
Detection of Community Trends Among YouTube's 'Iceberg' Videos

Anonymous Author(s)

Affiliation

Address

email

1 Introduction and Problem Statement

The 'iceberg chart' is an image format used to convey information about a certain subject in an easily digestible way. At the top of the image, or 'the tip of the iceberg', well known information about that subject is listed, and as you progress down the image, more obscure information about the subject is listed. Starting in 2020, YouTubers have been making 'iceberg explained' videos, simply explaining all the information found on an iceberg chart. In the five years since, iceberg videos have become a prominent genre on YouTube, with thousands of videos covering a range of different subjects.

However, unlike other YouTube genres such as commentary, gaming, or beauty, there is no commonly recognized 'community' among iceberg YouTubers, since while the video formats are the same, the video subjects differ vastly. However, anecdotal observations have shown that some people enjoy watching iceberg videos in general, no matter the subject. So, as someone who both creates and views iceberg videos, it is in my interest to ask- is there a community of people who enjoy iceberg videos in general? Or in other words, is there a significant network of people who watch a variety of iceberg YouTubers? Are there further sub-communities within a greater iceberg video community based around creator or genre? Are there specific videos or YouTubers which act as 'hubs' the rest of the community are centered around? Through answering these questions, valuable analytic information will be uncovered which can help iceberg YouTubers with collaboration and marketing strategies.

2 Literature Review

The first paper consulted in this project is "Community Detection in Graphs" by Santo Fortunato. This paper provides a broad explanation of various community detection techniques, including their best applications and shortcomings. [1] While Louvain, a method not directly mentioned in this paper, is used in this project, it still provides valuable context and possible alternative solutions.

The paper "From Louvain to Leiden: Guaranteeing Well-Connected Communities" by Traag et al. provides an explanation of the Louvain algorithm, which was consulted as this project will use an implementation of the algorithm. However, while the paper praises Louvain for being "simple and elegant", it also highlights the shortcomings of the algorithm, mainly regarding its tendency to find badly connected communities. The authors then propose the Leiden algorithm as an alternative which mitigates this problem by adding a refinement step to the Louvain algorithm after the nodes are grouped into communities. [2] While this project will use Louvain for its ease of implementation and fast computational speed, if improved results are needed, this paper will provide insight as to why the results provided by Louvain are unsatisfactory, as well as offer the Leiden algorithm as an alternative.

The paper "Large-Scale Community Detection on YouTube for Topic Discovery and Exploration" by Gargi et al. provides a relevant example of community detection among YouTube videos. Similar to this project, Gargi et al. represented YouTube as a graph, with each video being a node. However, the paper's graph defines edges in the graph as similarity between the videos, with similarity being determined by how many users co-watched those videos in anonymous sessions. The paper is

concerned with the creation of a new clustering algorithm, since current clustering algorithms do not work well on a graph the YouTube graph with tens of millions of nodes. The paper proposes running local clustering algorithms in parallel in combination with preprocessing and postprocessing steps to efficiently detect communities within the large graph. [3] While my project uses a dataset with only thousands of videos, this paper still provided valuable information on how to construct a graph of YouTube videos to detect communities among videos.

Finally, the paper "Surprise maximization reveals the community structure of complex networks" by Aldecoa and Marin offers a framework for analyzing the results of the community detection algorithms implemented. Are the communities detected just artifacts of the network structure which can occur in any similar graph, or do they offer meaningful insight? Due to the large variety in sizes and structures of graphs, traditional benchmark measures such as modularity are inconsistent in determining if a community detection algorithm successfully detects meaningful communities. However, the authors propose a new benchmark called Surprise, which measures the probability of edges within a graph's sub-community forming given that edges were assigned randomly in the graph. It is calculated with the following formula:

$$S = -\log \left(\sum_{j=p}^{\min(M,n)} \frac{\binom{M}{j} \binom{F-M}{n-j}}{\binom{F}{n}} \right)$$

Where n is the number of links in the graph, p is the number of intra-community links, F is the maximum number of links possible in the graph, and M is the maximum number of intra-community links possible. If the calculated value for S is low, that means it is likely for the subcommunity to have appeared randomly. On the contrary, a high S value means the subcommunity is unlikely, or 'surprising' to appear randomly [4]. This is useful for this project, as Surprise can be used to determine if the communities that arise in the graph of shared commenters among videos is statistically significant, offering insight into the relationships between these videos, or if the communities found are just random clusterings of videos.

3 Methodology

3.1 Graph Creation and Community Detection

This project will investigate the community structure among iceberg videos by constructing a graph where each node is a video, and weighted edges represent viewer overlap as approximated by shared commenters.

Video and comment data for a curated set of iceberg videos will be collected using the YouTube Data API. As stated before, this data will be used in the creation of a graph visualizing the relationships between different iceberg videos. For each pair of videos, the Jaccard Similarity is computed:

$$J(v_1, v_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

where C_1 and C_2 are the sets of commenters on videos v_1 and v_2 , respectively. Undirected edges are only added between nodes if the computed Jaccard similarity is greater than 0, or if the two videos share a commenter. In the future, the threshold may be increased if necessary. The edges are weighted with the computed Jaccard similarity, showing how strong the shared viewership between the two videos is.

There will be two community detection tests performed- one to test for general communities across different iceberg video channels, and another to see if there is a community among iceberg videos covering similar subjects.

The Louvain method for community detection will be used. This method calculates modularity using the density of edges between nodes, defined for a weighted graph as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Each node starts as its own community. Communities are found by optimizing node modularity, or checking if adding the node to a neighboring community increases modularity. Then, each local community is aggregated into a single node, with reduced edges between nodes. The process is recursively repeated on the new graph, until no further optimization can be made, meaning that the algorithm has identified all the major communities from the original graph, represented by each node in the final graph. [2]

While improvements upon the Louvain algorithm have been made, such as the Leiden algorithm, Louvain is still preferred initially, as it is easier to implement, and can still provide valuable initial insight. If necessary, more robust algorithms can be used for further investigation.

Since it is already expected that there will be an established community for a YouTuber, it is likely that the Louvain algorithm will first identify videos by the same YouTuber as part of a community. If the algorithm stops here and does not identify cross-creator communities, the edges representing shared commenters on videos uploaded by the same channel will have a filter applied to them after to reduce their weight. Visually, they will still be grouped together, but their importance to the algorithm will be reduced. This step can help Louvain identify inter-creator communities easier if intra-creator communities are too dense.

This experiment will also identify 'hub' videos that communities may form around or which connect different communities, various centrality measures will be used. Degree centrality will be used to measure videos which act as central hubs a community forms around, as this centrality measure identifies which nodes have the most connections to other nodes, or which videos will have the most shared commenters to other videos in its community. Betweenness centrality will be used to identify videos which connect communities, as this centrality measure identifies nodes which act as a bridge between two clusters.

3.2 Custom Dataset Creation

The dataset for this project is custom made. It will be collected using YouTube's Data API v3, provided by Google. This API will be used to gather video metadata such as titles, creators, and commentors.

The videos for the dataset come from two community playlists- the first being "My Favorite Iceberg Videos" by YouTube user Jack Hageman, containing 2,186 videos.¹ The second playlist is "Iceberg Videos" by YouTube user Saint Noah, containing 1,319 videos.² The playlists will be read in, with data such as video title, author, and comments being tracked. Any duplicate videos will be filtered out. Additionally, to ensure a feasible scope, an additional filter will be placed to ensure that only videos from channels with at least ten unique videos in the dataset will remain. With this, we avoid the dataset being dominated by channels with only one or two iceberg videos, and ensure that the analysis is performed on established Iceberg YouTubers. After running the playlists through the pipeline, the dataset consists of 1,727 videos and 198,824 comments. However, this number may change slightly in the future, as viewers are still commenting on the videos in the set, and the playlists are still being consistently updated.

The dataset for the test to determine if communities form around video subject will be the same as the initial test. However, to label the videos, the titles of all videos in the dataset will be ran through a pipeline utilizing the NLP model Bert's sentence transformer application, SBERT. This will perform semantic analysis on the title and classify them among pre-defined categories.

3.3 Evaluation Criteria

A quantitative and qualitative analysis will be performed of the communities detected within the graph. The quantitative analysis will be conducted via the 'Surprise' method described in the 'Related Literature' section of the paper. A random partition will be used as a baseline. For each sub-community detected at each depth by Louvaine, the size distribution of those communities will be used to generate 200 randomly assigned partitions. A Surprise value will be calculated for each of these partitions, which will be aggregated into a distribution of surprise values which will be used as a benchmark for communities of that size. Then, the Surprise value of the Louvain-detected community

¹<https://www.youtube.com/playlist?list=PLgxL0tm8FuoeCSN0KgPW4dq2LnzmrLBdZ>

²<https://www.youtube.com/playlist?list=PLt6oQcvfST4PT8csOpMhO6JbasOotBCet>

will be calculated, and it will be compared to the benchmark via Z-score. The higher the z-score, or the more standard deviations above the random distribution the Louvain community scored, the more significant the community is. For this experiment, a z-score greater than 10 will constitute a significantly strong community.

The qualitative analysis will be performed manually through visual inspection of the graph structure and human labeling of video genres. First, visual examination of the community layout will allow assessment of whether communities form around individual channels. Since nodes are color-coded by channel, a channel's community would appear as tight clusters of a single color. Additionally, the hub nodes identified through degree and betweenness centrality will be highlighted in a visualization. By examining which videos these hubs correspond to, I can determine, based on my personal knowledge of the iceberg videos on YouTube, whether they are plausible central connectors in the network. Additionally, validating the hypothesis of communities forming around genres of videos will be done via labeling. Each video in the sub-community will be labeled as one of six genres: "Horror"- covering true crime, creepypastas, and anything presented as 'scary', "Mystery"- covering any video with 'mystery' in the title, "Gaming"- covering video games, "Internet"- covering internet culture such as lost media, explorations of websites, influential internet figures, etc, "Culture"- covering general culture like movies, music, and television, and "Info", covering general informational topics like science or history. While these categories do have some overlap and do not perfectly cover every single Iceberg video, they are a good broad approximation. Upon looking at how videos in the sub-communities are labeled, if each sub-community has a majority of nodes sharing a label, it can be said that a community has formed around that genre. This also works as another evaluation of the validity of the communities detected by Louvain- if Louvain identified communities of genres, it means it identified significant communities.

151 4 Results

The networkx library was used for the construction and analysis of the YouTube video network. The spring layout is used, so that strongly connected nodes are closer together.

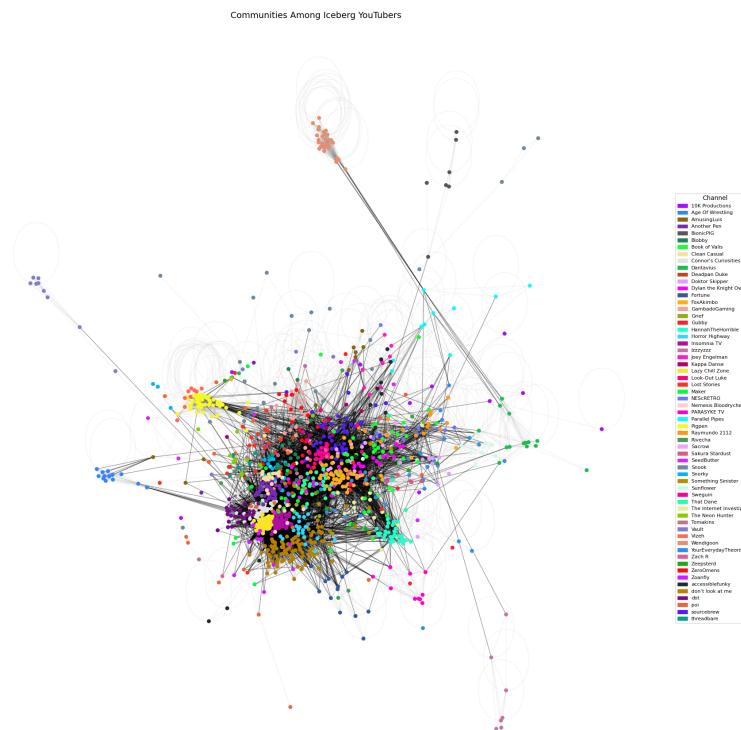


Figure 1: Network Representation of the YouTube Iceberg Community

Figure 1 shows a strong first step towards reaching the goals of the project. The large singular cluster in the middle showcases that there is a relevant community of commenters, and by extension, viewers of iceberg chart videos. Only a small amount of the total videos are isolated with little to no shared commenters. Additionally, with each color representing a channel, we see clusters of nodes with the same color. This indicates shared commenters within a channel's videos, aligning with the previous hypothesis that there would be strong communities videos uploaded by the same channel.

4.1 Louvain Community Detection

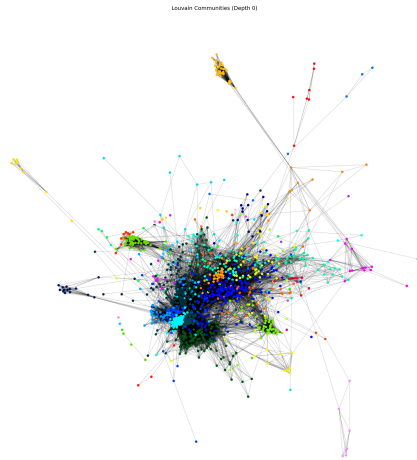


Figure 2:

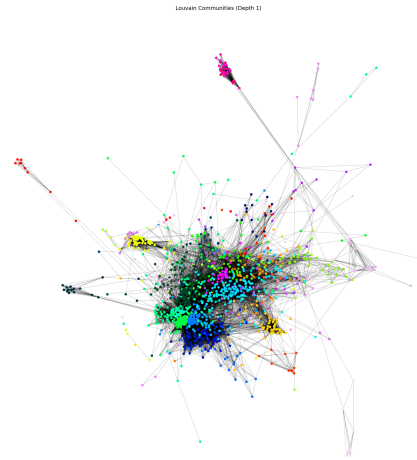


Figure 3:

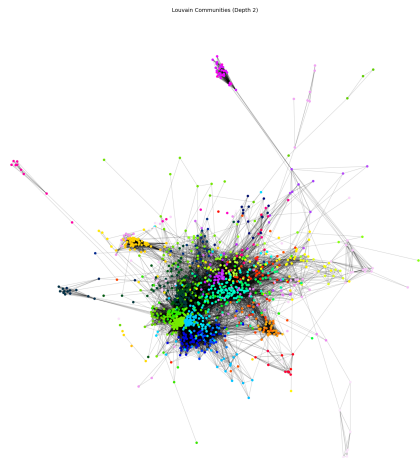


Figure 4:

Figures 2, 3, and 4 show the results of using Louvain community detection. Since the initial graph already featured a robust cluster that shows that a general "Iceberg videos" community exists on YouTube, Louvain should find relevant communities within the broader Iceberg community. Figure 2 showcases Louvain community detection at depth level 0. These results are promising- they show that large sub-communities do exist. While some sub-communities, mostly on the fringes of the graph, do line up with the channel clusters observed earlier. However, in the central clusters, we can observe that some sub-communities differ from the previously observed channel based communities, and even include nodes previously belonging to two different channels. This shows that cross-channel

communities do exist. Figures 3 and 4 show deeper iterations of Louvain community detection. Figure 3 shows promise, as while it is mostly similar some smaller communities nested within larger sub-communities form, potentially being smaller genres (eg. a videos about a specific video game within the larger gaming community). However, upon surface evaluation of depth 2 Louvain detection it seems to have the same modularity as the graph in Figure 3, with only a few new divisions made for single nodes, meaning that new significant sub-communities may not have been detected at this depth.

Next, surprise scores were calculated for the sub-communities detected at all three depths. Using the algorithms described in Section 3.3, a random distribution will be used as a baseline for Surprise score comparison. The comparison will be done via Z-Score, calculated using:

$$Z = \frac{S_{Louvain} - \mu_{random}}{\sigma_{random}}$$

Depth	Louvain Surprise	Mean Surprise (Random)	Std. Dev. (Random)	Z-score
0	41486.5472	1.8545	2.9040	14285.5882
1	6091.6834	2.6200	4.7598	1279.2686
2	0.3091	4.7051	8.5739	0.5127

Table 1: Surprise values and Z-scores for Louvain partitions at depths 0, 1, and 2.

The Z-scores align with the previous hypothesis regarding the strength of the detected communities. The Louvain Surprise values for depths 0 and 1 make sense in the context of the surprise scores reported in the paper Surprise maximization reveals the community structure of complex networks", which reported values ranging from approximately 0 to 100,000 as part of their benchmarks. However, depth 2 has a very low surprise score, indicating that the 'communities' detected at that depth are not significant and likely just random partitions. This aligns with the visual analysis of the graph showcasing communities detected at depth 2. Finally, even though depth 0 and 1 have high surprise scores, the surprise score of depth 0 is an order of magnitude greater than the surprise score of depth 1, meaning the communities detected at depth 0 are much more significant than those detected at depth 1. This also aligns with the hypothesis that depth 0 would have the most significant sub-communities. The Z-scores affirm the previous results, since the Z-scores for depths 0 and 1 exceed the threshold of 10 required for a significant community. However, the communities of depth 2 only have a z-score of 0.5, falling below the threshold.

The qualitative analysis was done via labeling the top five sub-communities detected at depth 0 with genres.

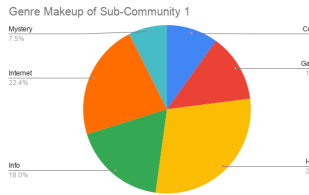


Figure 5:

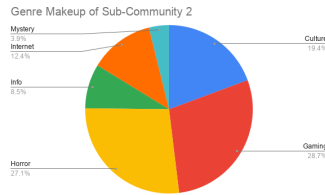


Figure 6:

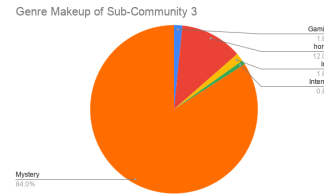


Figure 7:

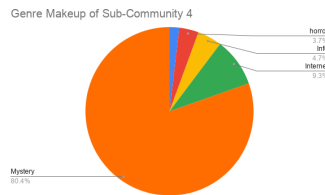


Figure 8:

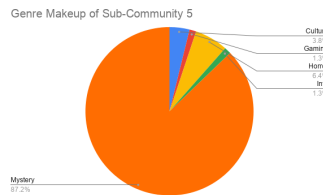


Figure 9:

As seen in figures 7, 8, and 9, most of the major communities formed around mystery videos from various channels, confirming the hypothesis that sub-communities would be detected around genre as well as channel. However, figures 5 and 6, upon first glance, appear to be communities of multi-genre videos. In figure 5 specifically, the largest genre is horror, making up 29.2 percent of the distribution. However, upon looking at the videos belonging to other genres in the group, it is found that many of the videos are horror-adjacent despite being in another genre. For example, the video "The Five Nights at Freddy's Iceberg Explained (Part 1)" was labeled under the video game category, as the subject is a video game. However, since it is a horror video game, it makes sense that it would be a part of the same community as most of the horror videos. Likewise, "The COMPLETE Disturbing Wikipedia Iceberg Explained" falls under the Internet category, but the 'disturbing' keyword in the title makes it horror-adjacent. Figure 6 has the Gaming and Culture category make up almost half of the community, a much larger share than those genres make up in any of the other communities. This could indicate that people interested in icebergs covering video games also enjoy icebergs covering other entertainment like films, shows, and music. Overall, the qualitative analysis supported the sub-communities detected by Louvain.

4.2 Centrality Detection

Degree and Betweenness centrality of nodes was computed using functions from the NetworkX Library. The top ten nodes for each category are listed in Tables 2 and 3. Figure 10 visualizes where on the graph the hub videos are located.

Looking at the most central videos defined by degree centrality, all of them fall under the mystery genre, and most are created by the Channels Insomnia TV and Lazy Chill Zone, which are two of the channels with the most videos in this graph. In Figure 10, most of these videos are located in the large central cluster, which is to be expected. Combined with the previous findings that many communities formed around the mystery genre, these videos having the highest degree centrality strengthens the community findings.

Title	Channel	Degree Centrality
obscure UNSOLVED MYSTERIES iceberg explained p...	Lazy Chill Zone	0.167036
Midwest Unsolved Mystery Iceberg Explained Part 2	Insomnia TV	0.159645
Insomnia Community Unsolved Mysteries Iceberg ...	Insomnia TV	0.153732
Unsolved Mystery Mega Iceberg Explained Part 20	Insomnia TV	0.150776
ULTIMATE Unsolved Mysteries Iceberg Explained ...	Lazy Chill Zone	0.149298
Puzzling UNSOLVED MYSTERIES That Defy Logic An...	Lazy Chill Zone	0.148559
The Ultimate Unsolved Mystery Iceberg Explaine...	Connor's Curiosities	0.147820
OBSCURE UNSOLVED: Iceberg Mysteries Revealed	Insomnia TV	0.144863
ULTIMATE Unsolved Mysteries Iceberg Explained ...	Lazy Chill Zone	0.144863
ULTIMATE Unsolved Mysteries Iceberg Explained ...	Lazy Chill Zone	0.144863

Table 2: Top Videos by Degree Centrality.

of videos of similar subjects. Degree centrality tests further showed how communities would form around videos of certain genres, and betweenness centrality tests showed which videos connected communities, whether they be genre or channel communities. While further recursive Louvain tests were performed, the second layer, despite being statistically significant, was still an order of magnitude less significant than the first layer. The communities found by the third Louvain test were found to be statistically insignificant, meaning that it's unlikely niche communities within communities exist.

However, this may also be due to flaws in the methodologies used, which can be areas to focus on in possible extensions to this work. The Louvain algorithm has a known drawback of identifying poorly connected fragments as 'communities' [2], which could be why the recursive steps in this experiment did not find communities as strong as the initial step. By using another algorithm like Leiden, it would be possible to find stronger sub-communities at greater depths within the graph, offering more insights. Additionally, a metric other than Surprise should be investigated for use in future work. While surprise did successfully show community significance, it is sensitive when used on larger graphs like this experiment, as the already existing strong clusters mean that strong communities are certain to exist, which is why the Surprise values for the detected communities were much higher than the random baseline. Algorithms like OSLOM could mitigate this issue, as it preserves the degree from the sub-community its testing means it would be less sensitive in larger graphs. Finally, there could be modifications done to dataset collection and curation. Of course, a larger dataset would always offer more insight. Future experiments could also add a filter to edges between two videos from the same channel to emphasize cross-channel communities. Additionally, the dataset also contained many videos that were split into parts (eg. Part 1, Part 2, ...), which would intuitively have a strong shared community. This is what was responsible for some of the strong communities found in the experiments of this project, so future work could also add a filter to reduce the importance of connections between videos part of the same series. Finally, the genre labels can also be refined. Many videos were in an area of overlap between multiple genres. If more specific labels like "horror gaming" were used, it would offer much more insight into the specific communities formed around videos. Overall, this project yielded promising results which establish a strong foundation for future work

References

- [1] Fortunato, S. (2010) Community detection in graphs. *Physics Reports* **486**(3–5):75–174.
- [2] Traag, V.A., Waltman, L. & van Eck, N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**:5233. <https://doi.org/10.1038/s41598-019-41695-z>
- [3] Gargi, U., Lu, W., Mirrokni, V. & Yoon, S. (2011) Large-scale community detection on YouTube for topic discovery and exploration. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 11–20.
- [4] R. Aldecoa and I. Marín, "Surprise maximization reveals the community structure of complex networks," *Scientific Reports*, vol. 3, no. 1060, 2013. :contentReference[oaicite:0]index=0