.

Cluster 4	Zombies, FPS, Puzzle, Horror, Classic, Third_Person, Point_&_Click, Psychological_Horror, Story_Rich, Exploration, First-Person, Atmospheric, Mystery, Female_Protagonist, VR, Stealth, Choices_Matter, Relaxing, Surreal, Short Dark, Walking_Simulator, Hidden_Object, Survival_Horror, Gore, Detective, Cult_Classic, 1990's.
Cluster 5	Action, Adventure, Indie, Free_to_Play, RPG, Massively_Multiplayer, Online_Co-Op, Tactical_RPG, MMORPG, CRPG, Action-Adventure, Beat_'em_up, Arena_Shooter, PvP, Competitive, Aliens, Rogue-like, Rogue-lite, Dungeon_Crawler, Choose_Your_Own_Adventure, Procedural_Generation, RPGMaker, Management, Level_Editor, Visual_Novel, Anime, JRPG, Character_Customization, Interactive_Fiction, Cartoony, Romance, Nudity, Hand-drawn, Lovecraftian, Episodic, Multiple_Endings, Cartoon, Narration, FMV, Mechs, Dating_Sim, Sports, Fighting, Match_3, Steampunk,, Economy, Perma_Death, Magic, Cyberpunk, Loot, Kickstarter, 3D_Platformer, Touch-Friendly, Otome, Soccer, Football, Noir, 2D_Fighter, Futuristic, Parkour, Dark_Humor, Experimental, Western, Pirates, Ninja, Mechs, 2.5D, Destruction, 1980s, Runner, Split_Screen, Score_Attack, Mature, Crime, FMV, Voxel, Dystopian, God_Game, Hacking, Science, Psychedelic.

D. *Data Visualization*

On Gephi, network visualization is much easier and faster. One problem expected in this step is the unbalance between the number of nodes and the number of edges. For merely 317 tags (excluding nine non-game tags such as Audio Production, with which no game is tagged), there are over a hundred thousand games. And a game usually has 5 to 10 tags (20 at maximum). Suppose all games are tagged exactly five times. Then the total possible number of edges is approximately 100,000 times 5C2, or 1,000,000. Of course, this number includes countless duplicates of the same edges, which should not be counted for the net number of edges. Still, one million is way too large a number compared to the number of tags there are. The data is anticipated to be effectively uniform and complete in general connectivity if we do not account for the edge weights. In a complete and uniform network, clustering is meaningless. Therefore, in order to make meaningful observations, the network needs to be much sparser. Since the edge weight distribution follows an extremely exponential function, even a weak filter greatly reduces the number of edges in the network. The edge weights are used to filter the edges of the network. The threshold is 20. As for the alluvial diagram, it will be discussed later.

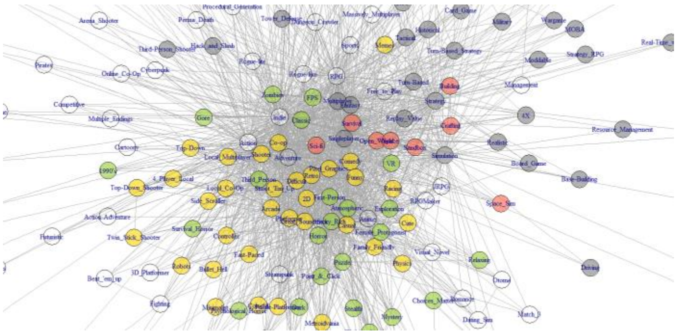


Figure 6: Visualization of 5 clusters

IV. Analysis

A. *Network Metrics*

After we know 5 game clusters, we figure out what is the statistical feature of each cluster. We are doing 5 types of metrics: degree, PageRank, transitivity, betweenness, and closeness. First, degree means the number of connection between tags. Second, PageRank is a way of measuring the importance of website pages. PageRank is similar to degree. However, it’s not same. For example, there are two tags, A and B. If A and B are connected to 5 other tags, the value of degree is equally 5 to tags. However, if A connected to weak tag, such as Dating_Sim and B connected to powerful tag, such as Indie, tag A of page rank is higher than tag B. Third, transitivity means possibility of when A is related to B and B is related to C, then A is related to C. Forth, betweenness means the number of cases when one node is located to shortest path between other two nodes. It can be mentioned the possibility of bridge between others. Fifth, closeness means the relation between two tags. For example, in social relation, if your closeness is higher, you can meet anybody in the world with just two bridges. The result of 5 metrics is <Fig 7>.

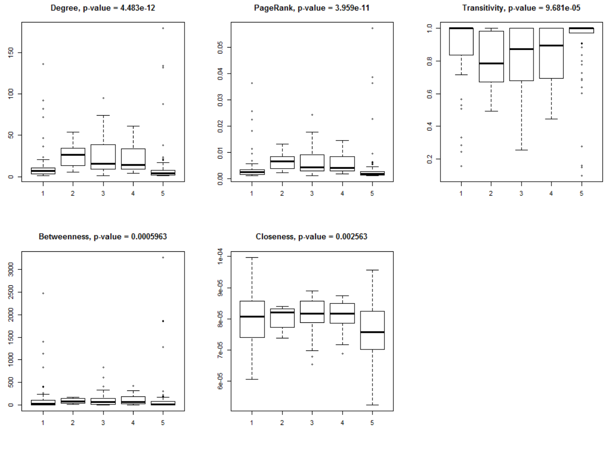


Figure 7: Result of network metrics

We can find similar pattern between degree and PageRank. However, cluster 3 and 5 show a meaningful difference. The degree of cluster 5 is close to 0 while its transitivity is close to

1. High transitivity implies a dense network structure. But low degree suggests the exact opposite. This happens because of the unique structure of cluster 5, in which majors nodes such as Indie and Action are tightly bonded and minor nodes such as Mature and FMV are connected to the central chunk of nodes by a single edge. The transitivity is high only because the algorithm cannot calculate transitivities for dyads. This is the very definition of a sparse network. We can say that games made within cluster 5 are structurally general because they will most likely share the central chunk of nodes, which will shape the fundamentals of the games. The only differences among the games will be determined by the suburb dyads. We can say that cluster 5 is the collection of tags for stereotypical games.

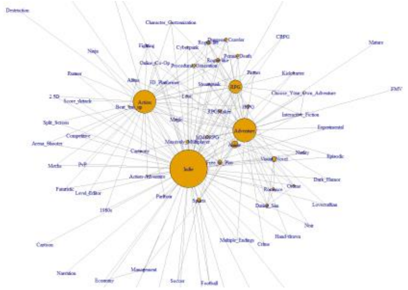
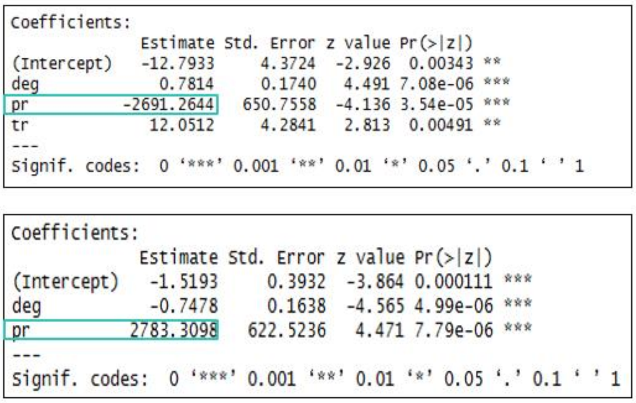


Figure 8: Cluster 5

B. *Logistic Regression and Lessons Learned*

Logistic regression for each cluster models the probability of a tag belonging to the cluster. If any significant pattern is observed in terms of the 5 metrics, then some characteristics of clusters can be elaborated. Surprisingly cluster 3 and 5 exhibited significances while others did not. The <Fig 9> is the result for cluster 3 and 5.



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -12.7933     4.3724  -2.926  0.00343 ** 
deg             0.7814     0.1740   4.491 7.08e-06 ***
pr          -2691.2644   650.7558  -4.136 3.54e-05 ***
tr             12.0512     4.2841   2.813  0.00491 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -1.5193     0.3932  -3.864 0.000111 ***
deg           -0.7478     0.1638  -4.565 4.99e-06 ***
pr          2783.3098   622.5236   4.471 7.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: The result of logistic regression of cluster 3 and 5.

On the cluster 3, the estimated coefficient for page rank is unusually low. One unit of increase in PageRank reduces the log-odds by -2691. On the cluster 5, in the contrary, the estimated coefficient for PageRank is unusually high. One unit of increase in PageRank increases the log-odds by 2783. This implies that an increase in the PageRank of a tag will shift that tag from cluster 3 to cluster 5, and vice versa. Intuitively this makes sense as well, because tags in cluster 3 are mostly found within indie games. Indie games may be larger in number, but the public is not exposed to indie games in general. The games made within cluster 3 are, in that sense, *atypical*. An increase in edge weight implies that a previously atypical (cluster 3) combination of tags becoming typical (cluster 5). Of course, both an increase and decrease of PageRank will also have an impact on the general network structure, too. It is even possible that the clusters will change. But it is still a valuable lesson in terms of understanding the current game market.

C. *Random Subnetwork*

Each complete subnetwork in the tag network can be considered as certain imaginary game. In this analysis, coded software randomly selects such subnetwork both from the entire network and each cluster. We rely on the idea social influence. Next, we assign scores for tags based on the positive or negative user review data. Score metrics for each edge can be defined in terms of positive and negative review numbers. We define score for each tag in terms of the edges that are connected to the tag like <Fig 10>.

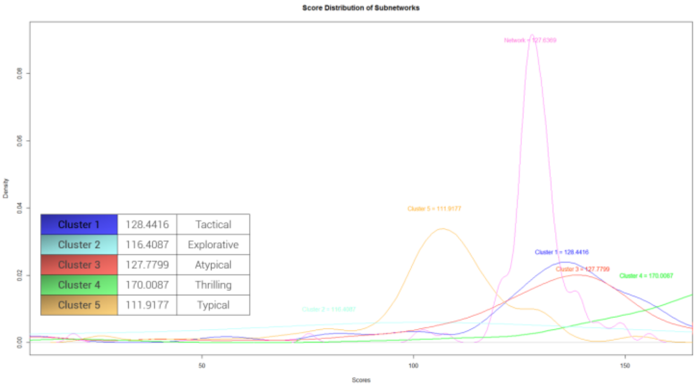


Figure 10: Score distribution of subnetworks

D. *Alluvial Diagram*

Temporal trend of clusters in our network is also of interest to our stakeholders. Clearly some real networks change their mappings over time, and different clusters diverge and converge at different points of time. Such phenomena can be well observed using alluvial diagrams. Below is one example from [1]:

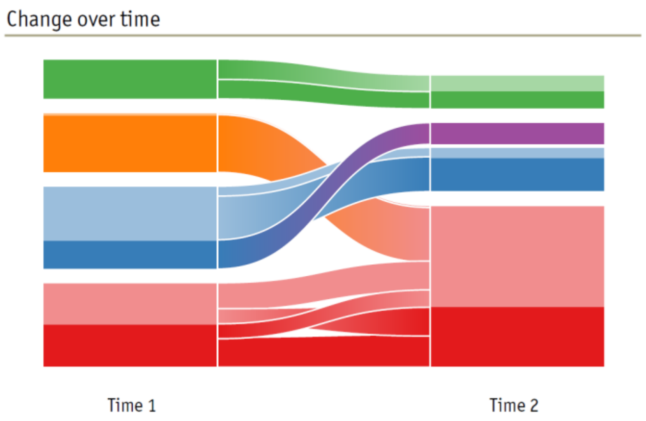


Figure 11: An illustration of alluvial diagram.

We use the release data of all games in the Steam by collecting when games are released. We extract 1984's games and to put 1993's games, and so on. When the released date is written such as "when the worlds end", we all set 2018. The timeline of diagram is from 1993 to 2018. Also, '0' means tags are not existed that period. Each cluster is distributed by color. We can see the result of diagram in <Fig 12>. From the diagram, we can see 5 insights. First, appearance of tags changed a lot during the timeline we set. Second, thrilling games are most preferred steadily. Also, thrilling games get the highest score in the subnetwork. Third, cluster 2 vanished in this diagram. Cluster 2 set in 2013 as 'Explorative', however, it is gone when 2018. We can see 'Explorative' involved to the 'Tactical' and also, cluster 2 has just 8 tags. Forth, we can find various segmentation of game in 2003 to 2013. It means user's preference to the games is changed extremely in that period. For example, in 1998, there was no FPS section. However, it appeared in 2003. It shows same conclusion compared to "Appearance of FPS after the release of Doom <Fig 1>."

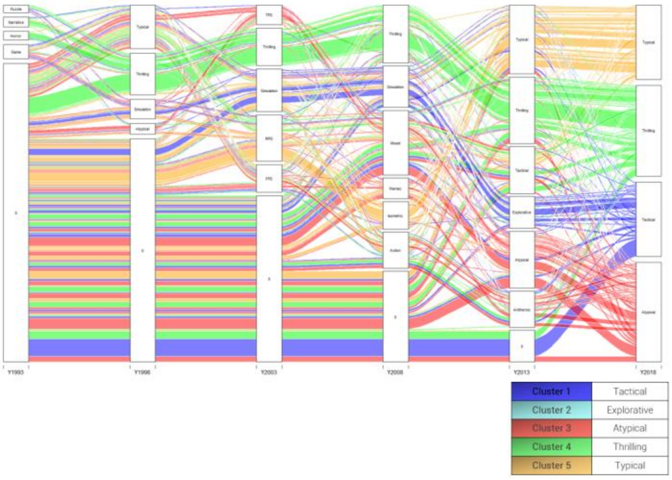


Figure 12: Alluvial diagram based on released date

E. Result

Through the 'Random Subnetwork' and 'Alluvial Diagram', we can interpret clusters to game communities based on user preference. Cluster 1 represents war games based on the strategy and simulation. Cluster 2 represents open world games that explore the space, building, and crafting. Therefore, cluster 1 shows strategy type and cluster 2 shows 'Explorative' type. However, as we can see in alluvial diagram, cluster 2 enter to the cluster 1. So, we can interpret cluster 1 to 'Tactical' type. Cluster 3 and 5 has similar in point of degree, however, it has a complete contrast in PageRank. Cluster 5 has most three powerful tags like 'Action', 'Indie', 'Adventure.' They represent 'Typical' tags. Cluster 3 is negative response about PageRank. However, score of review is higher than cluster 5. It means cluster 3 has 'Atypical' tags and also, user who like uniqueness of indie games, such as, 'funny', 'Great_soundtrack', 'Casual.' Lastly, tags of cluster 4 are 'horror', 'Story Rich', 'Mystery', 'Survival', and etc. They indicate 'Thrilling' type. We may interpret five clusters to user preference like <table 2> through the result of analysis.

Table 2: Game communities based on user preference

	Type of Game	Type of User
Cluster 1	Tactical	Structural Users based on the Rule
Cluster 2	Explorative	Free Users who like to explore
Cluster 3	Atypical	Users seeking the Uniqueness
Cluster 4	Thrilling	Maniac Users who like thrilling
Cluster 5	Typical	Users seeking the common

V. Conclusion

We analyze the game tag generated by users. It shows that there are game communities based on user preference. For analyzing this communities, we set 5 steps. Consequently, we distribute game communities through network analysis. First, we collect tag data from Steam, the largest digital distribution platform in the industry. Steam provide tag system that user can insert his thought or terms about certain games. Tags can be registered by many users, not just a few people. Therefore, analyzing Steam tag can be the user research based on user-generated tag. From Steam, we collect tags, release date, and positive or negative user reviews. Second, we transform the collected data. Making network sparse to view the relationship between tags clearly, we filter the edges to 20. Third, we are clustering these data by using Louvain method that maximizing the network modularity. Forth, to figure out what is the feature of clusters, we conduct 5 network metrics: degree, PageRank, transitivity, betweenness, and closeness. Fifth, we conduct random subnetwork based on score of user reviews and alluvial diagram from 1993 to 2018.

We can suggest the results to various game stake-holder, such as game developer and game designer. It is hard to say our results are correct in all parts. However, by analyzing user-generated tags, we can recognize insights to think about preference of users in the beginning level of making games. First, users give positive response to the unique games, not typical games. Second, current game trends are fixed. Game stream was divided to several segmentations in 2008 or 2013. In the contrary, game stream shows combination to 4 ways. We can predict, in the near future, streams may be distributed to several ways. Third, thrilling games have the high possibility of appealing to users, especially maniac users.

Our research has two limitations. First, we cannot guarantee the scores of actual games that have tags across clusters, because we set scores based on the imaginary games by random subnetwork. Second, we should evaluate it to the people who are working in the game industry. If we can hear the actual voice of game stake-holders, we can find a way how to use the user's preference by analyzing on the user-generated tags.

ACKNOWLEDGMENT

Our research has been progressed by the class (class code: GCT 544), Game Studies (prof. Doh) in GSCT, KAIST.

REFERENCES

[1] Rosvall, M., & Bergstrom, C. T. (2010). Mapping Change in Large Networks. PLoS ONE, 5(1). doi:10.1371/journal.pone.0008694.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. *Journal of Digital Information*, *10*(1).

[4] Arsenault, D. (2009). Video game genre, evolution and innovation. *Eludamos. Journal for Computer Game Culture*, *3*(2), 149-176.

[5] Steam official site: http://store.steampowered.com

[6] De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011, November). Generalized louvain method for community detection in large networks. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on* (pp. 88-93). IEEE.

[7] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826. doi:10.1073/pnas.122653799.