DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2018

Temporally Stable Clusters of Movie Series

A Machine Learning Approach to Content Segmentation

PATRICK MILLER

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Temporally Stable Clusters of Movie Series - A Machine Learning Approach to Content Segmentation

PATRICK MILLER

Master in Computer Science
Date: August 01, 2018
Supervisor: Pawel Herman
Examiner: Örjan Ekeberg
School of Electrical Engineering and Computer Science
Host company: HBO Europe

Abstract

Clustering techniques have been shown to provide insight in various domains and applications. *Adaptive evolutionary spectral clustering* is a state-of-the-art method to obtain temporally stable clustering results from time-stamped data. This thesis explores the use of adaptive evolutionary spectral clustering to perform a clustering of film series into groups based on video streaming data. The developed method successfully performs a stable segmentation of film series into groups and introduces a number of extensions to the framework within the context of video on demand. We find that the implemented method allows for reasoning about clusters from an evolutionary perspective and that the state-of-the-art can be extended to introduce a dynamic number of clusters without negatively impacting the stability of properties of clusters.

Sammanfattning

Klustringsalgoritmer har använts och implementerats inom ramen för en mängd olika sammanhang. *Adaptive evolutionary spectral clustering* är en klustringmetod som används för att uppnå kluster som är stabila över tid genom tidsstämplad data. Detta examensarbete implementerar och utforskar användning och påbyggnad av metoden för att segmentera filmserier i grupper genom *video streaming data*. Den utvecklade modellen segmenterar filmserier i stabila grupper och introducerar en mängd utvecklingar inom *video on demand* - domänen. Vi finner att den implementerade metoden möjliggör klusteranalys från ett evolutionärt perspektiv och att metoden kan utvecklas genom att introducera ett dynamiskt antal kluster utan att negativt påverka stabiliteten eller egenskaperna hos kluster.

Acknowledgements

I would like to thank my supervisors Mathias Althin and Fredrik Ragnar at HBO Nordic for their guidance, patience and time. Their knowledge, support and encouragement have been vital for the outcome of this project. Mathias, I would like to extend a special appreciation for the energy you have invested in this process. Your balance of informal and formal conversation has kept the thesis supervision very pleasant. I further wish to thank HBO Nordic for giving me the opportunity to explore my ideas and pursue my passion within analytics and digital media. Finally, I would like to acknowledge and thank my formal supervisor Pawel Herman for encouraging my ideation process and steering me in the right direction.

Contents

Chapter 1

Introduction

1.1 Background

The adoption of machine learning techniques by organizations has proceeded at a rapid pace in recent years. Machine learning plays an important role for systems in industries such as e-commerce, entertainment and insurance [1, 2, 3]. One particularly interesting topic is the study of user groups, which provides a foundation for efforts within marketing and customer relationship management [4, 5]. Correspondingly, the study of content groups helps improve the user experience and provides business value. Personalization systems within digital media leverage machine learning to tailor services to individual users.

Today, consumers of entertainment face a large number of easily accessible choices with regards to the services and content they consume. Similarly, providers of digital media have the opportunity to curate and develop their offering according to insights gained from behavioral data. However, user behavior is not static but evolves over time through trends and short term fluctuations in consumption patterns. Discovery of such patterns can be useful in forming marketing strategies [6]. Almost all real data sets connected to user behavior evolve over time, both in structure and in properties.

The main purpose of this work is to extend state-of-the-art methods for clustering and apply them in the video on demand (VOD) domain in order to provide foundation for segmentation efforts. VOD services allow users to have control over what they watch, when they watch and how they watch [7]. Such services have gained significant traction in recent years and provide opportunity for research into user behavior.

Though traditional clustering methods have been applied extensively and provide a good tool for performing different types of segmentation, they often fail to capture nuances in user behavioral data. Clusters often vary over time due to the noise in observed data and require additional work to reflect an accurate representation of groups. These characteristics contribute negatively to the interpretability of the results, and may therefore render a machine learning application less likely to be considered by a business manager [2].

Furthermore, many application rely on general variables since they are easy to implement [8]. However, the advantage of using behavioral data, as in this thesis, is that the results are more actionable within an organizational context. This is because user behavior can explicitly be leveraged to produce recommendations or predictions and thus the researcher may avoid the collection and processing of data that can be considered sensitive or that the user may be reluctant to endorse [9, 10]. Also, since objects within a group will reflect user behavior to a greater extent than any method reliant on general variables, marketing efforts could elicit a more homogeneous response. Lastly, most traditional methods provide only a static representation of groups. Many applications may benefit from studying evolutionary or structural aspects of clusters.

1.2 Research question

This thesis aims to perform a clustering of film series into groups that exhibit stable cluster membership assignments over time even though the segmentation base expresses dynamic properties over time. The clustering is based on video streaming data and may thereafter be interpreted based on remaining data about the users and series. In particular, within the scope of this thesis we investigate whether adaptive evolutionary spectral clustering can be applied in the VOD domain to segment film series based on user behavioral data. We investigate whether adaptive evolutionary spectral clustering produces results with properties that facilitate reasoning about the content mix, such as compactness and separation of clusters, or temporal stability of cluster membership assignments. Lastly, we explore a potential extension to the framework that may render the clustering results better suited for facilitating content segmentation within the VOD domain. This work boils down to addressing two major research questions:

- Can a segmentation of the content catalog into groups of film series that produces stable cluster membership assignments over time be obtained by applying adaptive evolutionary spectral clustering on video streaming data?
- Is it possible to enhance adaptive evolutionary spectral clustering to increase the compactness and separation of clusters without reducing the stability of the results?

1.3 Scope

The primary focus of this thesis is to explore the temporal stability of clustering results. We do not put emphasis on explicitly displaying or interpreting the clustering results. Although this analysis is built around user behavioral data, it will not attempt to formalize, model or explore behavioral cohorts. Further, the thesis is limited by certain confidentiality aspects. In the case that clustering results are interpreted in this thesis, it is through quantitative metrics.

1.4 Contributions

The main contributions of this thesis consist in:

- Applying adaptive evolutionary spectral clustering to the VOD domain
- Incorporating automated selection of the number of centroids for adaptive evolutionary spectral clustering
- Incorporating decay of latent similarities and penalty of specific entries for adaptive evolutionary spectral clustering

Chapter 2

Background

2.1 Video on Demand

The defining characteristic of a VOD service is the possibility for users to consume content whenever they choose. Typically, the service provides films and series from a catalog of content. A series is generally divided into one or several seasons, spanning a number of episodes. While each episode may vary in length, most episodes are typically under one hour in length. New content becomes available on a platform in varying intervals of time.

Different series tend to follow varying patterns in terms of how often they are released, and in what manner. For example, some series tend to have their episodes released in an incremental manner, often on a weekly basis. Other series may have an entire season released in bulk. VOD services typically offer a mix of original programming and licensed content. Due to asset rights constraints, licensed content may have several seasons released in bulk once the content is acquired. Licensing deals may constitute a significant portion of content on any VOD platform. It is therefore interesting to study research topics that can facilitate content purchase decisions

Many services recommend films and series to their users by email, smartphone notifications or through designated space on the platform itself. The recommendations are, as of recent years, often curated to each individual user. Machine learning techniques have gained significant traction in making this effort feasible. These techniques study behavioral data in order to produce recommendations. Grouping together complementary products can be a prominant technique in the effort to reduce customer churn [4]. The temporal properties of behavioral data express dynamic properties. For this research, there are three types of major changes in the data:

- New users and series joining the service or existing users and series leaving
- Similarities appearing or disappearing
- Changes in the number of content groups

These dynamics pose significant implications to the analysis of users and content. In turn, this also provides a natural opportunity for research. A fundamental assumption for this research is that the structural properties of content groups may change over time as a result of changes in the aggregate user behavior or service and catalog.

2.2 Host company

HBO, the host company, offers premium content through a wide catalog of film series. The service is provided to a range of countries within and outside of Europe. HBO Nordic also offers Toonix Kids content in the Nordics. The catalog consists of series produced by several different production companies, including original HBO programming. A segmentation of the content catalog into groups of series is interesting in order to facilitate reasoning about the content mix. In particular, a segmentation that is stable even though the segmentation basis, i.e. users and content, is dynamic, is of particular interest.

2.3 Market segmentation

Market segmentation is the process of dividing items or populations into different groups, and then targeting marketing efforts with this information. As an example, machine learning has become a popular approach to conduct segmentation of users [11]. Analogously, previous work at HBO Nordic has shown that it is possible to segment the catalog into groups of series based on machine learning and user streaming data.

2.4 Cluster analysis

Cluster analysis is a collection of methods to group objects or data points into groups. The general problems of clustering are concerned with selecting objects which are closer to each other than they are to the rest of the objects. The result from a clustering algorithm is a specific partitioning of the objects into a number of disjoint sets. Cluster analysis is a fundamental approach to unsupervised learning which gives information about the underlying structure of the examined data. Unsupervised learning studies data without explicit target outputs or environmental evaluations associated with each input [12].

Clustering results can be produced and interpreted in numerous ways. The method of interpretation is, ideally, consistent with the objective behind performing the clustering. It is often guided by a business decision or information need.

2.4.1 K-means

The K-means algorithm is a centroid-based method used to divide points into clusters such that the within-clusters sum of squares is minimized. Given K initial cluster centers, the following steps are repeated:

1. Each point is assigned to its closest cluster center
2. The new cluster centers are calculated as the mean of each cluster

This process is repeated until cluster assignments no longer change. The algorithm is then considered to have converged to a solution. K-means is sensitive to various effects, and therefore does not guarantee an optimal solution. In particular, effects resulting from initial placement of centroids as well as choice of distance metric have been studied.

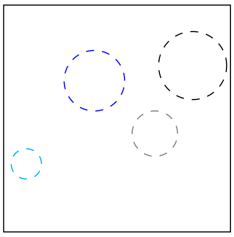

Figure 2.1: A two dimensional illustration of four groups.

2.4.2 Spectral clustering

Spectral clustering is a collection of techniques that use the eigenvectors of the graph Laplacian matrix to obtain a partition of the samples [13]. The graph Laplacian matrix is constructed from the affinity graph between sample points. The affinity graph is also referred to as the adjacency graph. Spectral clustering performs a dimensionality reduction before finding an optimal partition of the samples. As opposed to many other techniques, spectral clustering does not make strong assumptions on the form of the clusters. The researcher may or may not want to assume a particular a priori structure in the relations of the groups in question. This implies that spectral clustering can solve very general problems. Lastly, spectral clustering involves solving a linear problem once a similarity graph is chosen. This choice, however, is not a trivial problem. Below are the main general steps involved in performing spectral clustering:

1. Construct a similarity graph
2. Compute the Laplacian
3. Find the eigenvectors of the Laplacian
4. Compute k-means on the matrix with k eigenvectors as columns

For spectral clustering, the graph Laplacian is typically defined as follows [14]:

$$\text{Unnormalized: } L = D - W \tag{2.1}$$

$$\text{Normalized: } L = I - D^{-1/2} W D^{-1/2} \tag{2.2}$$

Where D is the degree matrix and can equivalently be expressed as $d_i = \sum_{j=1}^{n} w_{ij}$. W is the similarity graph's weighted adjacency matrix which can be formulated as $W(A, B) := \sum_{i \in A, j \in B} w_{ij}$.

Spectral clustering involves finding an optimal solution to a graph partitioning problem [15]. That is, finding a partition of the graph such that the edges within a group have high weights and the edges between groups have low weights. Similarly to the choice of similarity graph, determining the appropriate graph partitioning method is a nontrivial problem. Ideally, a precise criterion for a good partition of the graph should be formulated [16].

One must take care not to formulate a criteria such that it captures local properties of the graph but fails to extract global impressions of the data. In an organizational context, this notion is tied to the interpretability or actionability of the clustering results. For example, a cluster consisting of one lone series can be considered either desirable or unacceptable. Similarly, one large cluster capturing all the series in the catalog could possibly be considered a redundant solution. Using graph notation, we can express the different approaches for separating points into different groups according to their similarities.

For a given subset A, $\mid A \mid$ denotes the number of vertices and vol(A) denotes the weights of its edges. Either of these two measures can describe the size of A. Finally, $\bar{A}$ denotes the complement of A. The following are the most commonly studied graph cuts [13, 16, 17]:

$$mincut(A_1, ...A_k) := \frac{1}{2}\sum_{i=1}^{k} W(A_i, \bar{A}_i) \tag{2.3}$$

$$ratiocut(A_1, ...A_k) := \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \bar{A}_i)}{\mid A_i \mid} \tag{2.4}$$

$$normcut(A_1, ...A_k) := \frac{1}{2}\sum_{i=1}^{k} \frac{W(A_i, \bar{A}_i)}{vol(A_i)} \tag{2.5}$$

In the context of spectral clustering, Von Luxburg [13] presents the following relaxed versions of the solutions to the problem of finding the corresponding optimal cuts: average association, ratio cut and normalized cut.

$$AA(Z) = \max_{Z \in R^{n \times k}} tr(Z^T W Z) \text{ subject to } Z^T Z = I \tag{2.6}$$

$$RC(Z) = \min_{Z \in R^{n \times k}} tr(Z^T L Z) \text{ subject to } Z^T Z = I \tag{2.7}$$

$$NC(Z) = \min_{Z \in R^{n \times k}} tr(Z^T \mathcal{L} Z) \text{ subject to } Z^T Z = I \tag{2.8}$$

Here, L denotes the unnormalized Laplacian matrix. $\mathcal{L}$ denotes the normalized Laplacian matrix. Each variant aims to optimize a slightly different objective. Thus, the resulting partitioning from each respective cut have different properties. Figure 2.2 is an illustration of a graph partitioning in two dimensions:

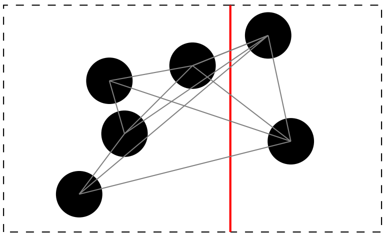

Figure 2.2: A graph partitioning into two partitions. The lines between the objects represent the weights

2.4.3 Kernel method

A similarity matrix is typically denoted by $S = (S_{ij})_{i,j=1...n}$ and is constructed by some similarity function which takes data points as input and outputs pairwise similarities between arbitrary objects. Entries in the similarity matrix range from zero to one, where zero is least similar and one is identical. Spectral clustering makes use of the kernel method, that is, uses the Gaussian kernel as similarity function. The Gaussian radial basis function maps the data to an infinite dimensional feature space. Essentially, the kernel trick allows the application of linear algorithms on non-linear data only when they are cast in terms of dot products. The choice of ρ is up to the researcher.

$$s(\vec{x}_i, \vec{x}_j) = \exp \frac{\| \vec{x}_i - \vec{x}_j \|^2}{2\rho^2} \tag{2.9}$$

2.4.4 Optimal number of clusters

It is often up to the researcher to choose the number of clusters. The choice may be derived from a business requirement or arbitrarily chosen. However, there are various methods for selecting an optimal number of clusters for a given dataset. The criterion of optimality varies depending on the method applied. Two common methods are the eigengap heuristic [13] and the average silhouette width [18].

The eigengap heuristic is a method that is particularly commonly applied to spectral clustering. This heuristic is applied a priori to performing a clustering on the data and measures the difference between successive eigenvalues to find an optimal number of clusters. In contrast, the average silhoeutte width criterion is applied on the clustering results of the data. It is particularly useful when one is seeking compact and clearly separated clusters. In the case of this thesis, this imposition is fundamental to the evaluation of clustering results and is therefore our prefered criterion of optimality. The silhouette can be expressed in one equation and only requires the partitioning of objects as well as the proximites between all objects.

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.10}$$

$a(i)$ is the average dissimilarity of any object i to all other objects of a. $b(i)$ is the minimum of all average dissimilarities of i to any other cluster.

2.4.5 Evolutionary clustering

Evolutionary clustering was first introduced by the work of Chakrabarti et al. [19]. Evolutionary clustering adds a temporal smoothness penalty to a static clustering algorithm. In the case of an offline algorithm, both historic and future data may be used. The total objective function follows, where the parameter alpha controls the amount of temporal smoothing:

$$C_{total} = \alpha\, C_{temporal} + (1 - \alpha)\, C_{snapshot} \tag{2.11}$$

The snapshot cost only measures the snapshot quality of the current clustering result with respect to the current data features. The temporal cost measures goodness-of-fit of the current clustering result with respect to either historic data features or historic clustering results.

Chi et al. [17] proposed two frameworks for evaluating evolutionary spectral clustering. Both frameworks incorporate the temporal smoothness into the overall clustering quality. In the first framework, the current partition is applied to historic data. The resulting cluster quality determines the temporal cost. In the second framework, the current partition is compared with the historic partition. The resulting difference determines the temporal cost.

2.4.6 Adaptive estimation of alpha

Xu et al. [20] extended the work of Chakrabarti et al. [19] to include a model for automatically selecting the forgetting factor. The choice of an automatically selected forgetting factor is a bias-variance trade off. W^t represents variance because it is the observed scene. W^{t-1} represents bias because it is historical data. A further explanation of the forgetting factor is found in section 3.5 and appendix A.

$$\alpha W^{t-1} + (1 - \alpha) W^t \tag{2.12}$$

2.5 Regularization

The use of regularization introduces statistical prior knowledge or parsimony to penalize an outcome for deviations from expected behavior. In particular, it is used to solve an ill-posed problem or to prevent overfitting. The simplest form of regularization is to use more data. The aim of regularization within this thesis work is to mitigate undesirable effects in the cluster results as a consequence of the temporal properties of the data domain. In particular, this refers to effects that can be argued to diminish the interpretability of results or introduce a temporary bias towards certain clustering properties that may not be representative long-term.

For example, two clusters may iteratively merge and separate as new data is fed into the model. The parameter α described in equation 2.12 is the shrinkage intensity which choice provides an answer to the question of how much historical structure is imposed on the model. An automated selection of α removes the ambiguity in the researchers choice of value. On any given data set and for any given purpose, there may be a different choice in shrinkage intensity which can be considered optimal.

2.6 Related Work

2.6.1 Cluster stability and temporal aspects in a dynamic segmentation basis setting

While the concept of clustering has been studied extensively in a wide range of areas and applications including the film entertainment industry [21], the study of temporal properties of clustering is attracting attention due to increased availability of streaming data [22, 23, 24]. Prior work on temporal aspects of clustering have been published [25, 26, 27, 28, 29]. This thesis aims to focus on exploring an application for the offline setting.

2.6.2 Clustering with supervised wrapper for evaluation

In a broad sense, the notion of stability of clusters has been discussed with a particular focus on exploring classification accuracy in a supervised setting resulting from choice of initial placement of centroids or distance metric [30]. Von Luxburg [30] stresses the conclusion that stability, in the context of k-means, can discriminate between different values of the number of centroids in the sense that the values that lead to stable clustering results have desirable properties. Stability can be considered an important property of a clustering solution, particularly with noisy data [31, 32]. In this thesis, it is presumed that a stable clustering produces greater confidence in segmentation results, and is therefore more likely to be considered by the enterprise. Stability is a major issue in data-driven market segmentation [33]. Wagner et al. [34] explore different distance metrics and introduce different measures for comparing clusterings that have been presented in other research. In particular, an attempt to formalize such measures is presented.

2.6.3 Recommendation systems

In the setting of personalization, Zhang et al. [35] explore neighbourhood and model-based approaches to recommendation systems. The stability of recommendations or predictions are explored by feeding different recommendation algorithms with data that is in line with the current predictions, and examining the implications to the recommendation output of the algorithms. Their research is a good starting point for examining implications from a dynamic segmentation basis. In particular, their research studies the stability under continuously incoming ratings to the system. The authors find that the output of neighbourhood based approaches are more sensitive to new input than model-based approaches in terms of producing consistent recommendations. Other work within the context of personalization and digital media [36] explores:

- implications of implicit feedback
- user features as a function of time, such as user biases and user preferences

Gower [36] finds that the addition of time dependence terms in the model can improve recommendation accuracy. The concept of time dependence terms is of interest to this thesis.

In a similar sense, Singleton et al. [37] study UK population census data and explore reassignment of geodemographic classes over time as a result of changes in values of census variables for given locations. While this study is applied in a different context than the one that is studied in this thesis, it provides valuable insights into how to possibly reason about cluster reassignments and a dynamic segmentation basis.

2.6.4 Temporal smoothing

In practice, there is usually some prior knowledge available for many clustering problems, such as cluster assignment of some data instances [38]. Ji et al. [38] present a semi-supervised model which incorporates prior knowledge in the clustering process. This method enables users to control the clustering process to either achieve desired cluster structures or accurate cluster results [38].

The notion of incorporating temporal constraints in a clustering model is of high interest to this thesis, and, in particular, can be considered a way of incorporating the user's supervision into the unsupervised clustering process [38]. Most notably, Chakrabarti et al. [19] explore the problem of processing time-stamped data to produce a sequence of clusterings in the online setting. The authors present the trade off between maintaining a consistent clustering over time against obtaining an accurate representation of the current data. Chakrabarti et al. [19] present the favorability of evolutionary clustering due to consistency, noise removal, smoothing and cluster correspondence. The k-means and agglomerative hierarchical clustering algorithms are evaluated within this setting. In particular, the setting of the study is in which today's data must be clustered before tomorrow's data is available. However, Chakrabarti et al. [19] argue that it can be used in a setting where clustering is done retroactively. Interpretability across time is an important consideration, this is indeed the context of this thesis work.

Chi et al. [39, 17] build upon the work of [19] and extend it to spectral clustering. The authors focus on the case when the similarities among existing data points vary over time and propose two different frameworks for measuring temporal cost. In particular, Chi et al. [39, 17] refine the notion of temporal smoothness and propose two frameworks that handle the issues of variation in cluster numbers and insertion or removal of nodes over time. Their PCM (named for *preserving cluster membership*) framework that incorporates all historical data using different weights is an interesting study for this thesis work [39, 17].

In the context of this thesis, similarities among different clusters and film series vary over time due to a number of reasons. For example, the pairwise similarity between two series that are released at similar points in time may temporarily be higher than during the long term. Insertion or removal of nodes corresponds to the film series dropping in and out of the catalog over time due to various reasons such as distribution rights. The handling of inactive and emerging nodes is of particular interest to this thesis. Chi et al. [39, 17] handle the dynamic property of time-stamped data by adjusting the feature vectors when computing the overall cost of a clustering at any given time.

The main topic of interest for this thesis is producing a non-trivial smooth evolution of cluster results. Although the study of the evolution of clusters itself may be of high value to an organization, it is not the main purpose of this study. Wang et al. [40] study data points that move, disappear and emerge within a clustering context. The aim of their research is to mine a smooth evolution of the clusters in order to provide an aggregated view of the numerous individual behaviors of clusters [40]. The authors propose a method which learns a hidden semi-Markov model. Their research provides an interesting perspective to this thesis work. While Chakrabarti et al. [19] "optimize" each individual clustering performed, which can be considered appropriate in the online setting, another perspective could potentially be one that is more similar to the work of Wang et al. [40]. That is, obtaining a temporally stable aggregate sequence of clusterings. This can be considered particularly interesting in the offline setting since it takes advantage of the entire history of data to achieve the overarching goal of stability. Chi et al. [17] similarly propose a total cost in the evolutionary spectral clustering setting.

2.6.5 Communities and networks

There is additional research on studying the evolution of clusters, most notably within analysis of communities and social networks [15, 41, 27, 26, 42, 28, 29]. Lin et al. propose a framework to generate evolutions of communities that are regularized by the temporal smoothness of evolutions [27]. Soft assignments to clusters are explored since it is argued to be an important imposition in community structures. In a similar manner, soft membership assignment of users to clusters can be considered appropriate for a segmentation based on behavioral analysis. However, in the case of this thesis, it is series rather than users that are grouped. The link between behavior and groups is thereby more ambiguous, since it is the objects rather than agents that are being studied. This thesis will leave analysis of soft membership assignment for future research, and instead focus on hard membership.

Tang et al. [26] study a similar setting, applying a spectral clustering method to analysis of community evolution. Similarly to Wang et al. [40], Tang et al. discuss how to handle dynamics in time-stamped data, and explore a method to obtain a temporally-smooth sequence of clusterings. Lin et al. [28] propose a direct temporal regularization on community membership, rather than indirect. In contrast to evolutionary spectral clustering, which learns parameters corresponding to eigenvectors and eigenvalues, Lin et al. find a probabilistic generative model that learns a parameter set X of community structure without requiring partition-matching methods.

Evolutionary methods such as the ones described above generally incorporate two cost factors:

- For the current representation of the data

- For measuring how well this representation is relative to history

Xu et al. [43] discuss the trade-off between smoothness of clustering results over time and a lag in detecting changes in clusters. The authors propose an automated method of choosing the weights between current and historical representation costs. The weight is presented as a "forgetting factor". This is done in the setting of evolutionary spectral clustering. Xu et al. [20] further this work and extend it to other static clustering algorithms. The main contribution of their research is providing a method to automate the process of heuristically selecting the "temporal trade-off", encapsulated by the forgetting factor. A manual selection of an optimal weighting could be considered to be infeasible for a large number of applications or introduce bias. Their AFFECT framework is a significant point of study for this thesis. It has, to the best of our knowledge, not been applied to the VOD domain.

2.6.6 Interpretability and actionability of clustering results

Often, the goal of clustering is to provide a foundation for business decisions concerning products and markets. The aim of such decisions is to satisfy customers with varying desires or needs. Therefore, we believe that a clustering must incorporate information that reflects nuances in the behavior of customers or attributes of a product, depending on the segmentation basis. Young et al. [44] argue that a segmentation on the basis of consumers' desired or sought benefits facilitates product planning, positioning and advertising communication. Similarly, this thesis is focused on facilitating content purchase decisions through behavioral based segmentation of film series. In contrast, rather than benefits desired, our segmentation is based on implicit user feedback as reflected by video streaming data. We believe that analysis of user behavioral data may provide a good foundation for business related decisions since it is monitored continuously and can be tracked while making potential changes to the VOD service. For such an effort to be feasible, the results must arguably be interpretable in the sense that they can be explicitly related to users or that the cluster formations exhibit desirable properties for analysis, such as being well separated from each other or stable with regards to cluster membership assignments as new data is fed into the model.

2.6.7 Behavioral variables as segmentation basis

Within the area of games, several works [45, 46, 47] have explored clustering with user behavioral variables as a segmentation basis. Much of this research has been focused around aggregate data sets. In particular, the focus of the research is on developing and evaluating clusters that in some manner are connected to user behavior such that design elements can be linked to the different clusters of users in order to facilitate game design choices. Drachen et al. [45] proceed by using Self Organizing Maps and several statistical features corresponding to high-level playing characteristics in order to produce clusters. Drachen et al. [45] describe the different clusters characteristics, label them and argue that the feature choice depends on the core mechanics of the game as well as the overall purpose of the analysis in question. Similar work has been done on telemetry data [46].

Notably, Sifa et al. [47] further this research by introducing a temporal aspect by examining behavioral clusters as they evolve during a game. Simplex volume maximization and archetypal analysis are used in their analysis to identify and describe how behavioral clusters evolve during the game. In particular, a high degree of variance in the percentage of behavioral clusters across levels in the game is found. Additionally, the authors express the interest for tracking the flow of individual users between different clusters as they progress through the game. This is indeed an interesting method of interpreting results.

So, for this thesis research, the clusters must ideally reflect nuances in user behavior that are interesting for guiding content purchase decisions. With this in mind, a certain minimum level of behavioral information would ideally be contained in the clustering model. Therefore, this thesis investigates how much such information that can be contained in the model with respect to the manner it affects the temporal stability of produced clusters. The above mentioned research shows the feasibility of such an analysis. Note, however, that much of the above mentioned research also explores the interpretability of clustering results.

One major difference in using census-data [37] in contrast to streaming data, which this thesis will study, is the detailed information of the agents, i.e. users metadata. For instance, the data set used during this thesis work does not contain age, gender or other properties that potentially could be interesting to contain in the feature space of the model.

Chien et al. [25] mention that search engine queries can be grouped according to temporal correlation, thereby reducing the need to understand the query terms at a linguistic level. This could be likened to clustering film series without understanding other than temporal correlations between different series. Arguably, the exploratory value of such an analysis is more limited relative other approaches which have a certain emphasis on interpretability.

This thesis research will embrace the notion that a stable and interpretable segmentation result inherently conveys greater credibility to a domain expert and is therefore more likely to be considered by the organization as a complement to existing methods of segmentation. For instance, this is similar to the notion that a user's trust and perception of personalization competence are key to the acceptance of a recommendation [48, 49, 35].

Chapter 3

Method

3.1 Model architecture

The framework introduced in this thesis will henceforth be referred to as the ANCL *(adaptive number of clusters)* framework. It is an extension of the AFFECT framework by Xu et al. [20]. In figure 3.1, an overview of the architecture is provided. Our main contributions of interest have been highlighted in the figure.

3.2 Data set

The studied data consists of a subset of video streams started by users as well as content and user metadata. Entries in the item-user matrix consist of the number of seconds that a user has watched series from the catalog. Several representations of the entries were studied such as binary or Gaussian such that any values were transformed and set within the range boundaries of 0 and 1.

The user started stream table was partitioned into intervals of 45 days in length in YYYY - MM - DD format. These intervals will henceforth be referred to as time windows. The interval-length was heuristically determined in order to represent a large enough time window that could allow users to watch at least more than one unique series during each interval of time. Then, this time window was varied. The content meta data included in this research is the release date of content. The release date corresponds to the date at which a series became available for users to watch.

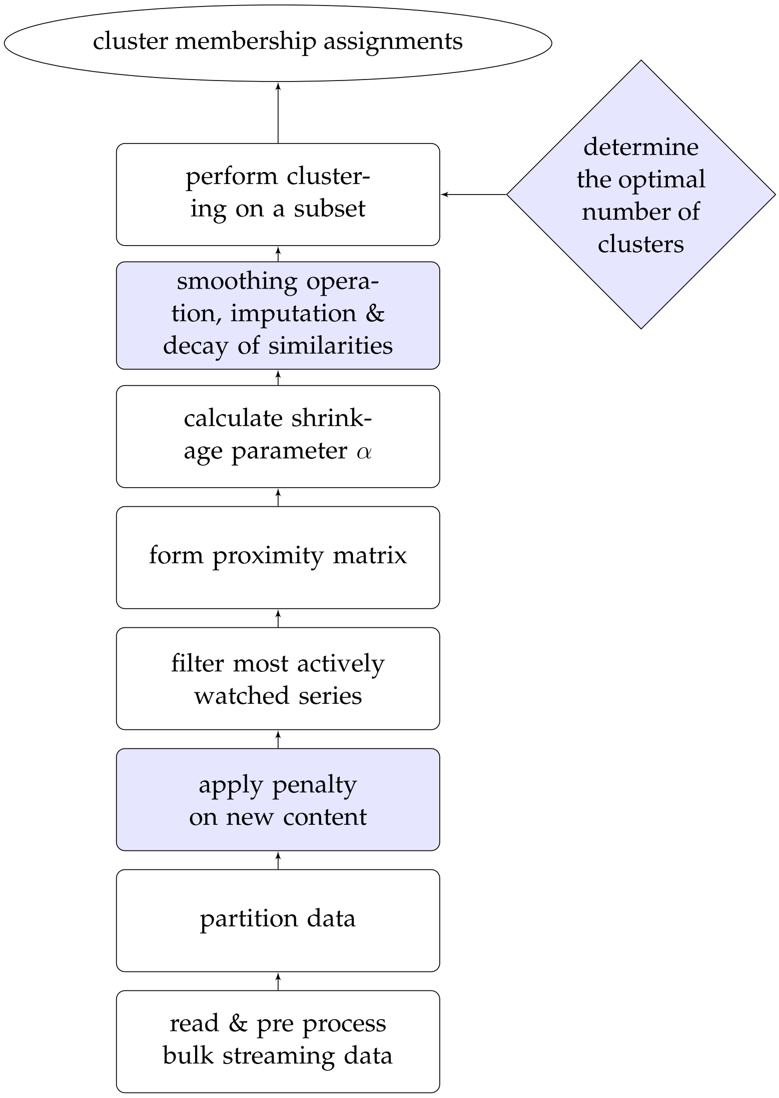
cluster membership assignments
perform clustering on a subset
determine the optimal number of clusters
smoothing operation, imputation & decay of similarities
calculate shrinkage parameter α
form proximity matrix
filter most actively watched series
apply penalty on new content
partition data
read & pre process bulk streaming data

Figure 3.1: Architecture for the ANCL framework.

Started video streams					
Account ID	Date	Time	Seconds	Content	Release date
12345	2018-01-01	17.00	3577	The Wire S01E01	20XX-XX-XX
12345	2018-01-01	17.30	3325	The Wire S01E02	20XX-XX-XX
12346	2018-01-01	17.00	3116	The Sopranos S04E03	20XX-XX-XX

Table 3.1: An illustration of the relevant data entries used.

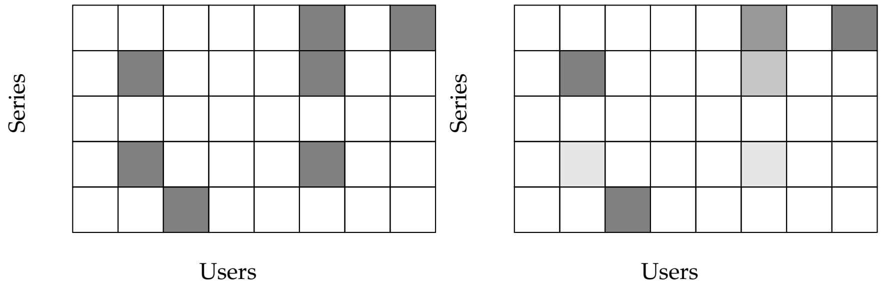

Figure 3.2: The left plot shows a binary representation of the data. The right plot shows a Gaussian representation of the data. Each column is a unique user. Each row is a unique series. The shades in the right plot represent different levels of intensity in consumption for each user and series. The left plot illustrates that levels of consumption are represented as either one (streamed) or zero (not streamed) for the binary representation of the data.

3.3 Preprocessing

3.3.1 Bulk processing

Preprocessing of the data was applied on two distinct levels. First, processing was applied to the bulk streaming data, see figure 3.1. Second, subsets of this data corresponding to each time window were individually processed.

The main concern when processing the data was to preserve behavioral information that could be considered important in contributing to the clustering results or for enabling detailed study of users. Any alteration of entries could be considered to corrupt the segmentation basis or could, for example, imply unintentional removal of potential user-groups. Both users as well as content were considered as potential candidates for containing unrepresentative characteristics.

Users and content were filtered on a total watch-time basis for the entire time period. Also, content was filtered on a user-spread basis, meaning the number of unique users watching each series. Series with a low relative number of unique users were removed. Series with a large number of unique users were not removed for two reasons. First, this would have removed a considerably large amount of behavioral information. Second, these series can be considered valuable since they were watched by a large portion of the user base.

3.3.2 Time window processing

Since the number of time windows was varied for the analysis, each time window was processed individually. Only the most watched series for each time window were included in the algorithm. This method was chosen because it could allow greater flexibility in choice of parameters for running tests without requiring manual processing. The motivation behind including only the most actively viewed series was to avoid incorporating behavioral information that could be considered unrepresentative for the overall user base in any time window. For example, if a new series is released the last day of a 45 day time window or an existing series removed the first day of that same time window, few users may potentially have the opportunity to watch it, and the limited behavioral information captured for that time window and series may be unrepresentative. In practice, this could imply many sparse entries in the proximity matrix, which is discussed later in the report. This method of processing implies that only active users contribute to the clustering results at each time window. Since the evolutionary aspects of clusters was considered valuable for analysis, this is an important detail. Finally, only a subset of the series in the item-user matrix were used in the clustering. All the most actively watched series were used to calculate the proximities, but only a subset of these would contribute directly to the clustering itself. This was decided in order to improve the interpretability and actionability of clustering results.

3.4 Block-approximation of proximities

The goal of the evolutionary clustering framework is to accurately estimate the true affinities between objects at each time t. The true affinity matrix is denoted Ψ, with observed affinities denoted with W and noise denoted by N. Since the observed affinities extracted from the item-user matrix are corrupted by noise, an estimate that is based solely on observed proximities will have a high variance. This is because, within the evolutionary framework, the affinities will only be based on the current observed data at time t.

$$W^t = \Psi^t + N^t, \quad t = 0, 1, 2, ... \tag{3.1}$$

A better estimate for the true proximity matrix, Ψ, is based on historic as well as current data. Thus, the true proximities are estimated at each timestep by the proximity matrix $\hat{\Psi}$ through the following smoothing operation:

$$\hat{\Psi}^t = \alpha \, \hat{\Psi}^{t-1} + (1 - \alpha) \, W^t \tag{3.2}$$

In the above case α is taken to be static, rather than dynamic. As proposed by Xu et al. [43], a suitable form for the target matrix, i.e. the true proximity matrix, is a block matrix. A blockmodel can be considered a hypothesis about a multirelational network which does not necessarily present information pertaining to individual nodes, instead, presenting general features of the network [50]. This structure implies that objects within a cluster have the same true proximities. From this posit it follows that the variances also inherit this same form.

3.4.1 Sample equivalents

The initial calculation of the true proximity matrix is derived by performing a static clustering on the data, then using the result to estimate the posited block form. The cluster membership assignments are used to obtain the sample equivalents of the proximities, means and variances. The initial true proximity matrix is thus obtained from the following equations, where w_{kl} denotes the pairwise similarity between two objects k and l in clusters c and d. N denotes the number of objects in a cluster:

$$\hat{\mathrm{E}}[w_{ii}^t] = \frac{1}{N_c} \sum_{k \in c} w_{kk}^t \tag{3.3}$$

$$\hat{\mathrm{E}}[w_{ij}^t] = \frac{1}{N_c(N_c - 1)} \sum_{k \in c} \sum_{l \in c} w_{kl}^t \tag{3.4}$$

$$\hat{\mathrm{E}}[w_{ij}^t] = \frac{1}{N_c N_d} \sum_{k \in c} \sum_{l \in d} w_{kl}^t \tag{3.5}$$

This implies that all different objects within a cluster have the same proximity between each other. For subsequent time windows, the corresponding terms are estimated using the previous time window clustering result as basis for the calculations. The diagonal entries of all proximity matrices studied in this thesis are set to the value of one. Since the proximity matrix is scaled with a Gaussian transform, the diagonal does not deviate largely from any entries. Therefore, inclusion or exclusion of diagonal entries when computing similarities are both feasible. Otherwise, arbitrarily large values on the diagonal enhance similarity between series that are in close proximity, and, correspondingly, small values on the diagonal increase the similarity of series that are not in close proximity [51].

3.4.2 Modeling missing proximities

After the smoothing operation described in 3.2 there may be several pairwise similarities missing in the resulting similarity matrix. This is because the two operands are derived from disjoint sets in time and the objects may not be consistently represented in the data. This scenario is particularly prominent if the researcher studies large time windows and can be considered a natural implication of studying data where user behavior and content base is dynamic. The computation of proximities is based on information about started streams from the users. Two different series must therefore be available for consumption in the same time window in order to produce an initial similarity measure between them. Unfortunately, this is not always the case. For example, the asset rights for one series can expire before a new series is introduced during the same time window. Then, missing values can either be imputed or ignored. Imputed similarities within this context will henceforth be referred to as synthetic, since they lend themselves to the researchers choice of imputation strategy rather than deriving from observed data.

There are many possible methods to impute missing values into the smoothed proximity matrix $\hat{\Psi}_t$. For example, one could impute the mean or median of the entire matrix. However, as it may have large implications on the clustering result, one must take care in the choice of imputation strategy. In the context of evolutionary clustering, imputed similarities are retained in the next time window as represented by Ψ_{t-1}. Thereby, imputed values will affect the next time windows clustering results, and so forth. This is because the clustering of that time window is performed on $\hat{\Psi}_t$, the smoothing between past and current proximities. Additionally, one must take particular care to study the magnitude of the similarities that are imputed, as these may vary largely between subsequent time windows or long term.

For example, one series may organically exhibit particularly low similarities with the rest of the catalog. If a missing similarity entry for this series is imputed as the mean of the smoothed proximity matrix, this may result in a value that is far larger than any of the observed values. This would render the content corresponding to that entry as the most similar content, although never observed in tandem with this series. In turn, this could form basis for a erroneous business decision. In a worst case scenario, most of the values for this series may be missing. Then, most similarities will be synthetic. Unless intentional, this effect should be avoided as it may hinder accurate analysis of the content mix.

It is here posited that a slightly lower imputed value is preferred over a higher value that trounces the rest of the series top similarities. This is in part because it will reduce the impact of synthetic similarities on the clustering results, but foremost because it is reasonable to assume that the researcher or decisionmaker may be biased towards studying the top similarities, rather than the bottom tail, for any given series.

In this thesis, several methods for the imputation of missing values have been explored. Within the ANCL framework, the appropriate imputation strategy relies on the most probable interpretation in the context of prior world knowledge. This imputation strategy is inspired by Bayesian reasoning and shares common characteristics with block modeling as follows.

1. If one of the objects has a cluster assignment, the proximity is modeled as the mean of proximities between that cluster and the other object.

2. Else, impute an adjusted version of the mean of the entire proximity matrix.

That is, equivalent series are assumed to have the same connection pattern to the same or different neighbors. It is assumed that equivalence is best described by cluster membership for this thesis. Missing values were not ignored because doing so would not contribute to interpretability of proximity calculation results. Imputing zeroes was also considered an inferior approach, as this would imply that two series have zero similarity, which, unless considered likely, is a poor default strategy.

A brute method of imputing the mean of the proximity matrix may lead synthetic similarities to trounce any observed values. Similarly, this presented ANCL framework may, to a certain extent, share this negative characteristic. It may, furthermore, enforce cluster formations, which may or may not be desirable. However, the advantage is that it is based on observed similarities, either explicitly or implicitly. That is, all other similarities are, in the same manner, prioritized based on that they are either observed or averaged out from observed values.

The aim in developing this order of operations was to preserve any observed values that could be tied to the object. Here, it is considered advantageous to use a value from the same cluster rather than taking the mean of the entire proximity matrix. To mitigate the effect of enforcing existing cluster structures without specific intent to do so, all imputed values were also weighted by the mean of the overall proximity matrix.

In practice, mean-imputation should not be done sequentially within a time-step, since that would imply that any missing entry before the first will be skewed by the previous one in the sequence. Finally, a note on filling proximities. All proximities are filled after performing the smoothing operation described in equation 3.2. Thereby, at each timestep, imputed values will contribute to the overall clustering result, yet will not impact the automatic choice of alpha in the current timestep. In other words, only observed similarities at each timestep will impact the selection of α, although both past, present and synthetic similarities will affect the overall clustering result. Overall, as few synthetic values as possible should ideally be present in the model at any given time.

3.5 Adaptive estimation of alpha

The derivation of the optimal shrinkage parameter alpha is included in appendix A. The expression for the optimal choice of the shrinkage parameter alpha [52] follows:

$$(\alpha^t)^* = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{var}(w_{ij}^t)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\{(\hat{\psi}_{ij}^{t-1} - \psi_{ij}^t)^2 + \operatorname{var}(w_{ij}^t)\}} \tag{3.6}$$

The true proximity matrix Ψ_{ij}^t is modeled as a block-matrix, as described earlier in this section. The expression for the smoothed proximity matrix can be extended for an arbitrary choice of alpha:

$$\hat{\Psi}^t = \alpha^t\,\hat{\Psi}^{t-1} + (1-\alpha^t)\,W^t \text{ for } t \geq 1 \tag{3.7}$$

As described earlier in this section, the initial estimate for the true proximity matrix is derived from a static clustering on the current data at time t_0. That is, $\hat{\Psi}^0 = W^0$.

3.6 Regularization

3.6.1 Penalization of new content

As mentioned throughout this thesis, temporal fluctuations as captured in the user data may be problematic for obtaining stable clustering results. Short-term fluctuations are particularly problematic as they often express abnormally skewed effects which is why effort was taken to reduce the impact of them. Viral effects are common in the VOD domain. Furthermore, content mix decisions are, arguably, most often based on long-term prospects. Therefore, methods were prioritized in order to reduce effects from short-term fluctuations in the data. In particular, as mentioned in section 3.3.2, when a new series season is released, short-term behavioral information can be considered unrepresentative from a long-term perspective. Some series may attract many new users to the platform but, in some manner, correlate less with effects on customer loyalty than other series or not live up to initial expectations of performance. A new release here refers to the date it became available on the platform for users to watch.

This thesis introduces a penalty on entries corresponding to content that is less than 30 days old. The method of penalization explicitly adds a static weight on entries in the user data corresponding to series released up to 30 days before the date of consumption.

3.6.2 Temporal decay of latent similarities

Latent similarities are here defined as entries in the affinity matrix that were available in a preceding time window but not in the current. Latent entries may arise from various reasons, as discussed previously. The ANCL framework introduces a decay of latent similarities.

The smoothing expression then becomes:

$$\hat{\Psi}^t = \alpha^t \, \hat{\Omega}^{t-1} + (1 - \alpha^t) \, W^t \text{ for } t \geq 1 \tag{3.8}$$

Where now

$$\hat{\omega}_{ij}^{t-1} = \begin{cases} \delta \hat{\omega}_{ij}^{t-1}, & \text{if } \hat{\omega}_{ij}^{t} \text{ not observed} \\ \hat{\omega}_{ij}^{t-1}, & \text{otherwise} \end{cases}$$

Now, the smoothed proximity matrix is denoted by Ω. The latent decay factor δ regulates the strength of similarities that are observed in previous timestep but not the current. That is, if there is a proximity value in previous time window but not the current, this value is decayed by a static factor, δ.

3.7 Clustering on a subset of all series

All available streaming data information was considered significant for the formation of similarities. Therefore, all series that were watched during each time window contributed to the formation of the proximity matrix. However, in order to improve the interpretability of results, only the 50 most actively watched series for each time window were clustered. This is an important detail.

3.8 Initialization of cluster centroids

While initialization of cluster centroids as the previous time-steps centroid coordinate could be considered an improvement to traditional initialization, the algorithm instead randomizes the positioning of centroids at each time-step according to the k-means++ algorithm [53]. Each time the initialization of centroids is done, a cluster-metric is evaluated. After a selected number of random initializations, the centroid placement that optimizes over this metric best is selected as the final initialization.

3.9 ANCL framework parameters

Average association was used as the graph partitioning method. Static values for new content penalty and latent similarities decay were used as well as a Gaussian representation of the data.

3.10 Estimated number of clusters

Considering that the content catalog and user base is dynamic in several aspects, it was assumed that the structure and properties of content groups vary over time. That is, for example, the size, the optimal number of centroids or the optimal value for new content penalty or latent decay penalties, with α already being accounted for in the AFFECT framework. The optimal number of centroids was determined using two different methods. These two respective methods were selected because they are applied in two distinct manners. The eigengap heuristic is applied on the Laplacian matrix before performing the clustering, and the average sihouette width is applied on the clustering results.

The estimated number of clusters was assumed to vary over time and was therefore estimated uniquely for each time window. The average silhouette width [18] was used to obtain the value for the estimated number of clusters. Tests using a forced initialization of cluster centroids were also performed. For these tests, the estimated number of clusters was not calculated. Tests using forced initialization of cluster centroids were omitted from this report because they were not included in the scope of the ANCL framework.

3.11 Evaluation of algorithm

This section provides information about the different evaluation methods used to study the cluster results. In particular, this thesis studies the temporal stability of cluster results. The qualitative evaluation of clusters was performed by a domain expert and is not thoroughly presented explicitly in this report. The results were compared to segments in previous marketing efforts as well as with previous clustering efforts at the company.

3.11.1 Historical Weighting

The values of the adaptive forgetting factor α provide a natural opportunity for interpretation of the final clustering results. Especially between intermediate time windows. In order to learn what weight has been placed on historical values to produce the current clustering results it may be interesting to study the alpha values that correspond to that time-partition.

3.11.2 Tracking performance

The goal here is to obtain posterior error estimates of the true proximity matrix. The unknown parameter, in this case, is the true proximity matrix. The tracking error is expressed as the difference between the true proximity matrix and the one that we have estimated, the smoothed proximity matrix. That is, the mean squared error between the true affinity matrix and the estimated affinity matrix as follows:

$$\mathrm{MSE} = \mathbf{E}[\| \hat{\Psi}^t - \Psi^t \|_F^2] \tag{3.9}$$

This value is measured for the distinct values of alpha at every timestep.

3.11.3 Clustering performance

The number of cluster reassignments between two time periods is measured through the Rand Index:

$$R_{t-1,t} = \frac{\# \, of Agreements}{\# \, of Agreements + \# \, of Disagreements} \tag{3.10}$$

The Rand index calculation requires pairwise calculations for all objects included in the clustering. The Rand index ranges from 0 to 1, where 1 indicates identical clusters and 0 indicates that no pair of items are classified consistently across two clusters. In the setting of this thesis, the Rand index is applied to all clusters across two subsequent time windows.

3.11.4 M-score for top similarities

The clustering performance measured by the Rand index as described above is a measure applied to the the output of the algorithm. Similarly, the performance of the input to the clustering algorithm can be measured. A metric based on an affinity matrix is presented from this research as follows:

$$Mscore_{t-1,t} = \sum \frac{\%\ of\ previous}{n\ partitions} \tag{3.11}$$

For each series and time window, this metric measures the percentage of series that are consistently top 5 between each time window. Its main contribution is qualitative. The M-score gives us insight into how much the titles corresponding to the top similarities vary from a qualitative perspective and its particularly insightful when the magnitude of similarities differ largely across time.

3.11.5 Volatility of similarities

The standard error of similarity formations was estimated as the sample standard deviation divided by the square root of the sample size.

Chapter 4

Results

In this section, the ANCL smoothing framework is compared to observed data and the AFFECT framework. The planned comparisons were run with a Gaussian representation of the data and static values for the decay of latent similarities as described in 3.6.2 and for the penalty of entries that are derived from new content as described in 3.6.1. AFFECT was run with 9 clusters (AFFECT9) unless otherwise specified because it yielded the highest average silhouette score among the compared versions of AFFECT. The Mann Whitney U test was used for testing for differences in the mean and spread in values that were produced between the ANCL and AFFECT frameworks. The illustrated 45 day time windows, i.e. timesteps, were grouped into four separate groups, consisting of five timesteps for each group. Planned comparisons were run for each group independently with a random subset of users. Group 1 consists of timesteps 1-5, group 2 consists of timesteps 6-10, group 3 consists of timesteps 11-15, group 4 consists of timesteps 16-20.

4.1 Dynamics of the data

4.1.1 Robustness of the models

As described in chapter 2, the similarities over time between different series are likely to vary due to the dynamics in the catalog and user behavior. In figures 4.1 and 4.2 it is shown that the mean of all observed pairwise similarities vary vastly for the catalog of series. The results are plotted for the most watched series as well as the entire catalog in order to illustrate the discrepancies between the values that are included in the clustering versus all values available at any time window. It is illustrated that both the AFFECT and ANCL frameworks smoothen the volatility in the magnitude of similarities. This indicates that both AFFECT and, in particular, the ANCL framework are capable of reducing spikes in observed similarities. Thereby, both models are more robust to temporal fluctuations of similarities than their static counterpart. Figure 4.2 shows that the mean of all similarities diverge between the observed values and the smoothed values for the ANCL framework when applied to the entire catalog of series. This is mainly since, by design, the ANCL framework decays latent similarities over time and fills sparse entries below the mean. Missing values for AFFECT are also filled below the mean. However, latent similarities are not decayed. This implies that observed similarities will become latent unless observed in a subsequent time window. Therefore, such a divergence is not as pronounced in figure 4.1. This effect is not expressed in figures 4.1 and 4.2 for the most watched series, likely because those series correspond to a higher percentile of users, and therefore are inherently more consistent over time since there are fewer latent entries in the proximity matrix.

For tables 4.1 to 4.4, a one-sided nonparametric test was run to determine whether there is a difference in means between the AFFECT and ANCL framework. The Mann-Whitney U statistic ranges from 0 to 625, indicating a low to high separation between the frameworks, respectively. For this test, the timesteps were partitioned into four groups. Table 4.1 indicates that we cannot reject the one-sided null hypothesis for the most watched series. Table 4.2 indicates that we can reject the one-sided null hypothesis for three out of four groups. Tables 4.3 and 4.4 indicate that we can reject the one-sided null hypothesis for all groups.

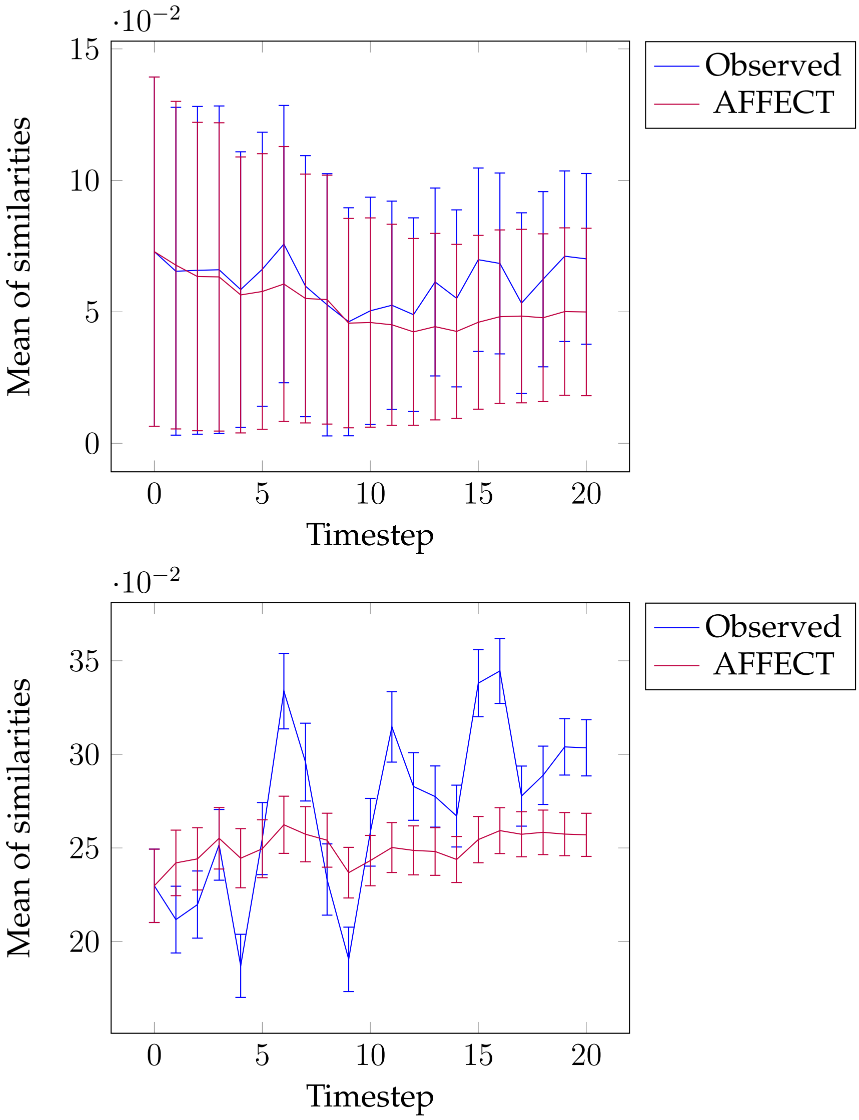

Figure 4.1: Comparison of the mean of all pairwise similarities for the most watched series (top plot) and for the entire series catalog (bottom plot). The mean and standard error is included for observed values from the data as well as for smoothed values from the AFFECT framework in each respective plot. (Number of sampled users = 2000)

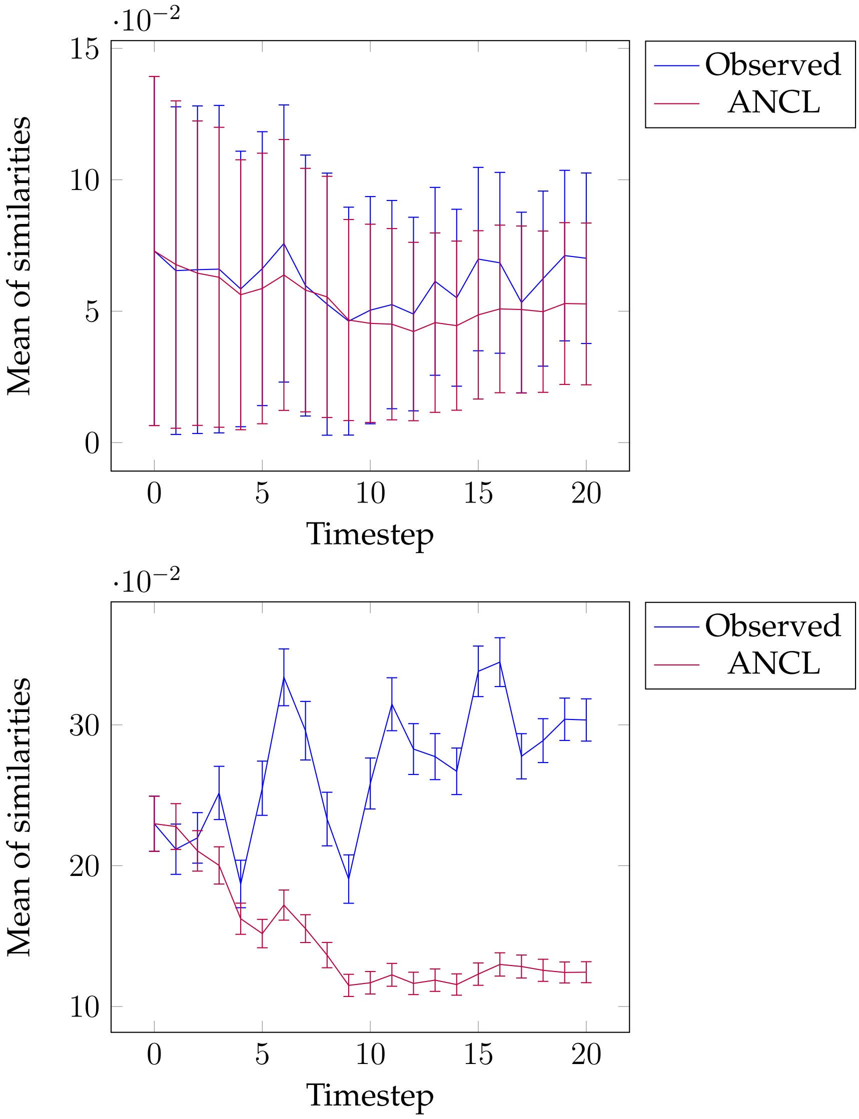

Figure 4.2: Comparison of the mean of all pairwise similarities for the most watched series (top plot) and for the entire series catalog (bottom plot). The mean and standard error is included for observed values from the data as well as for smoothed values from the ANCL framework in each respective plot. (Number of sampled users = 2000)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	335	0.33	No
group 2	374	0.12	No
group 3	392	0.063	No
group 4	279	0.75	No

Table 4.1: Mann-Whitney U test for the values of the mean of similarities for the most watched series. The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	376	0.11	No
group 2	396	0.048	Yes
group 3	516	4.095e-05	Yes
group 4	577	1.508e-07	Yes

Table 4.2: Mann-Whitney U test for the values of the standard deviation of similarities for the most watched series. The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	558	9.98e-07	Yes
group 2	620	1.29e-09	Yes
group 3	625	7.078e-10	Yes
group 4	625	7.078e-10	Yes

Table 4.3: Mann-Whitney U test for the values of the mean of similarities for the entire catalog. The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	561	7.47e-07	Yes
group 2	625	7.078e-10	Yes
group 3	625	7.078e-10	Yes
group 4	625	7.078e-10	Yes

Table 4.4: Mann-Whitney U test for the values of the standard deviation of similarities for the entire catalog. The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

4.1.2 Flexibility of cluster formations

The estimated number of clusters was determined through the silhouette score in order to impose compact and well separated clusters through the ANCL framework. The eigengap heuristic was also used to calculate the estimated number of clusters. Since the eigengap heuristic is calculated before filtering of most watched series, see figure 3.1, the corresponding values were calculated a posteriori for illustrative purposes. The estimated number of clusters was determined for observed values as well as for AFFECT and ANCL. In figure 4.3 it is shown that there is a difference in the estimated number of clusters determined for observed similarities and smoothed similarities.

Notably, the smoothed similarities tend to yield a higher number of estimated clusters than a static counterpart, and fewer occurrences at the low threshold of a minimum of two clusters. This result indicates that the structural properties of similarities are affected by the smoothing operation. The estimated number of clusters was found to vary over time for observed, AFFECT and ANCL. This finding is an indication that the structural properties of user behavior vary over time. One must take note that the estimated number of clusters was only enforced for the ANCL framework, by design. The results indicate that the estimated number of clusters is different for ANCL and AFFECT. This finding is expected because the adaptive number of clusters is enforced in the ANCL framework and has implications for the calculation of the block model, and, in turn, alpha. In contrast, the number of clusters in AFFECT is static and has been calculated and plotted for illustrative purposes in this analysis. Both AFFECT and ANCL have a different estimated number of clusters than observed values, indicating that both frameworks impose a certain structure on the data. This finding is expected. Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing [54].

One should take note that the minimum number of clusters is restricted to two. The eigengap, therefore, yields the minimum number of clusters when applied to the entire catalog. This result may indicate that no superior structure can be found with this method.

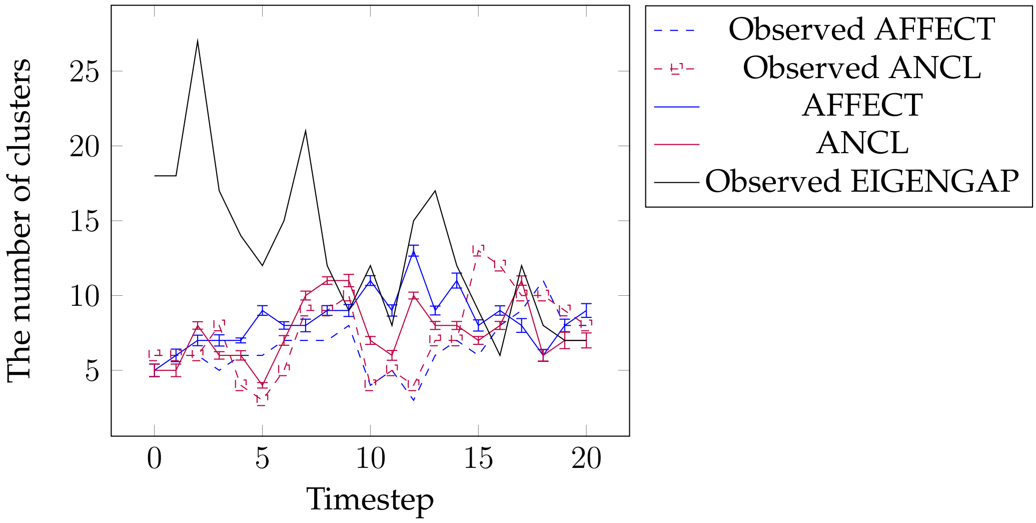

Figure 4.3: The estimated number of clusters as determined through the silhouette score and eigengap heuristic for the most watched series. The estimated number of clusters are included for the smoothing frameworks ANCL and AFFECT, as well as for observed data. (Number of sampled users = 2000)

4.1.3 Temporal stability

The cluster membership reassignments over time is the proxy for cluster stability within the context of this thesis work. The choice of α was made based on equation 3.6, but can also be chosen arbitrarily. The implication on cluster membership reassignments over time resulting from different choices of α is presented from a testrun below. Figure 4.4 shows that the measure of stability of clusters over time, the Rand index, is more volatile and consistently lower for the two static choices of α. The adaptive choice of α is shown to yield a consistently higher Rand index in this testrun. The result is presented to highlight the performance of the automated method of parameter selection, in this case α, with regards to the stability of results. The adaptive selection of alpha implies fewer cluster membership reassignments over subsequent time windows.

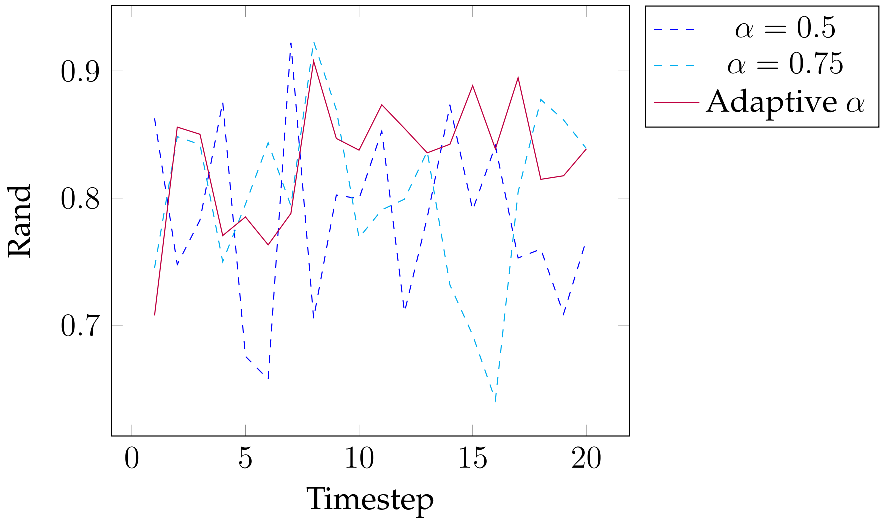

Figure 4.4: *Observed* cluster reassignments for different choices of alpha as measured by the Rand index. A higher Rand index indicates fewer cluster membership reassignments between two subsequent time windows. (Number of sampled users = 2000)

4.1.4 MSE

As described in section 3.11.2, the loss function is derived as the difference between the true affinity matrix and the estimated affinity matrix. In figure 4.5, a testrun is shown where the adaptively selected α consistently yields a lower MSE than the two static choices of α. This is an expected result since the model explicitly optimizes for this criterion. As above, this result is presented to highlight the performance of the automated method of selecting α.

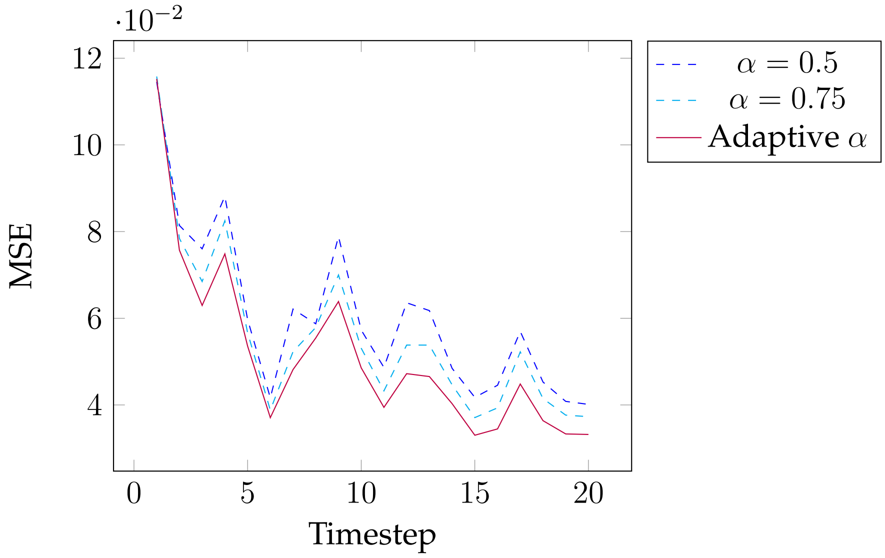

Figure 4.5: *Observed* tracking error (MSE) for different choices of alpha. (Number of sampled users = 2000)

4.2 Further comparisons

4.2.1 Alpha over time

Figure 4.6 shows the difference in historical weighting for the AFFECT and ANCL frameworks. The ANCL framework produces lower values for alpha than AFFECT. This means that the ANCL framework puts a relatively lower emphasis at any given time window on similarities derived from historical data, rather than observed values. As illustrated in table 4.5, a nonparametric test was run to determine whether there is a difference between the AFFECT and ANCL framework. The U statistic ranges from 0 to 625, indicating a low to high separation between the frameworks, respectively. For this test, the time interval was partitioned into 4 segments. Table 4.5 indicates evidence against the null hypothesis that the values obtained from AFFECT and ANCL were drawn from the same distribution. The alternative hypothesis is that the AFFECT framework is stochastically larger than the ANCL framework.

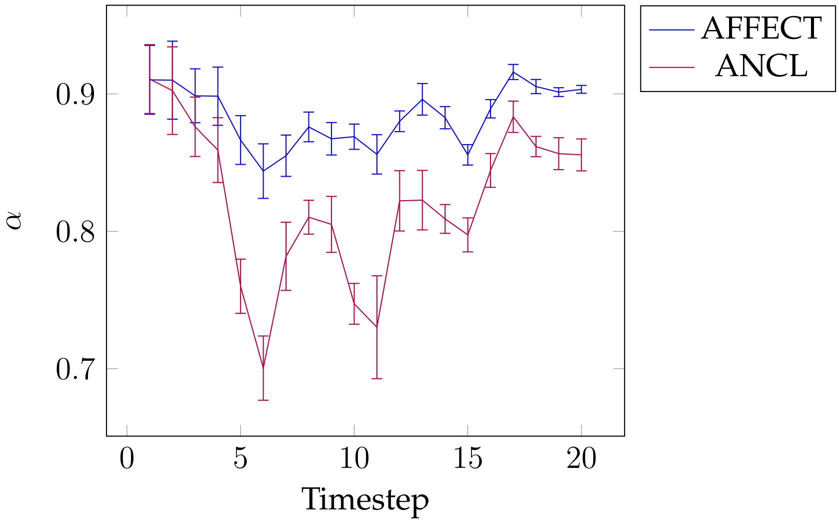

Figure 4.6: Estimates for the optimal shrinkage intensity α. (Number of sampled users = 2000)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	423	0.016	Yes
group 2	459	0.0023	Yes
group 3	529	1.39e-05	Yes
group 4	568	3.75e-07	Yes

Table 4.5: Mann-Whitney U test for the parameter alpha between AFFECT and ANCL. The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

4.2.2 Rand and MSE over time

As summarized in the top plot of figure 4.7, a static clustering was performed on every time window. The Rand index between time windows was calculated for the static and the two frameworks. The results demonstrate that the ANCL framework is able to control the trade off between historical and current cost. As depicted in 4.7, the cluster membership reassignments over time are more volatile for the

ANCL framework than AFFECT. Notably, both frameworks have a lower number of cluster membership reassignments than the static counterpart. The ANCL framework consistently produces a lower tracking error (MSE) than the AFFECT framework.

As illustrated in tables 4.1 to 4.4, a nonparametric test was run to determine whether there is a difference between the AFFECT and ANCL framework. The U statistic ranges from 0 to 625, indicating a low to high separation between the frameworks, respectively. For this test, the time interval was partitioned into 4 segments. Table 4.6 indicates evidence against the null hypothesis that the values obtained from AFFECT and ANCL were drawn from the same distribution. The alternative hypothesis is that the AFFECT framework is stochastically larger than the ANCL framework.

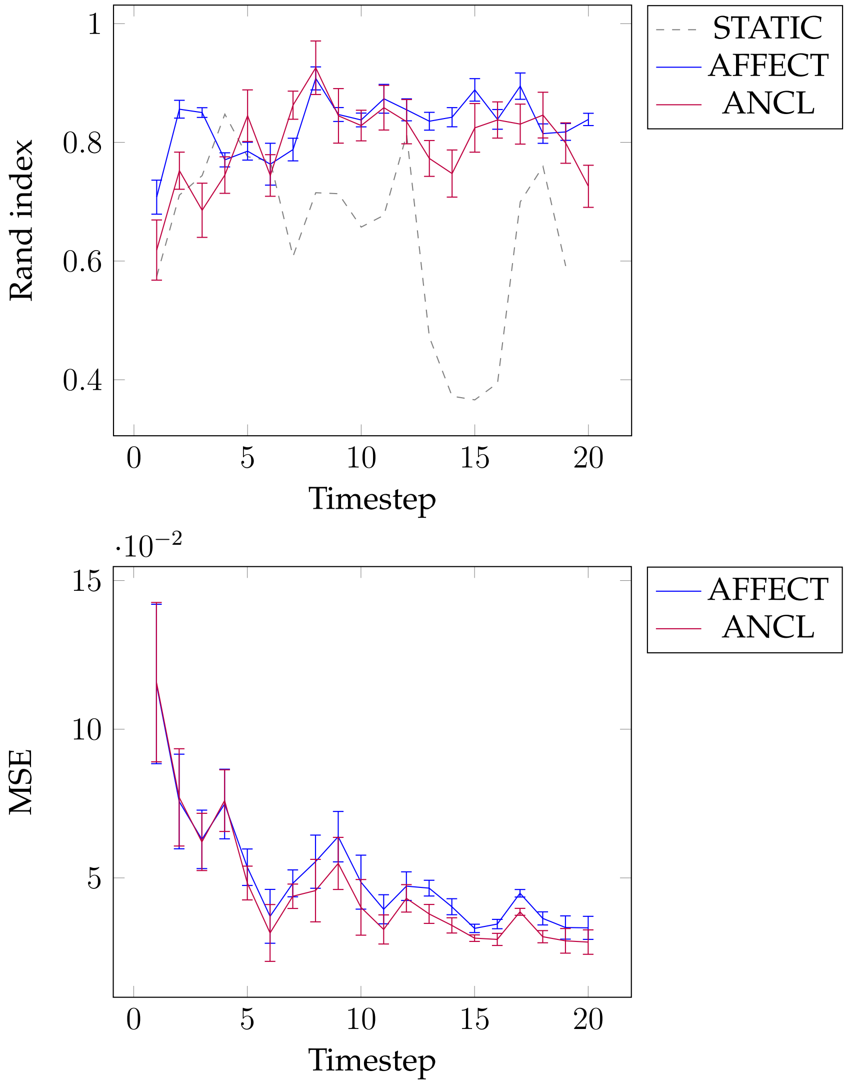

Figure 4.7: The top plot shows the reassignment similarity measured by the Rand index. A higher Rand index indicates less cluster membership reassignments between two subsequent time windows. The bottom plot shows the loss function as measured by the mean squared error. (Number of sampled users = 2000)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	554	1.46e-06	Yes
group 2	565	5.048e-07	Yes
group 3	527	1.64e-05	Yes
group 4	531	1.12e-05	Yes

Table 4.6: Mann-Whitney U test for the Rand index between AFFECT and ANCL. The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

Timesteps	U statistic	p-value	Reject Null at 5%
group 1	323	< 0	No
group 2	128	< 0	No
group 3	200	< 0	No
group 4	45	< 0	No

Table 4.7: Mann-Whitney U test for the tracking error (MSE). The null hypothesis is that the mean of the values of AFFECT are not larger than for the ANCL framework. The alternative hypothesis is that the mean of the values from AFFECT are larger than for the ANCL framework. (25 samples were evaluated for each group of timesteps. Each sampled entry has a randomly drawn subset of 2000 users.)

4.2.3 Optimizing for MSE

The shrinkage parameter α, and the number of clusters n, are computed to optimize for two different criteria. α is optimized to place a weight between historical and observed similarities as to minimize the MSE between the true affinity matrix and the estimated affinity matrix. In contrast, the selection of the number of clusters is aimed at yielding compact and well separated clusters. Therefore, the optimal shrinkage factor will not necessarily yield the highest silhouette score. The testrun in figure 4.8 illustrates that the optimal α, although not explicitly optimizing for compactness and separation, may yield a higher silhouette score than the first estimation of α.

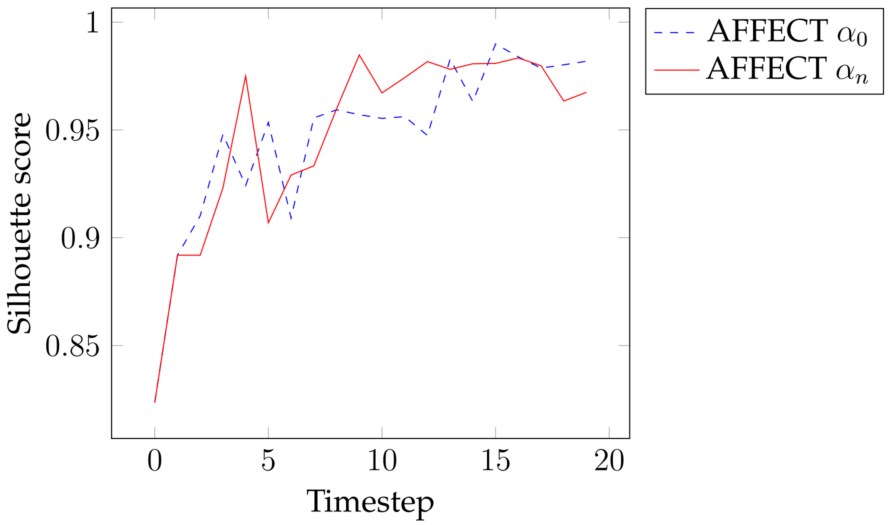

Figure 4.8: The silhouette score of the clustering result using the first estimation of the shrinkage parameter α is compared to the corresponding values for the improved values of α. (Number of sampled users = 2000)

Framework	t statistic	p-value	Reject Null at 5%
AFFECT3	10.55	5.68e-10	Yes
AFFECT5	4.26	0.00028	Yes
AFFECT9	0.94	0.35	No
AFFECT12	7.97	9.07e-08	Yes

Table 4.8: T-test for the silhouette score means of each respective framework. The null hypothesis that AFFECT and ANCL have identical average expected values can be rejected for all except AFFECT9 on the 5% level.

4.2.4 Consistency of top similarities over time

As illustrated through the testrun in figure 4.9, the series corresponding to the five highest similarities for each series in the catalog changes less between each time window for the ANCL framework than for the AFFECT framework, with observed values of the M-score far lower than both. It was observed that, for AFFECT, large similarity values were formed at any given time window and became latent over time. Since it is highly unlikely that observed values for pairwise are nearidentical between time windows, which also proved to hold false upon further study, it was deemed a weakness of this particular implementation of the AFFECT algorithm. These similarities became latent either because of the large amount of processing which filtered out series, thus rendering any updates of their pairwise similarities null, or through changes in availability of content in the platform, as discussed earlier in this report.

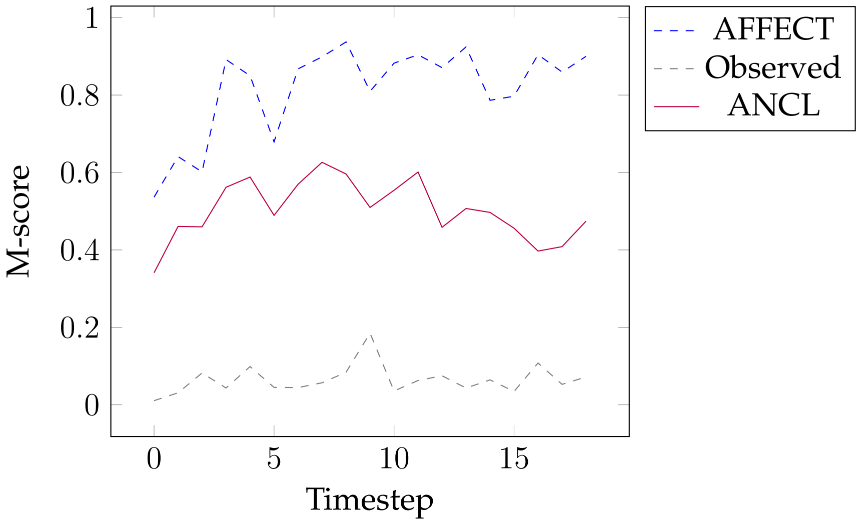

Figure 4.9: M-score plotted for observed values as well as for the AFFECT and ANCL frameworks. A high M-score indicates that, for all series, the five series corresponding to the highest produced similarity values of all series is consistent between two time windows. (Number of sampled users = 2000)

4.3 Evaluation of segments

The groups of series were evaluated by a domain expert and heuristically compared to other clustering results that have been performed at the company. A large number of series in the ANCL framework were found to be classified into similar groups as the other segmentation efforts and also largely in agreement with series segments formulated by the domain expert.

Chapter 5

Discussion

5.1 Key findings

The purpose of this thesis was to perform a clustering of film series that produces stable cluster membership assignments over time and which captures nuances in user behavior in order to provide a foundation for reasoning about the content mix. For reasons of confidentiality, clustering results are not explicitly illustrated in this report. In the absence of defined qualitative metrics, the ANCL framework was compared to previous clustering results at the company. While previous clustering efforts have expressed similarities with series segments formulated by a domain expert, the clusters have not been stable over time. The clusters produced by the ANCL framework are compact, clearly separated and exhibit stable cluster membership assignments over time. In contrast to static clustering which tends to lend itself towards undesirable properties in terms of cluster structure, the structure produced by the ANCL framework can be considered superior within the context studied in this thesis. For example, no cluster contains a vast majority of the objects being clustered, which is an effect that could be observed during the thesis process for static clustering applied to certain time windows. Such a result could impede the interpretability or actionability of segments, as discussed previously in this thesis. In general, the ANCL framework handles the major obstacles of the data domain that have been presented, such as the temporary effects on pairwise similarities arising from a release of a new series, or from users and series dropping and joining. The major contribution of this research is the development of a state-of-the-art segmentation model that enables a detailed study of user and content groups over time. The model is designed to be flexible such that preference for time interval length, historical bias etc. can be encoded to optimize for a business objective. Although this research exhibits potential to introduce a new method of evaluating clusters of series, the results cannot be fully evaluated within the context of this thesis. Exploratory research is often evaluated by applying the results in a business context and observing potential effects.

5.2 The data domain

The manner in which users interact with a streaming platform and consume content is highly dynamic and dependent on a number of factors such as UX design or marketing efforts [55]. Therefore, the observed temporal fluctuations of aggregate similarities as illustrated in figures 4.1 and 4.2 most likely represent a common characteristic for video streaming data as it reflects a dynamic user interaction with a VOD service and content. The analysis and smoothing of such irregularities is important for facilitating the analysis of groups [56]. In addition to the aggregate smoothing of similarities, we believe that the smoothing between individual pairwise similarities is of equal importance. Many applications tailor a service on an individual level and thus place emphasis on pairwise relations in isolation. The smoothing of individual pairwise similarities is further discussed in section 5.5.

The divergence as described in figure 4.2 is a result of intentionally increasing the dependence of observed values for the clustering results, rather than synthetic or latent values. It is assumed that decision making based on observed values hold higher credibility than decisions based on synthetic or latent similarities. Therefore, this effect is assumed to improve the actionability of results. This is because, arguably, decisions are more likely to be based on observed values rather than imputed. In turn, this implies greater interpretability since observed values can be directly related to user behavior over time. This should not be mistaken with introducing a preference for current rather than historical similarities.

The divergence expressed in 4.2 could also be interpreted as a consolidation of views towards core series in the catalog, as time passes and user behavior or preference possibly matures. However, this result is not expressed for the AFFECT framework, and can thus possibly be concluded to be an effect from the ANCL framework implementation. In table 4.3 we can indeed note a high separability between AFFECT and ANCL with regards to the similarities in the proximity matrix. Lastly, the observed values between top series and the entire catalog do not express this effect, which further weakens this hypothesis.

The ANCL framework imputes missing values in order to improve the interpretability of results. It is assumed that a sparse matrix is more difficult to interpret since values will be missing for many pairwise comparisons. In contrast, with the imputation strategy described in 3.4.2, there will exist an estimate for any such comparison. In other words, this result allows for greater flexibility in a posteriori analysis, since far more comparisons and evaluations can be performed.

5.3 An optimal number of clusters

Chi et al. [17] make the assumption that the number of clusters n remains the same over all time. It is mentioned that this is a very strong restriction, although that the evolutionary framework can handle a variation in the number of clusters, in theory. A major contribution of the ANCL framework is the variation and optimization of the number of clusters at each time window, as illustrated in figure 4.3. This allows for greater flexibility in the formation of clusters which could arguably yield results that more accurately reflect user behavior. This is because groupings are not constrained by an a priori decision of the number of clusters. Any such decision of n may otherwise be sub optimal with regards to the cluster properties, either resulting from a poor arbitrary choice of n or since the optimal number of clusters may vary between any given time window [20]. Figure 4.8 illustrates this characteristic.

The importance of this development deserves emphasis. This extension can more easily allow structural changes of clusters to be captured in the results, regardless of the data domain that is studied. In the setting of this thesis, any variation in the number of clusters reflects changes in the inherent structure that is derived from user behavior. Specifically, a change in n implies a change in the number of content groups, as constrained by compactness and separability. The silhouette score imposes this constraint in the results and therefore represents a trade-off between representing the data as accurately as possible versus improving the interpretability and actionability of the content groups that are produced from the model. Since the major topic of interest of this research is to capture and handle dynamic user behavior in the results to a greater extent than static clustering, there exists a clear raison d'être for this extension. The evolutionary properties of clusters, such as the number of nodes, have furthermore been emphasized as an important perspective in providing insight from research [40, 28, 57, 29].

Nonetheless, a static number of clusters may be desired in certain applications, and the ANCL framework can then be adjusted for this purpose. This can be considered a way to specify real-valued preferences in order to constrain the solution space [58, 59]. Similarly, the ANCL framework allows for different choices for the decay of latent similarities or penalty of entries derived from new content. The latter arguably provides a particularly interesting opportunity to specify preference in the amount of regularization on characteristics that are specific to VOD data and related to spikes in popularity for the release of new series or seasons.

While the AFFECT framework is shown to be able to produce compact and well separated clusters, table 4.8 illustrates that this effect is dependent on an optimal choice in the number of clusters. For many applications, manual selection of n may be unfeasible. Together with the freedom of choosing the factors for decay of latent similarities and penalization of entries deriving from new content, this development contributes to the flexibility of the ANCL framework relative to AFFECT.

5.4 Responsiveness

Alpha can be interpreted as the extent to which the model can detect changes that are reflected in the data [19, 17]. A high value of alpha puts larger weight on historical values, and therefore more slowly incorporates changes into the cluster results. Correspondingly, a low value of alpha places smaller weight on historical values in any given time window. In figure 4.6, the ANCL framework is shown to consistently produce lower values of alpha than AFFECT does, and can therefore be argued to be more responsive to changes in behavior. Additionally, the values of alpha vary more for ANCL than AFFECT, indicating that the calculation of alpha itself is more responsive. Since the value of alpha may have implications on cluster membership reassignments, this result may be concerning. However, it was found that the Rand index of the ANCL framework lies within range of the different AFFECT variations that were run. This result indicates that the ANCL framework can handle a dynamic number of clusters without sacrificing the responsiveness of the model to changes in the data. However, since no optimal number of cluster membership reassignments has been declared, it is difficult to determine which model performs best in this regard.

Interestingly, while both the ANCL and AFFECT frameworks produce more stable results than the static counterpart, see figure 4.7, the ANCL framework is more responsive. This could perhaps indicate that the model captures user behavior in a more nuanced manner. If we believe that user behavior evolves over time, so must the calculation of alpha, to a certain extent.

5.5 Interpretability

A high M-score may, *ceteris paribus*, lead to temporally stable clustering results, but is not necessarily a good metric for the quality of the clustering itself at any time window. For example, a redundant solution may consist of an identical clustering at any given time window. This is easy to achieve, for instance by setting α to the value of one. Furthermore, a perfectly consistent sequence of clusters contradicts the fundamental assumption that the user behavior and content catalog expresses dynamic properties, as this should be reflected in the groupings of series. With the observed M-score values in mind, it was assumed that the AFFECT framework produced unreasonably high values for the M-score. It was assumed that an M-score that is closer to observed values would yield a more actionable result, because it would capture changes to a greater extent. As illustrated in figure 4.9, the ANCL framework manages to produce this effect implicitly by decaying latent values in the similarity matrix. As previously mentioned in this thesis, latent values were not assumed to contribute to the clustering results in a desirable manner, since they may give a false impression of the pairwise relationship between two series. For series that drop and join the catalog in a cyclical manner, often due to distribution rights, this may be an unfortunate effect. This is because observed similarities will be decayed until the next time the series is available. However, for the series that have a low number of users, and are therefore either filtered out or not watched during a time window, this implication is not considered a large compromise.

The M-score is an important aspect of interpretability. Just as understanding of recommendation is important to users [55], the basis on which it is formed may be important for decision making. It is here assumed that the researcher has a bias towards studying the most similar series for any given title.

5.6 Ethics and sustainability

Machine learning systems leverage data to perform a wide array of tasks which have the potential to benefit or harm society. It is the responsibility of the scientific community as well as each individual researcher to take into account ethics and sustainability in their work.

This thesis leverages user information as the segmentation basis. Users may not have any desire for their information to be used for scientific research or marketing efforts. In particular, features such as gender or race may be considered sensitive and have not been available or considered in any manner. The data used in this thesis work is anonymous and does not contain any other information than the user interaction with the platform and the meta-data of series. Behavioral variables have been used exclusively which avoids the processing of any potentially sensitive information. The overall aim of this thesis is to provide basis to improve the core product offering.

Chapter 6

Conclusions

In this thesis work it is shown that a segmentation of film series that produces stable cluster membership assignments can be obtained by applying adaptive evolutionary spectral clustering on video streaming data. By introducing an automated selection of the number of clusters in each time window, compactness and separation of clusters are improved consistently over time. Also, by introducing temporal decay of latent similarities as well as static penalty of new content, clustering results are, as judged heuristically by a domain expert, significantly improved.

6.1 Outcomes

- Both the AFFECT and ANCL frameworks provide significant improvements over static clustering with regards to robustness, flexibility and interpretability.

- The AFFECT framework, while well-posed, is not robust enough to the video on demand problem domain. The ANCL framework introduces a number of changes that, as judged by a domain expert, improve the results in this particular context. Principally, the ANCL framework improves the structural properties and interpretability of clustering results as compared to the AFFECT framework.

- The ANCL framework improves on AFFECT to allow greater freedom in cluster formations by optimizing the number of clusters at each time window. This can potentially better capture nuances in the data, as well as provide a better foundation for analysis.

- The ANCL framework enables further segmentation efforts by providing an evolution of content groups, thereby allowing benchmarking and analysis of user and content groups over time. This will allow study of content groups over time, rather than individual series over time.

6.2 Further Work

Time and computational resources were the largest limitations for this thesis. For future research, it would be interesting to include a larger subset of users in the data. With regards to the presented ANCL framework, future research could consider how to optimize and automate selection of the decay of latent similarities or penalization of entries deriving from new content. The decay could be adjusted to exponentially decay towards the mean, for example. Furthermore, the method of penalizing entries deriving from new content should be revised as it, arguably, corrupts the item-user matrix data in its current form of implementation. It is perhaps the most apparent way of improving the ANCL framework.

Some final interesting perspectives for future research are the optimization of cluster centroid initialization with a dynamic number of clusters or the study of different number of time windows. Also, since the evolution of the clusters originate from the initial calculation of the true proximity matrix, it would be interesting to vary the length of the first time window and observe if and how the clusters are affected. This could, for example, help determine if there is a minimum size of data needed before running the framework.

Bibliography

[1] Pat Langley and Herbert A Simon. "Applications of machine learning and rule induction". In: *Communications of the ACM* 38.11 (1995), pp. 54–64.

[2] Indranil Bose and Radha K Mahapatra. "Business data mining—a machine learning perspective". In: *Information & management* 39.3 (2001), pp. 211–225.

[3] Ian H Witten et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[4] Russell S Winer. "A framework for customer relationship management". In: *California management review* 43.4 (2001), pp. 89–105.

[5] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification". In: *Expert systems with applications* 36.2 (2009), pp. 2592–2602.

[6] Yen-Liang Chen et al. "Market basket analysis in a multiple store environment". In: *Decision support systems* 40.2 (2005), pp. 339–354.

[7] The Nielsen Company. *Video On Demand: How worldwide viewing habits are changing in the evolving media landscape*. Ed. by The Nielsen Company (US). [Online; posted 16-March-2016]. Mar. 2016. URL: http://www.nielsen.com/content/dam/nielsenglobal/eu/docs/pdf/Nielsen-global-video-on-demand.pdf.

[8] C-Y Tsai and C-C Chiu. "A purchase-based market segmentation methodology". In: *Expert Systems with Applications* 27.2 (2004), pp. 265–276.

[9] Mary J Culnan. “Protecting privacy online: Is self-regulation working?” In: *Journal of Public Policy & Marketing* 19.1 (2000), pp. 20–26.

[10] Ramnath K Chellappa and Raymond G Sin. “Personalization versus privacy: An empirical examination of the online consumer's dilemma”. In: *Information technology and management* 6.2-3 (2005), pp. 181–202.

[11] Michel Wedel and Wagner A Kamakura. *Market segmentation: Conceptual and methodological foundations*. Vol. 8. Springer Science & Business Media, 2012.

[12] Peter Dayan, Maneesh Sahani, and Grégoire Deback. “Unsupervised learning”. In: *The MIT encyclopedia of the cognitive sciences* (1999).

[13] Ulrike Von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and computing* 17.4 (2007), pp. 395–416.

[14] Fan RK Chung. *Spectral graph theory*. 92. American Mathematical Soc., 1997.

[15] Mark EJ Newman. “Modularity and community structure in networks”. In: *Proceedings of the national academy of sciences* 103.23 (2006), pp. 8577–8582.

[16] Jianbo Shi and Jitendra Malik. “Normalized cuts and image segmentation”. In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), pp. 888–905.

[17] Yun Chi et al. “On evolutionary spectral clustering”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3.4 (2009), p. 17.

[18] Peter J Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[19] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. “Evolutionary clustering”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006, pp. 554–560.

[20] Kevin S Xu, Mark Kliger, and Alfred O Hero Iii. “Adaptive evolutionary clustering”. In: *Data Mining and Knowledge Discovery* 28.2 (2014), pp. 304–336.

[21] Urszula Kuzelewska. “Clustering algorithms in hybrid recommender system on MovieLens data”. In: *Studies in logic, grammar and rhetoric* 37.1 (2014), pp. 125–139.

[22] S Guha et al. “Clustering data streams”. In: *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society. 2000, p. 359.

[23] Charu C Aggarwal et al. “-A Framework for Clustering Evolving Data Streams”. In: *Proceedings 2003 VLDB Conference*. Elsevier. 2003, pp. 81–92.

[24] Rocco Langone, Carlos Alzate, and Johan AK Suykens. “Kernel spectral clustering with memory effect”. In: *Physica A: Statistical Mechanics and its Applications* 392.10 (2013), pp. 2588–2606.

[25] Steve Chien and Nicole Immorlica. “Semantic similarity between search engine queries using temporal correlation”. In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 2–11.

[26] Lei Tang et al. “Community evolution in dynamic multi-mode networks”. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 677–685.

[27] Yu-Ru Lin et al. “Facetnet: a framework for analyzing communities and their evolutions in dynamic networks”. In: *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, pp. 685–694.

[28] Yu-Ru Lin et al. “Analyzing communities and their evolutions in dynamic social networks”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3.2 (2009), p. 8.

[29] Derek Greene, Donal Doyle, and Padraig Cunningham. “Tracking the evolution of communities in dynamic social networks”. In: *Advances in social networks analysis and mining (ASONAM), 2010 international conference on*. IEEE. 2010, pp. 176–183.

[30] Ulrike Von Luxburg et al. “Clustering stability: An overview”. In: *Foundations and Trends® in Machine Learning* 2.3 (2010), pp. 235–274.

[31] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. “A stability based method for discovering structure in clustered data”. In: *Biocomputing 2002*. World Scientific, 2001, pp. 6–17.

[32] Asa Ben-Hur and Isabelle Guyon. “Detecting stable clusters using principal component analysis”. In: *Functional genomics*. Springer, 2003, pp. 159–182.

[33] Sara Dolnicar. “Using cluster analysis for market segmentation-typical misconceptions, established methodological weaknesses and some recommendations for improvement”. In: (2003).

[34] Silke Wagner and Dorothea Wagner. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[35] Gediminas Adomavicius and Jingjing Zhang. “Stability of recommendation algorithms”. In: *ACM Transactions on Information Systems (TOIS)* 30.4 (2012), p. 23.

[36] Stephen Gower. *Netflix Prize and SVD*. 2014.

[37] Alexander Singleton, Michail Pavlis, and Paul A Longley. “The stability of geodemographic cluster assignments over an intercensal period”. In: *Journal of Geographical Systems* 18.2 (2016), pp. 97–123.

[38] Xiang Ji and Wei Xu. “Document clustering with prior knowledge”. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2006, pp. 405–412.

[39] Yun Chi et al. “Evolutionary spectral clustering by incorporating temporal smoothness”. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2007, pp. 153–162.

[40] Yi Wang et al. “Mining naturally smooth evolution of clusters from dynamic data”. In: *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM. 2007, pp. 125–134.

[41] Purnamrita Sarkar and Andrew W Moore. “Dynamic social network analysis using latent space models”. In: *Advances in Neural Information Processing Systems*. 2006, pp. 1145–1152.

[42] Aoying Zhou et al. "Tracking clusters in evolving data streams over sliding windows". In: *Knowledge and Information Systems* 15.2 (2008), pp. 181–214.

[43] Kevin S Xu, Mark Kliger, and Alfred O Hero. "Evolutionary spectral clustering with adaptive forgetting factor". In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE. 2010, pp. 2174–2177.

[44] Shirley Young, Leland Ott, and Barbara Feigin. "Some practical considerations in market segmentation". In: *Journal of Marketing Research* (1978), pp. 405–412.

[45] Anders Drachen, Alessandro Canossa, and Georgios N Yannakakis. "Player modeling using self-organization in Tomb Raider: Underworld". In: *Computational Intelligence and Games, 2009. CIG 2009. IEEE Symposium on*. IEEE. 2009, pp. 1–8.

[46] Anders Drachen et al. "Guns, swords and data: Clustering of player behavior in computer games in the wild". In: *Computational Intelligence and Games (CIG), 2012 IEEE Conference on*. IEEE. 2012, pp. 163–170.

[47] Rafet Sifa et al. "Behavior evolution in tomb raider underworld". In: *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*. IEEE. 2013, pp. 1–8.

[48] Sherrie YX Komiak and Izak Benbasat. "The effects of personalization and familiarity on trust and adoption of recommendation agents". In: *MIS quarterly* (2006), pp. 941–960.

[49] John O'Donovan and Barry Smyth. "Is trust robust?: an analysis of trust-based recommendation". In: *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM. 2006, pp. 101–108.

[50] Katherine Faust and Stanley Wasserman. "Blockmodels: Interpretation and evaluation". In: *Social networks* 14.1-2 (1992), pp. 5–61.

[51] Joseph E Schwartz. "An examination of CONCOR and related methods for blocking sociometric data". In: *Sociological methodology* 8 (1977), pp. 255–282.

[52] Olivier Ledoit and Michael Wolf. “Improved estimation of the covariance matrix of stock returns with an application to portfolio selection”. In: *Journal of empirical finance* 10.5 (2003), pp. 603–621.

[53] David Arthur and Sergei Vassilvitskii. “k-means++: The advantages of careful seeding”. In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[54] Mark S Aldenderfer and Roger K Blashfield. “Cluster analysis: Quantitative applications in the social sciences”. In: *Beverly Hills: Sage Publication* (1984).

[55] James Davidson et al. “The YouTube video recommendation system”. In: *Proceedings of the fourth ACM conference on Recommender systems*. ACM. 2010, pp. 293–296.

[56] Kun Tu et al. “Temporal Clustering in Time-varying Networks with Time Series Analysis”. In: ().

[57] Huazhong Ning et al. “Incremental spectral clustering by efficiently updating the eigen-system”. In: *Pattern Recognition* 43.1 (2010), pp. 113–127.

[58] X Yu Stella and Jianbo Shi. “Grouping with bias”. In: *Advances in neural information processing systems*. 2002, pp. 1327–1334.

[59] Xiang Wang and Ian Davidson. “Flexible constrained spectral clustering”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, pp. 563–572.

Appendix A

Derivation of alpha

In order to reach an expression for α we first consider the Frobenius norm between the true and the estimated proximities. The quadratic loss function is expressed as a function of α, the parameter which we are optimizing over.

$$L(\alpha^t) = \parallel \hat{\Psi}^t - \Psi^t \parallel_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (\hat{\psi}_{ij}^t - \psi_{ij}^t)^2$$

Resultingly, the risk is expressed conditionally on previously observed similarities, where $W^{(t-1)}$ denotes the set $\{W^{t-1}, W^{t-2}, ..., W^0\}$:

$$R(\alpha^t) = \mathrm{E}[\parallel \hat{\Psi}^t - \Psi^t \parallel_F^2 |W^{(t-1)}]$$

$$\mathrm{E}[W^t|W^{(t-1)}] = E[W^t] = \Psi^t$$

$$\mathrm{var}(W^t|W^{(t-1)}) = \mathrm{var}(W^t) = \mathrm{var}(N^t)$$

$N^t N^{(t-1)}, ..., N^0$ are mutually independent with zero mean

Now the risk is expressed as follows:

$$R(\alpha^t) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}[(\alpha^t\,\hat{\psi}_{ij}^{t-1} + (1-\alpha^t)\,w_{ij}^t - \psi_{ij}^t)^2|W^{(t-1)}]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\{\mathrm{var}(\alpha^t\,\hat{\psi}_{ij}^{t-1}+(1-\alpha^t)\,w_{ij}^t-\psi_{ij}^t|W^{(t-1)})+\mathrm{E}[\alpha^t\,\hat{\psi}_{ij}^{t-1}+(1-\alpha^t)\,w_{ij}^t-\psi_{ij}^t|W^{(t-1)}]^2\}$$

$$R(\alpha^t) = \sum_{i=1}^{n}\sum_{j=1}^{n}\{(1-\alpha^t)^2\,\mathrm{var}(w_{ij}^t) + (\alpha^t)^2(\hat{\psi}_{ij}^{t-1} - \psi_{ij}^t)^2\}$$

Then, to arrive at an expression that minimizes the risk $R(\alpha)$ with respect to α, the first two derivatives must be calculated:

$$R'(\alpha^t) = 2\sum_{i=1}^{n}\sum_{j=1}^{n}\{(\alpha^t-1)\,\mathrm{var}(w_{ij}^t) + \alpha^t(\hat{\psi}_{ij}^{t-1} - \psi_{ij}^t)^2\}$$

$$R^{''}(\alpha^t) \geq 0 \text{ for all } \alpha^t$$

In this case, alpha, the forgetting factor, is set automatically as a result of minimizing the expected loss. By setting $\mathrm{R}(\alpha)^{''} = 0$ and rearranging the resulting equation, we arrive at an expression for the optimal choice of the shrinkage parameter alpha [52]:

$$(\alpha^t)^* = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{var}(w_{ij}^t)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\{(\hat{\psi}_{ij}^{t-1} - \psi_{ij}^t)^2 + \mathrm{var}(w_{ij}^t)\}}$$

TRITA -EECS-EX-2019:843