

COLUMBIA UNIVERSITY  
GRADUATE SCHOOL OF BUSINESS

MARKOV CLUSTERING ON PERSON-TO-PERSON SIMILARITY GRAPH  
ATTRIBUTION OF MOVIES' BOX OFFICE RESULTS TO PREFERENCES OF  
VIEWER COMMUNITIES

A Thesis by

Yegor (Igor) Tkachenko

Department of Marketing

Submitted in partial fulfillment of the requirements for the degree of

Master of Science

in Marketing

Advisor:

Professor Kamel Jedidi

September 15, 2013

## **ABSTRACT**

Search for methods of deriving actionable marketing segmentation has a long history in the marketing literature. This work proposes the use of Markov clustering algorithm on person-to-person similarity graph, where similarity between individuals is based on their similarity in rating assignments. This allows the detection of taste-based communities of users. Simple regression analysis is subsequently applied to detect the dependencies of box office results of movies of various genres on the preferences of specific viewer communities. The resulting analysis permitted identification of communities that drive box office results of specific movie genres.

**Keywords:** marketing segmentation, collaborative filtering, Markov clustering, movie ratings, box office results.

## TABLE OF CONTENTS

LIST OF TABLES AND FIGURES .....	4
1. INTRODUCTION .....	5
2. LITERATURE REVIEW AND METHODOLOGY	
2.1. RATINGS DATA AND PERSON-TO-PERSON SIMILARITY .....	7
2.2. GRAPH REPRESENTATION OF ADJACENCY MATRIX AND MARKOV CLUSTERING	
2.2.1. Graph Representation .....	8
2.2.2. Markov Clustering Algorithm .....	9
2.3. THE EFFECTS OF VIEWER COMMUNITIES' PREFERENCES ON MOVIES' BOX OFFICE .....	12
3. RESULTS	
3.1. DATA & SOFTWARE	
3.1.1. Data .....	14
3.1.2. Software .....	14
3.2. IMPLEMENTATION .....	14
4. CONCLUSIONS .....	21
5. REFERENCES .....	21

## LIST OF TABLES AND FIGURES

### Tables

Table 1. Dependency of segment granularity on MCL inflation parameter ( $r$ ) .....	15
Table 2. MCL cluster composition .....	17
Table 3. MCL cluster description by age and location of their members .....	17
Table 4. Results of attribution regression analysis for selected movie genres .....	20

### Figures

Figure 1. Markov clustering process .....	10
Figure 2. Dependency of mean squared error on the number of retained edges ( $m$ ).....	15
Figure 3. Dependency of segment granularity on MCL inflation parameter ( $r$ ) .....	15
Figure 4. MCL segmentation results .....	16

## 1. INTRODUCTION

Within every field of scientific enquiry there is some issue, which is at the center of experts' undivided attention, but solution to which is never perfect enough. In marketing such a problem is market segmentation.

Fundamentally, market segmentation is a process of subdividing a market into segments, which are homogenous with respect to needs/demand characteristics/consumption patterns etc. of consumers.

A natural, most basic counterpart of segmentation in statistics and machine learning is clustering – application of various statistical and computational techniques to detect groups of observations, where observations are relatively close together, when they belong to one group, and relatively distanced, when they belong to different groups, in the hyper-dimensional space formed by various variables recorded for each observation.

Once a matrix consisting of variables characterizing observations (clients) has been obtained, numerous techniques can be applied: k-means clustering, k-nearest neighbors clustering, hierarchical clustering and others, which would yield a particular split of observations into groups.

The main issue with conventional segmentation techniques, however, is not with the statistical techniques themselves, but with the data such techniques tend to be applied to.

The typical data available to marketers on their consumers consists of demographic information (sex, age, occupation, nationality, income, location) and some history of transactions with the company. In some markets such information is sufficient to group consumers into meaningful segments, based on which efficient marketing campaigns can be carried out. More often, though, the resulting segmentation is not granular enough.

Simply put, the problem is that even if 2 individuals are completely the same in terms of the variables the researcher possesses to describe them they might still behave differently and be of different importance to the company. [1].

Movie market could serve as an illustration of the issues a market researcher frequently faces. If we take a group of middle-income white men 25 – 30 years old, we will find that different members of this group can have cardinally opposite tastes when it comes to movies. For example, some of them might prefer drama movies, others – horror. Therefore, some of the members of the noted group could be critical to a horror movie success, whereas others might have no impact on its box office.

Thus, the great dilemma for a marketer is that all the data he might possess on movie viewers might not provide him with a good-enough criteria for actionable segmentation, that is, segmentation, which would allow him to identify segments driving box office of various movie-types.

It seems, however, that the issue might have a solution thanks to the new kind of data being collected online through various e-commerce and review websites, namely, ratings data reflecting the tastes of individual consumers. Such data can be used to estimate the similarity between users, which, in turn, can serve as a fertile basis for user segmentation, as will be argued below.

The rest of the thesis is dedicated to description of methodology for eliciting meaningful users segments based on person-to-person similarity data. A time-efficient and noise-resistant Markov clustering algorithm, previously successfully applied to proteins' clustering, is proposed. Finally, a simple regression method for detection of communities, whose preferences affect box office of movies of particular genres, is implemented.

## 2. LITERATURE REVIEW AND METHODOLOGY

### 2.1. RATINGS DATA AND PERSON-TO-PERSON SIMILARITY

Use of rating data to estimate similarity between rated objects and/or between individuals assigning the ratings - in order to subsequently use such data to produce recommendations - has become a growing field of research in machine learning since late 90s, when large recommender systems, such as Amazon, Yahoo! Music, and Netflix, came to being. [2].

Though the primary purpose of similarity calculations is typically the recommendations themselves, it is the by-product of the recommendation process – the resulting similarity matrix between individuals (in case of person-to-person similarity-based recommendations) – that holds great potential for market segmentation.

The similarity between 2 customers, A & B, can be measured in a variety of ways; however, the cosine of the angle between 2 vectors (where each vector corresponds to a customer) has been the most widely accepted measure. [3].

$$similarity(\vec{A}, \vec{B}) = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \bullet \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|}$$

The noted measure is quite similar to correlation; however, when correlation is computed, each of the respective vectors gets additionally normalized to have the mean of zero.

One of the consequences of such difference is that cosine similarity measure is always larger or equal to zero if the entries of the compared vectors are non-negative. This is not true for correlation measure.

This distinction is not crucial for the algorithms producing recommendations (which are beyond the scope of this work). However, it is a useful property that the adjacency matrix resulting from application of cosine similarity calculation to ratings data is always non-negative (assuming that ratings are non-negative). This property will be utilized during application of the proposed clustering method in the next section.

The content of the vectors is a separate very important issue.

Let us consider a matrix of  $k$  movies and  $n$  individuals, where each column contains the ratings assigned by a particular individual to movies he had rated. The missing values are represented with zeros. One possibility to compute similarity between 2 individuals would be to apply the cosine similarity measure to 2 whole vectors with ratings by the first and the second individuals respectively.

The problem, however, is that in the case, when there are missing ratings, the hypothetical situation, in which one individual watched a particular movie and gave it the maximum rating of 5, whereas the other individual didn't watch the movie (and thus his rating for that movie is equal to zero), would result in great implied difference between individuals, as estimated by cosine similarity.

Though one could argue that such situation should indeed be interpreted as signifying dissimilarity between individuals, it could also be the case that if the second individual had watched the movie, he would have given it the highest score too. Unfortunately, there is no way to know what situation precisely we are in.

The solution to this problem is to calculate the cosine similarity only based on ratings for the objects rated by both individuals. That is, in calculation of similarity between every 2 individuals only the set of movies they both watched and rated would be considered. Empirically, such approach has been shown to yield stable results. [3,4].

Iterative application of this calculation to every pair of individuals yields an  $n \times n$  person-to-person adjacency matrix  $S$ .

It is the argument in the foundation of this work that the noted adjacency matrix can serve as a basis for true taste-based segmentation of individuals, as the cosine similarity computed for each pair of individuals represents precisely how close the tastes of the 2 individuals are, given their rating patterns.

Thus, if one could arrive at a good segmentation based on such a matrix, one would possess information on the true "taste" communities within the population.

## **2.2. GRAPH REPRESENTATION OF ADJACENCY MATRIX AND MARKOV CLUSTERING**

### **2.2.1. Graph Representation**

Once person-to-person adjacency matrix has been obtained, an important decision to be made remains – how one shall retrieve the clusters. One way is to consider individuals



as both observations (rows) and variables (columns) and then proceed with application of conventional clustering techniques.

There are 3 issues with such an approach. First, if the number of users is large, the researcher faces the curse of dimensionality. That is, all users are very far apart from each other in the studied hyper-dimensional space. [4,5].

Second, frequently missing values are an issue, so the matrix can be rather sparse. This introduces extra noise and may pose a challenge for the orthodox techniques. In particular, it has been noted by many researchers that such a conventional clustering technique as k-means tends to create clusters centering on outliers, thus possibly yielding a sub-optimal clustering. [6].

Third, many of the traditional techniques are not computationally efficient when dealing with large volumes of data. [5,7,8].

One way out is to consider the graph representation of the adjacency matrix, where users are represented as nodes, whereas the similarity between them takes the form of the graph edges. When the data on similarity between 2 particular users is not available, the respective edge is simply not added to the graph.

Graph representation makes the dimensionality issue go away, as we no longer consider the framework where observations are placed within high-dimensional space, but instead we work with a simple, easily manageable graph structure.

Careful consideration, however, is necessary to select the graph clustering method, which would prove noise-resistant and computationally efficient.

### **2.2.2. Markov Clustering Algorithm**

The above discussion motivates the proposal of the novel Markov clustering (MCL) algorithm for clustering of the person-to-person similarity graph.

Markov clustering algorithm, which is based on stochastic modeling of flows on the graph, has been successfully applied to protein cluster detection and has been found to deliver rather stable results, while being very computationally efficient. [9,10].

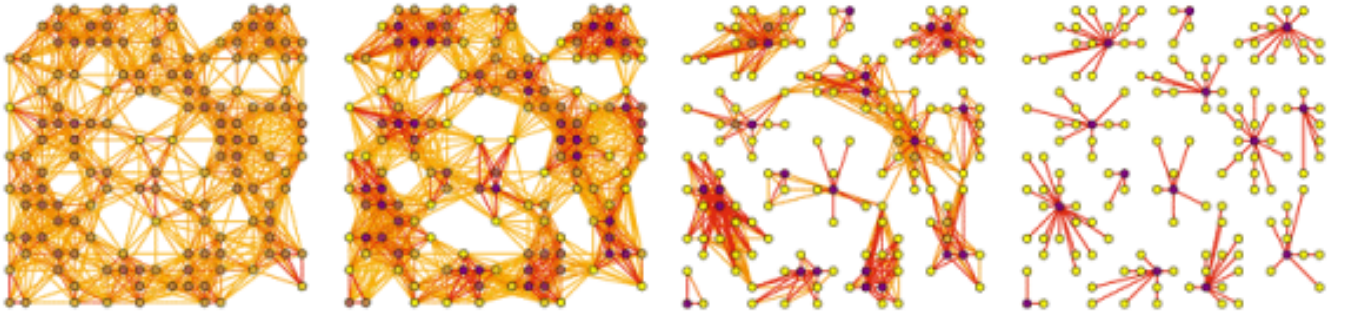
It has been shown that it is much more noise-resistant than affinity propagation – a similar flow-modulation-based graph clustering algorithm. Moreover, it has been shown to be more time efficient compared to such techniques as DBSCAN or spectral clustering, when applied to high-dimensional arrays. [7,11].

2 additional important advantages of MCL, which make this method especially attractive, is that (a) it does not require the researcher to pre-specify the number of clusters, and (b) it is able to detect outliers by identifying so-called singletons – clusters, consisting of a single element. [9,10].

The former property is very useful, as researchers rarely possess any preliminary knowledge of the exact number of segments within a given population. The latter property allows a researcher to account for naturally occurring outliers – individuals, whose taste are unlike tastes of others.

The broad idea behind MCL is that there should be more links with greater weights within clusters and fewer – between clusters. Thus, if one were to start at a node and then randomly travel to a connected node, one would be more likely to stay within a cluster than travel outside of it. By doing random walks on a graph MCL detects where the flow “gathers” and identifies such areas as clusters, strengthening the links within cluster and setting the links between clusters to zero.

**Figure 1. Markov clustering process. [9].**



The mechanics of MCL involves several simple steps.

At the initial stage Markov clustering requires the transformation of the adjacency matrix into stochastic matrix, where each row is normalized for its elements to add up to 1 – thus representing the transition probabilities from one state to every other state.

Formally, for an  $n \times n$  matrix  $S$ , where  $s_{ij}$  is its element:

$$s_{ij}^{normalized} = s_{ij} / \sum_{j=1}^n s_{ij}$$

This, of course, demands that the edges of the graph be larger or equal to zero, and this is precisely where cosine similarity measure plays out particularly well, as it is

guaranteed to produce scores larger or equal to zero as long as the values it is applied to are non-negative, which is exactly the situation with movie ratings in question.

However, one modification to adjacency matrix before doing the normalization is necessary for Markov clustering to yield proper results (not producing one huge cluster). This modification is an artificial increase of matrix's sparsity, as advised by method's author Van Dongen. More specifically, only  $m$ -largest edges for transition from each particular node (that is,  $m$  largest edges in each row) must be kept. [9].

The value of  $m$  is chosen based on how many neighboring nodes allow for the most accurate users' characteristics estimation. For each node  $m$ -closest neighboring nodes are selected; then, a particular metric associated with the nodes is picked, and the average of the neighbors' metrics is taken (weighted by the edges between the node in question and its neighbors). This is done for each node of the similarity matrix. Such  $m$  is selected, which minimizes the MSE (mean squared error) between the estimated ( $\hat{Y}$ ) and the true ( $Y$ ) values of the explored metric.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

At the second stage, once the unnecessary edges have been set to zero, and the stochastic matrix has been obtained through row-wise normalization, the MCL boils down to iteration of 2 steps: *expansion* and *inflation*.

*Expansion* represents a step in a Markov chain, that is, taking the stochastic matrix to a power  $e$  (multiplying it by itself  $e$  times). Informally, we allow the flow to proceed along the edges of the studied graph  $e$  times, retrieving the probabilities of ending up at each of the nodes.

For matrix  $S$ :

$$S^{expanded} = S^e = \{S * \dots * S\} \text{ } e \text{ times}$$

*Inflation* constitutes taking each element of the matrix produced at the expansion stage to a power  $r$  ( $>1$ ), and then re-normalizing the matrix row-wise, for the elements in each row to add up to 1. The effect of this procedure is that the strong edges get stronger, whereas the weaker edges decline, until after several iterations they become equal zero.

$$s_{ij}^{inflated} = (s_{ij}^{expanded})^r / \sum_{j=1}^n (s_{ij}^{expanded})^r$$

Expansion parameter  $e$  and inflation parameter  $r$  are thus the two main tuning parameters for MCL.

Expansion parameter is frequently set to 2 by convention, as empirically its changes above 2 do not tend to lead to any significant alterations in the resulting segment structure. [9,10,11].

Inflation parameter  $r$ , however, plays a very important role, as it directly controls granularity of the segmentation. Its values typically lie in the range from 2 to 10, and the larger its value, the more granular and numerous segments one would expect as a result. So far, there has been no literature on optimality of the parameter, and the choice is often arbitrary, with researcher's satisfaction with the resulting segment structure often being the only criterion in place. [9,10,11].

The resulting segments have a star-like structure with a single attractor node at the center of each segment, and all other nodes connected to the attractor, but not to each other. As it has already been mentioned, some resulting clusters may contain only one element.

Overlapping clusters (with several attractors) have been reported as a very rare phenomenon occurring solely in graphs with particular symmetry properties.

Finally, it is important to note that the convergence of the algorithm has not been proven in the seminal work, which introduced MCL, but empirically, it has been observed to occur in most of the cases. [9].

### **2.3. THE EFFECTS OF VIEWER COMMUNITIES' PREFERENCES ON MOVIES' BOX OFFICE**

It has been shown in the literature that ratings can be an important factor in predicting movies' box office success. However, to the best knowledge of the author, there has so far been no attempt to estimate how the preferences of particular viewer segments affect the box office results of movies of different genres, where ratings are considered to be indicative of preferences. [12].

Once the clusters have been retrieved, one needs to devise the way of attributing box office to preferences of particular communities.

First, select the communities of interest. Given that segments can be rather granular and numerous, a researcher should be interested in the communities large enough to affect

the movie box office. For this purpose a threshold of a certain number of users in a community could be used.

Second, for each of  $k$  movies in the dataset calculate the average score assigned to the respective movie by each of the communities. If not a single person in the community watched a particular movie, the respective rating is set to zero.

Third, run a regression with log box office of movies of a particular genre as a dependent variable, and ratings for such movies by the communities as independent variables.

The following model is estimated for movies belonging to each genre of interest.

$$\log(y_i) = \beta_0 + \beta_1 x_i^1 + \dots + \beta_k x_i^k + \varepsilon_i$$

Where  $y_i$  is box office of movie  $i$ , and  $x_i^k$  is a mean rating assigned to movie  $i$  by cluster  $k$ .

Communities, whose ratings have a significant positive effect on the box office data for a particular movie genre, would be deemed to be drivers of such movies' box office success.

Identifying such communities would allow movie producers to target specifically the tastes of the audiences, which drive box office of the movies of genre in question.

Such approach could as well be extended beyond movie industry, to any field where both ratings and sales data are available, such as fashion, consumer electronics and other areas within e-commerce world.

### **3. RESULTS**

#### **3.1. DATA & SOFTWARE**

##### **3.1.1. Data**

The data consists of 371,792 ratings assigned by 5,905 users to 890 movies (all produced before or in 2000), with at least 20 ratings per user. The data originates from GroupLens Movie Dataset collection, which is a frequently used source for recommender system research. [13].

For each of the movies its genre as well as the US box office (scraped from Internet Movie Database – IMDB on August 15, 2013) is available. [14].

User data includes sex, age and zip code variables.

##### **3.1.2. Software**

All analyses were run using Python, R and Gephi (network visualization software). [15, 16, 17].

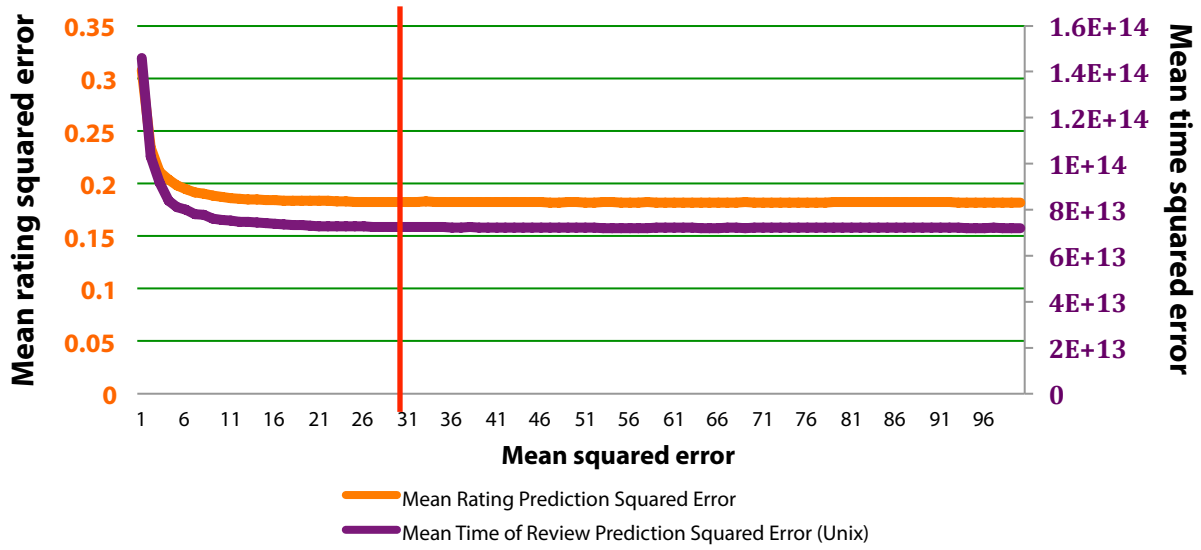
#### **3.2. IMPLEMENTATION**

We begin with the ratings data and compute the  $5,905 \times 5,905$  person-to-person similarity matrix. The missing values (when individuals have not rated a single same movie) are set to zero.

Before applying Markov clustering algorithm, we transform the matrix, keeping only  $m$  largest entries in each row.

To estimate the  $m$  we run experiments on the graph to discover the optimal number of neighboring nodes accurately predicting various metrics of the node in question. In particular, we test the mean squared error of the prediction of average time of review and the average rating for each person, based on same data of its  $m$  neighbors.

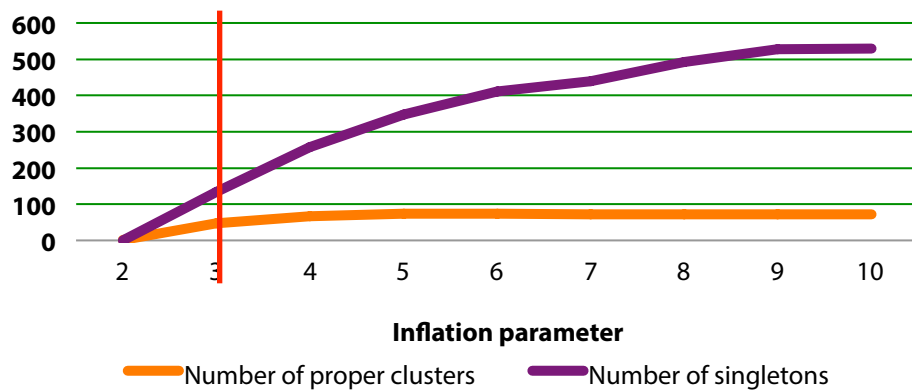
**Figure 2. Dependency of mean squared error on the number of retained edges ( $m$ ).**



Based on the plot above, we pick  $m$  equal to 30, as with this value the size of both kinds of prediction errors reaches the minimum and remains stable as  $m$  grows.

The next step is application of MCL to the resulting matrix. We set the expansion parameter  $e$  to 2, which is a conventional choice, as noted before. In order to determine the optimal inflation parameter experiments are run and the segment structures for inflation parameter values ranging from 2 to 10 are retrieved.

**Figure 3. Dependency of segment granularity on MCL inflation parameter ( $r$ ).**



**Table 1. Dependency of segment granularity on MCL inflation parameter ( $r$ ).**

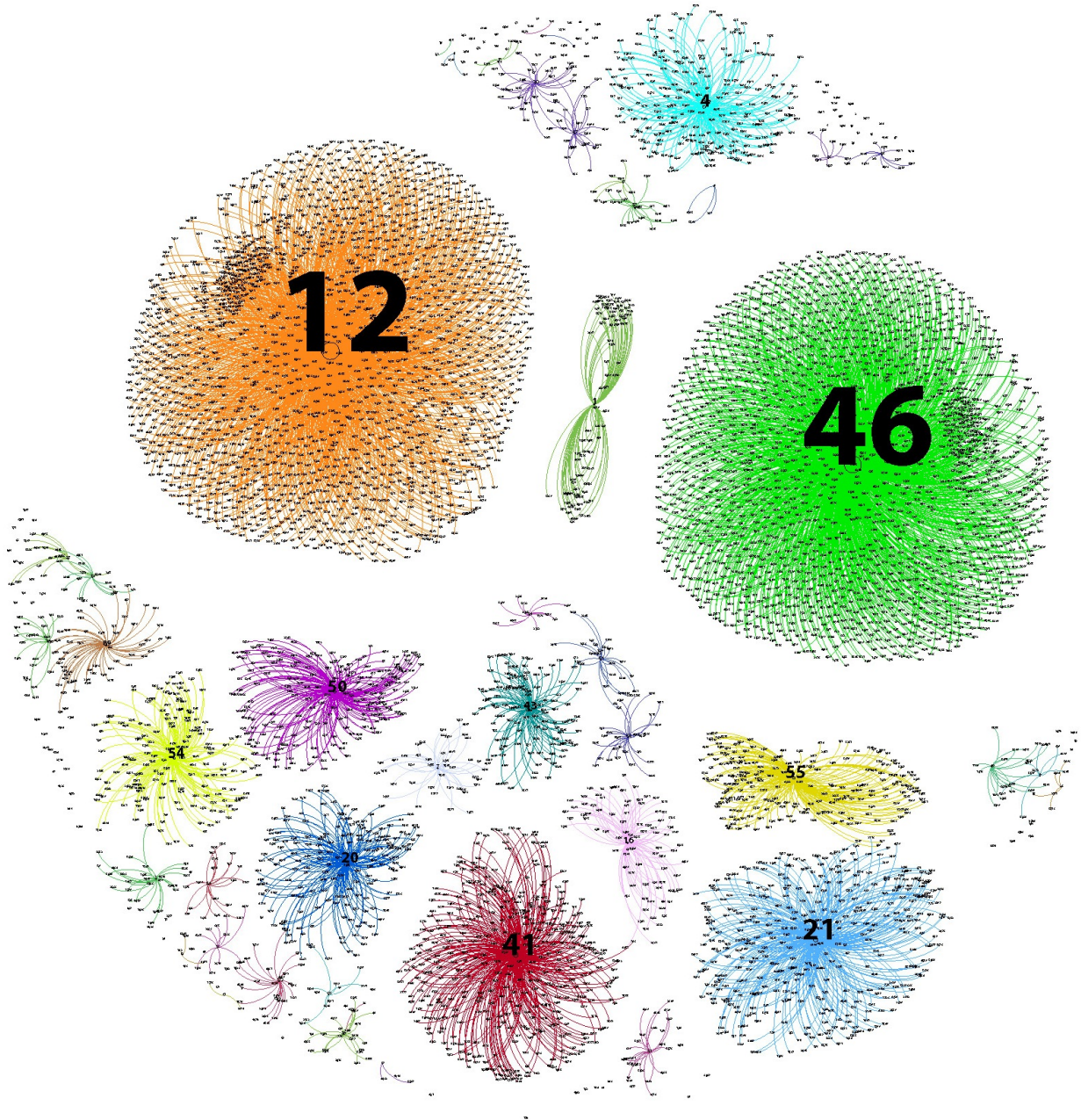
Inflation parameter	2	3	4	5	6	7	8	9	10
Number of proper clusters	1	47	67	73	73	72	72	72	72
Number of singletons	0	133	257	347	412	440	493	528	529
Total	1	180	324	420	485	512	565	600	601



We pick the inflation parameter  $r$  equal to 3, which yields the minimum number of singletons (180), while producing more than 1 cluster (that is, an *interesting* segment structure).

The resulting clustering can be seen in Figure 4.

**Figure 4. MCL segmentation results.**



Each of the segment labels represents the id of the user, whose node played the role of the attractor for its respective segment.

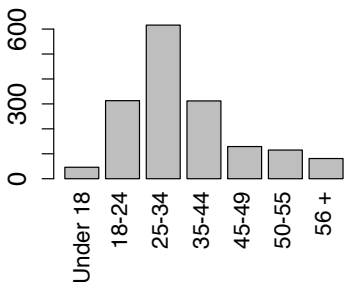
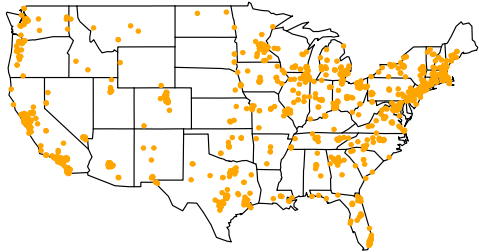
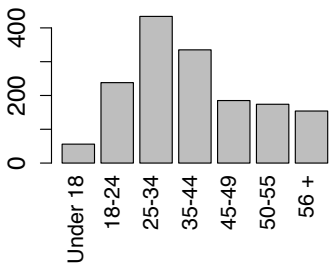
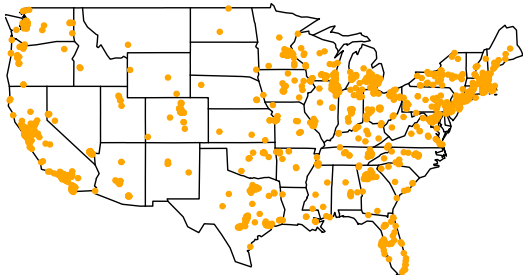


**Table 2. MCL cluster composition.**

Cluster ID	# of individuals	Size as % of population	Proportion of males
X12	1612	27%	72%
X46	1576	27%	65%
X41	400	7%	81%
X21	383	6%	69%
X4	219	4%	77%
X55	212	4%	78%
X50	200	3%	77%
X20	178	3%	78%
X54	172	3%	69%
X43	119	2%	69%
...	701	12%	-
Singletons	187	3%	-

For the purpose of further analysis we only keep 10 largest segments (the ones with more than 100 individuals). The description of these clusters by age and zip code distributions of the individuals can be seen below (Table 3).

**Table 3. MCL cluster description by age and location of their members.**

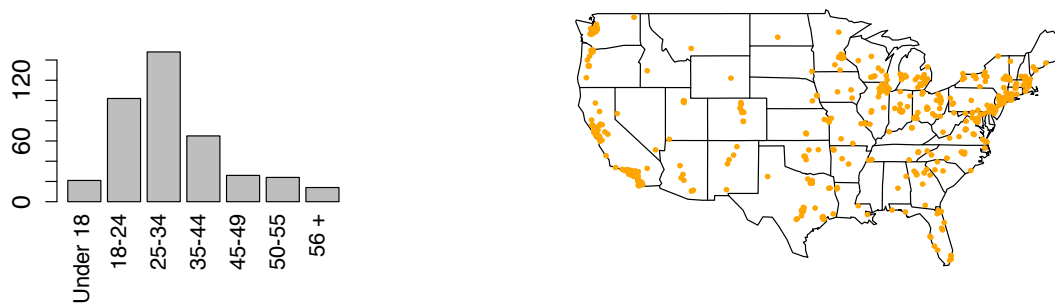
Age distribution	Zip code distribution
<b>X12</b>	
	
<b>X46</b>	
	

---

---

**X41**

---

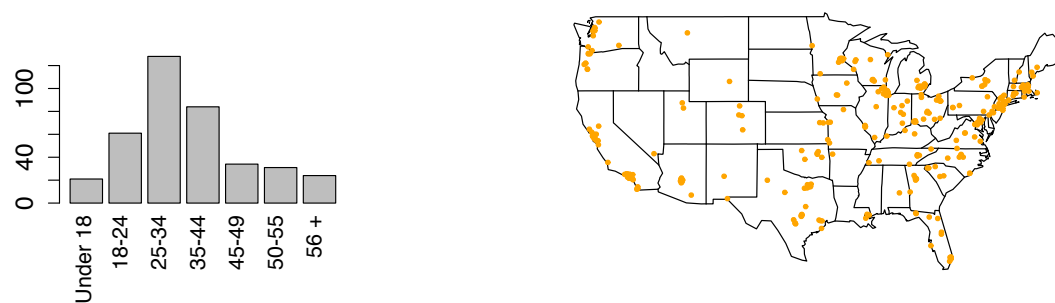


---

---

**X21**

---

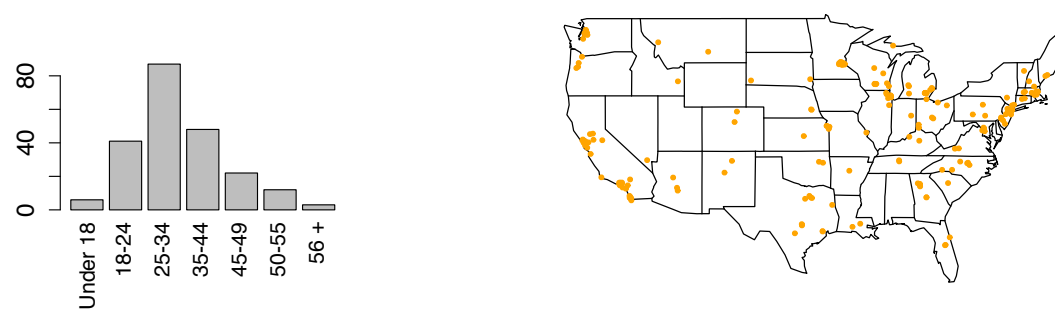


---

---

**X4**

---



---

---

**X55**

---

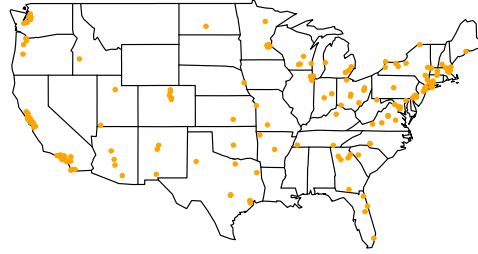
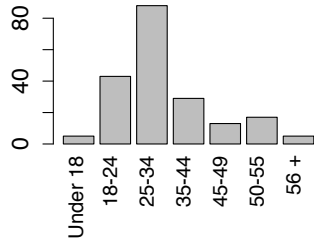


---

---

**X50**

---

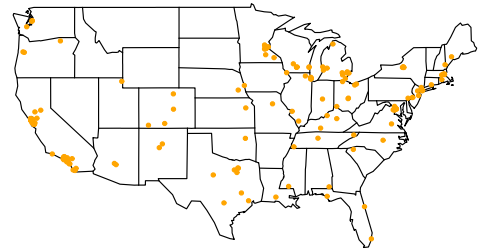
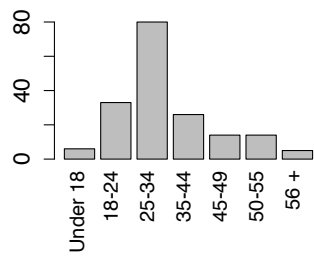


---

---

**X20**

---

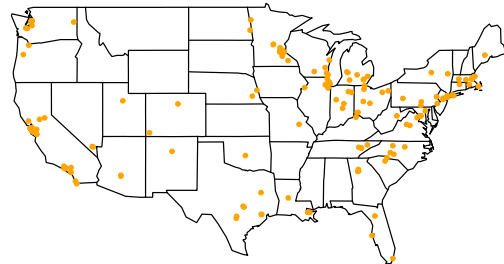
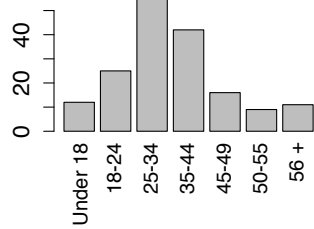


---

---

**X54**

---

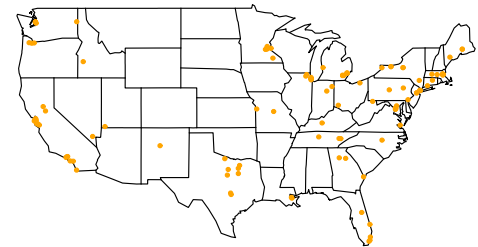
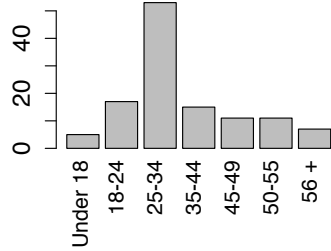


---

---

**X43**

---



As described before, for each of the selected 10 largest segments the mean rating assigned to each of the 890 movies is calculated. When no one in the segment has rated a particular movie, the rating for this movie by the respective segment is set to zero.

Finally, genre-specific regression models are estimated, where the log box office of movies of selected genres (Drama, Action, Animation and Musical) is a dependent variable, and the segment-specific ratings calculated as described above are the independent variables. The results are presented in Table 4.

**Table 4. Results of attribution regression analysis for selected movie genres.**

<b>Drama</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>		<b>Animation</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>	
(Intercept)	15.66605	0.55468	28.244	2.00E-16	***	(Intercept)	23.95292	3.09205	7.747	0.0015	**
X12	-0.40231	0.30373	-1.325	0.18619		X12	-5.52105	2.06633	-2.672	0.0557	.
X46	0.19873	0.14316	1.388	0.166		X46	-4.90811	2.33063	-2.106	0.103	
X41	-0.11781	0.14602	-0.807	0.42035		X41	3.32526	2.43107	1.368	0.2432	
X21	-0.04073	0.12543	-0.325	0.74558		<b>X21</b>	<b>4.55132</b>	<b>1.18888</b>	<b>3.828</b>	<b>0.0186</b>	*
X4	-0.17401	0.19512	-0.892	0.3731		X4	-2.28833	1.50187	-1.524	0.2023	
X55	0.08923	0.14072	0.634	0.52643		X55	-0.41672	1.38361	-0.301	0.7783	
<b>X50</b>	<b>0.2459</b>	<b>0.09074</b>	<b>2.71</b>	<b>0.00706</b>	**	X50	-0.15782	1.27504	-0.124	0.9075	
X20	0.07129	0.08308	0.858	0.39143		X20	2.89325	2.41631	1.197	0.2973	
<b>X54</b>	<b>0.17714</b>	<b>0.07694</b>	<b>2.302</b>	<b>0.02191</b>	*	X54	0.76122	1.65583	0.46	0.6696	
<b>X43</b>	<b>0.21379</b>	<b>0.06497</b>	<b>3.291</b>	<b>0.0011</b>	**	X43	0.09452	0.93898	0.101	0.9247	
R-squared					0.1334	R-squared					0.9091
<b>Action</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>		<b>Musical</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z-value</b>	<b>Pr(&gt; z )</b>	
(Intercept)	16.36612	0.47211	34.666	2.00E-16	***	(Intercept)	15.784	2.2112	7.138	7.61E-06	***
X12	-0.7622	0.49631	-1.536	0.12663		X12	-1.8524	1.8631	-0.994	0.3382	
X46	-0.31052	0.25896	-1.199	0.23231		X46	-2.6415	1.5058	-1.754	0.1029	
<b>X41</b>	<b>0.60439</b>	<b>0.33739</b>	<b>1.791</b>	<b>0.07518</b>	.	X41	-0.4261	1.0661	-0.4	0.6959	
X21	-0.41133	0.26174	-1.572	0.11808		X21	0.4771	1.3084	0.365	0.7212	
<b>X4</b>	<b>0.52431</b>	<b>0.27614</b>	<b>1.899</b>	<b>0.05945</b>	.	X4	-1.0117	0.7922	-1.277	0.2239	
X55	-0.34452	0.25113	-1.372	0.17207		X55	-0.3242	1.1385	-0.285	0.7803	
<b>X50</b>	<b>0.66538</b>	<b>0.31622</b>	<b>2.104</b>	<b>0.03697</b>	*	<b>X50</b>	<b>1.7221</b>	<b>0.7958</b>	<b>2.164</b>	<b>0.0497</b>	*
X20	-0.08725	0.2201	-0.396	0.69235		X20	0.2598	0.9958	0.261	0.7983	
X54	0.1218	0.16716	0.729	0.46731		<b>X54</b>	<b>1.5667</b>	<b>0.6171</b>	<b>2.539</b>	<b>0.0247</b>	*
<b>X43</b>	<b>0.39708</b>	<b>0.11933</b>	<b>3.328</b>	<b>0.00109</b>	**	<b>X43</b>	<b>2.7132</b>	<b>0.9194</b>	<b>2.951</b>	<b>0.0113</b>	*
R-squared					0.2594	R-squared					0.6665

The emerging patterns provide a lot of food for thought. For example, higher ratings by communities X50 and X43 tend to lead to significantly higher box-office across Drama, Action and Musical genres, which might indicate that the movie producers should attempt to target the tastes of these communities to ensure the success of the movie in one of the specified mass-market genres.

However, in case of the Animation genre the tastes of neither of these communities have any impact on movies' box office. Rather, it is community X21, which seems to be driving the box office of such movies.

The fact that the data explains 90% of box office variance in case of animation movies is particularly surprising, given that the same data explains only a small portion of variance in drama and action movies' box office.

Fundamentally, this might mean that the monetary success of the animation movies depends on their reception by a taste-based community X21, which constitutes only around 6% of the total population. Such information could be potentially invaluable to producers attempting to tailor the movie to the tastes of the future viewers.

Moreover, the fact that the age distribution of community X21 is not significantly different from the age distribution of other communities serves as an indication that the traditional approach to building segmentation using solely demographic data would most likely fail, as it wouldn't be able to differentiate between individuals, whose tastes are of varying importance to animation movies' box office.

## 4. CONCLUSIONS

Development of actionable segmentation has been one of the priorities of marketing research as a science. This work investigates one of the methods for creating such segmentation by applying Markov clustering algorithm to person-to-person similarity graph, constructed based on movie rating data. Such approach produces granular taste-based segmentation of consumers. Application of simple regression analysis reveals that preferences of some of the resulting communities can be identified as drivers of box office results for specific movie genres.

## 5. REFERENCES

- [1]. McDonald, Malcolm. **Why so much segmentation is rubbish**. Research-live. March 2011. <<http://www.research-live.com/comment/why-so-much-segmentation-is-rubbish/4004672.article>>. Retrieved on September 13, 2013.
- [2]. Takacs, Gabor et al. **Scalable Collaborative Filtering Approaches for Large Recommender Systems**. Journal of Machine Learning Research. March 2009.
- [3]. Linden, Smith, and York. **Amazon.com Recommendations: Item-to-Item Collaborative Filtering**. 2003.
- [4]. Leskovec, Jure. **Recommender Systems: Content-based Systems and Collaborative Filtering**. CS246: Mining Massive Datasets Slides. 2013.
- [5]. Steinbach, Michael; Ertöz, Levent; Kumar, Vipin. **The Challenges of Clustering High-Dimensional Data**. In L. T. Wille, editor, New Vistas in Statistical Physics – Applications in Econophysics, Bioinformatics, and Pattern Recognition. Springer-Verlag. 2003.
- [6]. Hautamaki, Ville et al. **Improving K-Means by Outlier Removal**. SCIA 2005.

- [7]. Viswanath, P.; Babu, V. Suresh. **Rough-DBSCAN: A fast hybrid density based clustering method for large data sets**. Pattern Recognition Letters. August 2009.
- [8]. Schaeffer, Satu Elisa. **Graph clustering. Survey**. Computer Science Review. August 2007.
- [9]. Van Dongen, S. **Graph Clustering by Flow Simulation**. PhD Thesis, University of Utrecht, The Netherlands. 2000.
- [10]. Stuluri, Venu; Parthasarathy, Srinivasan; Ucar, Duygu. **Markov Clustering of Protein Interaction Networks with Improved Balance and Scalability**. BCB 2010. Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology.
- [11]. Vlasblom, James; Wodak, Shoshana J. **Markov clustering versus affinity propagation for the partitioning of protein interaction graphs**. BMC Bioinformatics. March 2009.
- [12]. Terry, Neil; Butler, Michael; De'Armond, De'Arno. **The Determinants of Domestic Box Office Performance in the Motion Picture Industry**. Southwestern Economic Review. 2005.
- [13]. GroupLens Research. **MovieLens 1M Dataset**. <<http://www.grouplens.org/node/12>>. Retrieved on June 2013.
- [14]. The Internet Movie Database (IMDB). **US box office data**. <<http://www.imdb.com>>.
- [15]. R Core Team (2012). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. <<http://www.R-project.org/>>
- [16]. G. van Rossum and F.L. Drake (eds). **Python Reference Manual**. PythonLabs. USA, 2001. <<http://www.python.org>>
- [17]. Bastian M., Heymann S., Jacomy M. (2009). **Gephi: an open source software for exploring and manipulating networks**. International AAAI Conference on Weblogs and Social Media.