**Master Thesis**

Analysis of Niche Communities in ZDF Mediathek

Ioannis Konstantakopoulos 0960047

Contents

[Chapter 1 4](#_Toc71103132)

[Introduction 4](#_Toc71103133)

[1.1. Motivation and Context 4](#_Toc71103134)

[1.2. Literature Review 4](#_Toc71103135)

[1.3. Research Question 5](#_Toc71103136)

[1.3.1. Methods 6](#_Toc71103137)

[1.4. Limitations 8](#_Toc71103138)

[Chapter 2 10](#_Toc71103139)

[Data 10](#_Toc71103140)

[2.1. Selected Data Exploration Results 10](#_Toc71103141)

[2.2. Data Preparation for Analysis including motivation (integration, missing data analysis, etc..) 10](#_Toc71103142)

[2.3. Ethical and legal considerations of the data 10](#_Toc71103143)

[Chapter 3 11](#_Toc71103144)

[Methods 11](#_Toc71103145)

[3.1. Translation of the research question to a data science question 11](#_Toc71103146)

[3.2. Motivated selection of method(s) for analysis 11](#_Toc71103147)

[3.3. Motivated settings for selected method(s) 11](#_Toc71103148)

[Chapter 4 12](#_Toc71103149)

[Results 12](#_Toc71103150)

[4.1. Selected Analysis results 12](#_Toc71103151)

[Chapter 5 13](#_Toc71103152)

[Conclusion and Discussion 13](#_Toc71103153)

[5.1. Answering the data science question 13](#_Toc71103154)

[5.2. Answering the research question 13](#_Toc71103155)

[5.3. Describing implications for the proper domain setting 13](#_Toc71103156)

[5.4. Discuss ethical implications and consideration 13](#_Toc71103157)

[Chapter 6 14](#_Toc71103158)

[Appendix 14](#_Toc71103159)

[6.1. Annotated scripts of analyses and method settings 14](#_Toc71103160)

[6.2. Full data exploration results 14](#_Toc71103161)

[6.3. Full analysis results 14](#_Toc71103162)

[Chapter 7 15](#_Toc71103163)

[References 15](#_Toc71103164)

Chapter 1

1. Introduction
   1. Motivation and Context

ZDF is one of the biggest publicly funded German broadcasters, having monthly fees subscriptions from all Germans. In order to serve the public user best, good content needs to be provided while keeping ZDF ~~and it’s~~ profitable. The definition of what content is actually good however, is strongly dependent on how it is presented to each user. Each group of similar users, or within similar tastes can be defined as taste clusters.

Through the creation of new consumer categories such as taste communities, ZDF can also adapt what (Rogers, 2009) defines as post demographic profiling. By exclusively generating insights from users’ behavior, algorithms allegedly are freed from traditional markers of identity. Ultimately, the recommenders intends to deliver a hyper-personalized experience, but little is known about the specific mechanisms it uses.

Taste clusters, aka taste communities, is a term that has previously been mentioned in research (Barret, 2016) and (Adalian, 2018). Taste clusters are similar to group of text, but they are detected by analyzing viewing behavior. Barrett provides one of the most concise examples, describing them as "communities" of titles based on what subscribers want, adding that "Netflix assigns each subscriber three to five of these clusters, weighted by the degree to which each suits their taste."

If taste clusters are intersections of titles, one can imagine taste communities as aggregations of people around particular types of titles in those clusters. None of those ideas are appropriate for SVODs (Subscription Video On Demand), once we have overlaps on these communities. SVODs, on the other hand, can provide tailored recommendations via customized interfaces. But, more significantly, such suggestions tend to please audiences based on previously established criteria. But what kind of criteria? What metric defines a taste community?

To answer these questions, this work aims to investigate **Taste Communities in the ZDF Mediathek** (German Public Broadcast).

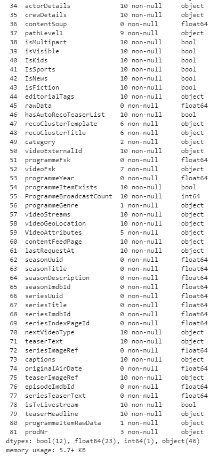
* 1. Literature Review

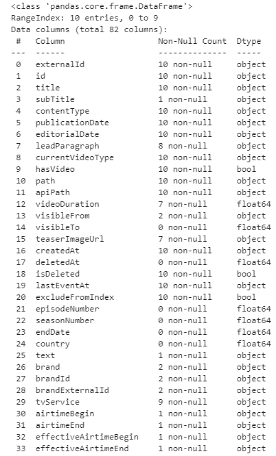
Taste has many different definitions in literature, ranging from (Bordieu, 1984) who defines taste as a disposition, acquired and cultivated through social and cultural capital. Others say taste is also an act embodied through consumption (Bordieu, 1984) (Peterson, 1992). These different definitions constitutes taste and renders it unstable and plural in nature (Bucher, 2012). On the other side, algorithms construct taste following their own computational logics, as illustrated by the Netflix global algorithms (Rodriguez, 2017). Taste is extracted from users data, which is then clustered into categories such as sub-genres and taste communities. The recommender then predicts taste through recommendations, which when consumed, creates more data about individuals’ tastes (Gaw, 2019).

One should not investigate taste only by theoretical importance but historically it has been a mechanism to define and classify individuals and cultures according to dominant social structures (Bordieu, 1984). Works of (Bucher, 2012) (Gillespie, 2016) (Cohn, 2019) discuss in which ways algorithms biases particular tastes and neglect others? They calculated group of users that may normalize presumptions about social identities (i.e labelling such as “hippies”)? Lastly, how it influences individuals’ taste on what (Lash, 2007) terms post-hegemonic power, which refers to the recursive reproduction of social structures within culture? These questions requires a critical understanding on how algorithms can shape our cultural experiences and identities.

Literature suggest that both models of taste and algorithms, involve processes that presuppose and produce categories or groups of “similarities” that organize the world (Bordieu, 1984) (Gillespie, 2016). However, models of taste attach these distinctions on individuals’ social positions, to which algorithms are claimed to be indifferent (Cohn, 2019). How to identify taste in the recommender logics is a focus of this research.

* 1. Research Question

ZDF have provided 180 days of anonymized usage data (online, linear) for the whole ZDF *Mediathek*, within content metadata. ZDF can (maybe) provide anonymized usage data of all of our registered users. The dataset provided until date can be view below (figure1) consists of 82 columns within key/values in some of these fields in a hierarchical structure, the lack of real data (values or samples) limits the analysis further on.



**Figure1**. Overview of ZDF dataset

Given ZDF Data, one can assume it is possible to find taste communities, leading to the main research question: **What taste communities can be identified in the ZDF data**?

This research question leads to further **sub questions**:

* What are taste communities?
* How do I will define them ?
* Which of the found taste communities fits the ZDF problem best?

The data provided will be analyzed applying clustering techniques to find communities, once we don’t have labels and a clear definition of how to model a taste communities. Therefore, unsupervised learning is required (Unsupervised Learning [Wiki], 2021). The **following subsection** will describe methods to answer this research question.

* + 1. Methods

As the goal is to model taste communities within ZDF data, one important aspect to mention is the best method depends heavily on quality/type of the data. Therefore it is needed to test different methods.

Cluster analysis is used in **unsupervised learning** to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships (Unsupervised Learning [Wiki], 2021). It is a machine learning technique that groups the data that has not been labelled (i.e classified or categorized). Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

**Similarity**, is a key aspect on clustering problems it can be between movies or between members, and can be in multiple dimensions such as metadata, ratings, or viewing data. Also need to take into account factors such as context, title popularity, interest, evidence, novelty, diversity, and freshness. Therefore methods such as NLP (Natural Language Processing) will be required to get context for bag of words or TF\_IDF (i.e Title, tags) and be used as input for cluster algorithms.

Another way to detect community structure, that is, the type of communities where nodes are more likely to connect to each other if they belong to the same community, and thus there are more edges within communities than between. This is a very common intuitive definition of communities which is incorporated in many community detection criteria, for example, **modularity** (Zhang, Levina, & Zhu, 2016).

Other inputs for the method would be as **explicit taste** preferences refer to user actions that directly provide feedback to the algorithm about individual tastes. This includes answering a new profile survey, creating titles queues, and rating content. **Implicit taste** preferences, on the other hand, are user signals that indicate consumption behavior within the system. Plays, watch time, searches and navigation are all implicit data. ZDF data seems to have both type of feedbacks.

As the best method depends on the data, this work will use the methods described below and compare the best performance to suggest what fits ZDF’s best.

* + - 1. K-Means (centroid based)

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean  (cluster centers or cluster centroid), serving as a prototype of the cluster. K-means is fast and simple to implement, but it lacks requires a previous knowledge of cluster numbers or “defined by hand” (Roman, 2019).

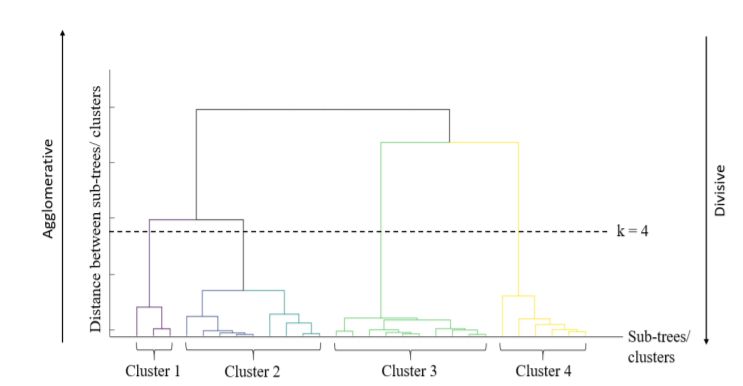
Algorithm XXX K-Means clustering

1. Chose K points as initial centroids
2. Assign each point to its closest centroids to form K clusters
3. Re-evaluate the centroid of each cluster using some distance metric (i.e. Euclidian)
4. Repeat step 2 and 3 until centroids no longer change
   * + 1. Dendrograms (Hierarchical Clustering)

The main advantage of Hierarchical clustering is that one does not need to specify the number of clusters, as they are found during the process (Roman, 2019). In addition, it enables the plotting of dendrograms, they are visualizations of a binary hierarchical clustering. Dendrograms provide an interesting and informative way of visualization. They are specially powerful when the dataset contains real **hierarchical relationships**. On the other hand, they are very sensitive to outliers and, in their presence, the model performance decreases significantly. Additionally, they are very computationally expensive.

Algorithm XXX Agglomerative Hierarchical clustering

1. Create n cluster with observation
2. Compute the proximity matrix
3. Merge the closest two clusters
4. Re-Evaluate the distance between clusters
5. Repeat step 3 and 4 until only one cluster remains



FigureX. A dendrogram representing the clustering technique of hierarchical clustering algorithm

* + - 1. SVD – PCA (Principal Components)

The singular value decomposition (SVD) provides a way to factorize a matrix, into singular vectors and singular values (SVD [WIKI], 2021). The SVD is used widely both in the calculation of other matrix operations, such as matrix inverse, but also as a data reduction method in machine learning. Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Jolliffe & Cadima, 2016). As mentioned in (Liu, 2007), recently used PCA to identify the principal components (PCs) emergent from interest token data collected from user profiles in LiveJournal blogs. Bourdieu (1984) also used statistics akin to PCA to make sense of data from a lifestyle survey of French residents. Key to using SVD is selecting an acceptable rank-k approximation to the original sparse matrix. One should carry out some exploratory analysis to see how much variance can be explained, using the singular values in the diagonal matrix Sigma.

* + - 1. Network Analysis

Definition of a social network: “a graph made up of a set of social actors (such as individuals or organizations), and other social interactions between actors”. Significance of social network analysis: provides a mathematical statements of some social concepts, and make theory more testable and introduced the formal concepts of social network analysis (e.g., density, spam, connectedness, transitivity/cluster…).

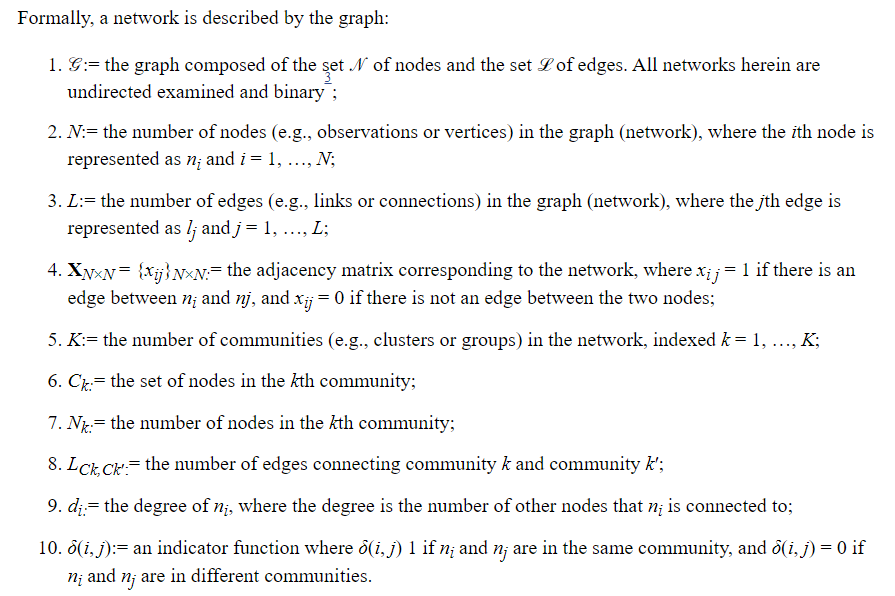
It is composed of **Nodes** are usually representing entities in the network, and can hold self-properties (such as weight, size, position and any other attribute) and network-based properties (such as Degree- number of neighbors or Cluster- a connected component the node belongs to etc.). **Edges**, represent the connections between the nodes, and might hold properties as well (such as weight representing the strength of the connection, direction in case of asymmetric relation or time if applicable). These two basic elements can describe multiple phenomena, such as social connections, virtual routing network, physical electricity networks, roads network, biology relations network and many other relationships, like **taste communities** (Goldenberg, 2019).

These measures in a network can help identify, analyze and define taste communities in a graphical way increasing the explanation power of why user is inserted in specific community, therefore increasing transparency and acceptance and helping content producers to have insights of trends in community level.

Works such as (Zhang, Levina, & Zhu, 2016), (Bedi & Sharma, 2016) and (Vineeth, RamKarthik, Reddy, Surya, & Deepthi, 2020), were used to detect Taste Communities using Network analysis. They used different metrics to detect communities, such as modularity, cluster-size, transitivity, centrality measures and nodes/edges parameter’s.

These works states to evaluate the communities found one would use metrics known as **modularity** to judge the quality of partitions or communities formed. Modularity has been widely accepted and used by researchers to evaluate the goodness of the clusters obtained from the community detection algorithms. Modularity was defined as Σ𝒆𝒊𝒊𝒊−𝒂𝒊𝟐 , where 𝒆𝒊𝒊 denotes fraction of the edges that connect vertices in community 𝒊, 𝒆𝒊𝒋 denotes fraction of the edges connecting vertices in two different communities 𝒊 and 𝒋 while 𝒂𝒊=Σ𝒆𝒊𝒋𝒋 is the fraction of edges that connect to vertices in community 𝒊. High values of modularity means indicates a network with strong community structure.

As defined in (Hoffman, Steinley, Gates, Prinstein, & Bruscoc, 2018),a network is described by the graph:



* 1. Limitations

This work doesn’t focus on improving accuracy of recommendation system neither business metrics. The works depends heavily on the quality of data provided by ZDF.

**Diversity** is a value that must be considered in cluster communities to avoid echo chambers and allow users access to broad content and different ideas. The measure of how well the taste is identified should be assessed through visual check and business validation (ZDF members) and metrics, such as: total variance, distances from communities or the number of members in each cluster.

To evaluate cluster will be used **The Silhouette Coefficient**, which is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a,b). Further discussion on Methods Chapter.

Chapter 2

1. Data
   1. Selected Data Exploration Results
   2. Data Preparation for Analysis including motivation (integration, missing data analysis, etc..)
   3. Ethical and legal considerations of the data

Chapter 3

1. Methods
   1. Translation of the research question to a data science question

As states on [link](https://www.researchgate.net/publication/295549761_An_Improved_Collaborative_Filtering_Recommendation_Algorithm_Based_on_Community_Detection_in_Social_Networks), Recommender Systems use data of similar users or similar items to generate recommendations. This is analogous to the identification of groups, or similar nodes in a graph. Hence community detection holds an immense potential for recommendation algorithms. Cao et al114 have used a community detection based approach to improve the traditional collaborative filtering process of Recommender Systems. The process starts with the mapping of user-item matrix to user similarity structure. On this matrix, a discrete PSO (particle swarm optimization) algorithm is applied to detect communities. The items are then recommended to the user based on the discovered communities.

* 1. Motivated selection of method(s) for analysis
  2. Motivated settings for selected method(s)

Chapter 4

1. Results
   1. Selected Analysis results

Chapter 5

1. Conclusion and Discussion
   1. Answering the data science question
   2. Answering the research question
   3. Describing implications for the proper domain setting
   4. Discuss ethical implications and consideration

Chapter 6

1. Appendix
   1. Annotated scripts of analyses and method settings
   2. Full data exploration results
   3. Full analysis results

Chapter 7

# **References**

Adalian, J. (2018). *How Netflix swallowed tv industry.* New York: Vulture. Retrieved from https://www.vulture.com/2018/06/how-netflix-swallowed-tv-industry.html

Barret, B. (2016). *Netflix's Grand, Daring, Maybe Crazy Plan to Conquer the World.* Wired. Retrieved from https://www.wired.com/2016/03/netflixs-grand-maybe-crazy-plan-conquer-world/

Bedi, P., & Sharma, C. (2016). COMMUNITY DETECTION IN SOCIAL NETWORKS. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. doi:DOI: 10.1002/widm.1178

Bordieu, P. (1984). Distinction: A social critique of the judgement of taste.

Bourdieu, P. (1992). An Invitation to Reflexive Sociology. *Cambridge: Polity*.

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. . *New media & society, 14(7), 1164-1180.*

Cohn, J. (2019). The Burden of Choice: Recommendations, Subversion, and Algorithmic Culture. . *New Brunswick: Rutgers University Press.*

Dhaenens, F. &. (2019). ‘Press play for pride’: The cultural logics of LGBTQ-themed playlists on Spotify. *New Media and Society, 21(6).* doi:https://doi.org/10.1177/1461444818808094

Elkins, E. (2019). Algorithmic cosmopolitanism: on the global claims of digital entertainment platforms. *Critical Studies in Media Communication, 36(4)*, 376–389. Retrieved from https://doi.org/10.1080/15295036.2019.1630743

Gaw, M. F. (2019). Algorithmic logics of taste: Cultural taste and the Netflix.

Gillespie, T. (2016). # trendingistrending: When algorithms become culture. . *In Algorithmic cultures (pp. 64-87). Abingdon, Oxon: Routledge.*

Gilmore, J. N. (2020). To affinity and beyond: Clicking as communicative gesture on the experimentation platform. *Communication, Culture and Critique, 13(3).* Retrieved from https://doi.org/10.1093/CCC/TCAA005

Goldenberg, D. (2019). Social Network Analysis: From Graph Theory to Applications with Python. *Towards Data Science*. Retrieved from https://towardsdatascience.com/social-network-analysis-from-theory-to-applications-with-python-d12e9a34c2c7

Havens, T. J. (2020). Algorithmic Audience Modeling and the Fate of African American Audiences. *JCMS: Journal of Cinema and Media Studies, 60(1).* Retrieved from https://doi.org/10.1353/cj.2020.0071

Hennion, A. (2004). Pragmatics of taste. *In M. Jacobs & N. W. Hanrahan (Eds.),The Blackwell companion to the sociology of culture (pp. 131-144).*

Hoffman, M., Steinley, D., Gates, K., Prinstein, M., & Bruscoc, M. (2018). Detecting Clusters/Communities in Social Networks. doi:10.1080/00273171.2017.1391682

Jaffali, S. J. (2020). Like-tasted user groups to predict ratings in recommender systems. *Social Network Analysis and Mining, 10(1).* Retrieved from https://doi.org/10.1007/s13278-020-00643-w

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci. 374(2065): 20150202.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/

Kunaver, M. &. (2017). Diversity in recommender systems – A survey. *Knowledge-Based Systems 123*. Retrieved from https://doi.org/10.1016/j.knosys.2017.02.009

Lash, S. (2007). Power after Hegemony: Cultural Studies in Mutation? . *Theory, Culture & Society, 24(3), 55-78.*

Liu, H. (2007). Social network profiles as taste performances. *Journal of Computer-Mediated Communication, 13(1).* Retrieved from https://doi.org/10.1111/j.1083-6101.2007.00395.x

Lotz, A. (2021). Unpopularity and cultural power in the age of Netflix: New questions for cultural studies’ approaches to television texts. *European Journal of Cultural Studies.* Retrieved from https://doi.org/10.1177/1367549421994578

Paßmann, J. &. (2020). Liking as taste making: Social media practices as generators of aesthetic valuation and distinction. *New Media and Society.* Retrieved from https://doi.org/10.1177/1461444820939458

Peterson, R. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. . *Poetics, 21(4)*, 243-258.

Rodriguez, A. (2017). *Netflix divides its 93 million users around the world into 1,300 “taste communities”.* Quartz. Retrieved from https://qz.com/939195/netflix-nflx-divides-its-93-million-users-around-the-world-not-by-geography-but-into-1300-taste-communities/

Rogers, R. (2009). Post-demographic machines. *Walled Garden, 38.*

Roman, V. (2019). "Unsupervised Machine Learning: Clustering Analysis". *Medium*. Retrieved from https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e

Seaver, N. (2021). Seeing like an infrastructure: avidity and difference in algorithmic recommendation. *Cultural Studies.* Retrieved from https://doi.org/10.1080/09502386.2021.1895248

SVD [WIKI]. (2021). Retrieved from https://en.wikipedia.org/wiki/Singular\_value\_decomposition

(2021). *Unsupervised Learning [Wiki].* Retrieved from https://en.wikipedia.org/wiki/Unsupervised\_learning

Vineeth, M. S., RamKarthik, K., Reddy, M., Surya, N., & Deepthi, L. (2020). Comparative Analysis of Graph Clustering Algorithms for Detecting Communities in Social Networks. In *Ambient Communications and Computer Systems pp 15-24.* Retrieved from https://link.springer.com/chapter/10.1007%2F978-981-15-1518-7\_2

Zhang, Y., Levina, E., & Zhu, J. (2016). Community detection in networks with. *Electronic Journal of Statistics*, 3153–3178.