Clustering users of public access venues

Analysis results featuring the Global Impact Study

Lucas Koepke

2014

TECHNOLOGY & SOCIAL CHANGE GROUP (TASCHA)

The Technology & Social Change Group (TASCHA) at the University of Washington Information School explores the design, use, and effects of information and communication technologies in communities facing social and economic challenges. With experience in 50 countries, TASCHA brings together a multidisciplinary network of social scientists, engineers, and development practitioners to conduct research, advance knowledge, create public resources, and improve policy and program design. Our purpose? To spark innovation and opportunities for those who need it most.

CONTACT

Technology & Social Change Group
University of Washington Information School
Box 354985
Seattle, WA 98195

Telephone: +1.206.616.9101
Email: tascha@uw.edu
Web: ***tascha.uw.edu***

GLOBAL IMPACT STUDY

The Global Impact Study of Public Access to Information & Communication Technologies was a five-year project (2007-2012) to generate evidence about the scale, character, and impacts of public access to information and communication technologies. Looking at libraries, telecenters, and cybercafes, the study investigated impact in a number of areas, including Communications & Leisure, Culture & Language, Education, Employment & Income, Governance, and Health. The Global Impact Study was implemented by the University of Washington's Technology & Social Change Group with support from the International Development Research Centre (IDRC) and a grant to IDRC from the Bill & Melinda Gates Foundation.

Learn more at **tascha.uw.edu/projects/global-impact-study/**.

ABOUT THE AUTHOR

Lucas Koepke is a Data Analyst for the Technology & Social Change Group, providing analytical, data processing, and statistical support for a variety of projects. Recent work includes cluster analysis for the Global Impact Study, data mining mobile phone logs, and logistic regression analysis of survey data from Bermuda. Research interests include analyzing large data sets efficiently, optimizing software algorithms for computational speed, and GPGPU computing. Lucas holds a Master of Science in Statistics from the University of Washington, with bachelors degrees in Mathematics and Germanic Studies.

COPYRIGHT, LICENCING, DISCLAIMER

ABSTRACT

This paper examines the application and subsequent findings from the use of cluster analysis on data from the Global Impact Study. Specifically focusing on the user survey data, the results successfully show that complex and interesting structure exists in this data. Making connections between features of the data is an exciting and integral part of analysis. Using cluster analysis as an exploratory analysis tool, additional insight is gained into the complicated relationship between usage patterns and perceived impact for users of public access venues. Additional insight is also gained into the ongoing debate about constructive uses of ICTs, and whether gaming and related activities have solely a detrimental effect. These findings are useful in providing insight beyond that gained by tables and crosstabs, and at the very least, suggest areas for discussion or further study.

The data the cluster analysis was performed on and that is referenced in this paper can be found at **tascha.uw.edu/publications/global-impact-study-user-survey-data-csv-format/**.

KEYWORDS

Public access venues, users, cybercafés, internet cafes, libraries, telecenters, internet, ICT, ICTD, ICT4D, latent class analysis, cluster analysis, similarity measure, Global Impact Study

RECOMMENDED CITATION

Koepke, L. (2014). *Clustering users of public access venues: Analysis results featuring the Global Impact Study.* Seattle: Technology & Social Change Group, University of Washington Information School.

Contents

List of Figures

List of Tables

1 Introduction

Frequency tables and other data summaries such as crosstabs are often the first tools of attack when analyzing new survey data. Crosstabs can provide a useful view of two (and up to three) variables at a time, but quite often a feature of interest will span 5, 10, or more variables. Analysis via crosstabs is then limited to manually sifting through all applicable crosstabs to build an overall picture. With so many separate pieces of information, it is difficult to make connections within the results, especially as these connections are an important part of analysis. More complex analysis techniques, such as cluster analysis, can help highlight connections in a larger data set.

The reasons for this are best illustrated with a brief example. Consider two hypothetical variables: 'use of computer to search for information' and 'use of computer for gaming'. Each has two possible responses: 'used' and 'did not use'. Although reporting the percentages responding 'used' for each question gives a general sense of the data, it is far more interesting to connect the variables. The variables **together** give **four** possible response groups: those who used computers for **both** information searches and gaming, those who used only one or the other, and those who used neither. The proportions in each of these four categories is a noticeably more nuanced and informative view of the data. Now discussion can focus on the percent X who engaged in **both** gaming and information searches, versus the percent Y who **only** engaged in gaming, and the demographics of these two user groups can be contrasted. This opens up exciting possibilities in terms of analysis, but a key challenge remains: doing this for more than 2 variables is increasingly (and prohibitively) complex.

In this paper, new results obtained from the application of cluster analysis to data from the Global Impact Study are presented. These results are placed in context with previous findings, and offer additional insight into this complex data set. A full discussion of methods is reserved until Section 6, where interested readers can obtain detailed technical information. Section 7 contains the complete analysis steps to produce the paper results.

2 Background

The goal of the Global Impact Study of Public Access to Information & Communication Technologies was to generate evidence about the scale, character, and impacts of public access to information and communication technologies. This large-scale collection of data spanned 8 countries over 5 years. In this study users of public access venues (such as cybercafés, telecenters, and connected libraries), non-users, and venue operators were all surveyed. Of these three data sets, only the User data is considered here. This data alone contains nearly 500 variables for over 5000 subjects, and was extensively analyzed using frequency tables and crosstabs for the final Global Impact Study project report [Sey et al., 2013]. The aim of this paper is to illustrate how cluster analysis can be applied to this data, and how the results are both interesting and useful.

Two key factors lead to the consideration and use of cluster analysis: the lack of a definitive outcome variable, plus the fact that the survey comprised primarily categorical variables. These made a regression formulation more difficult, due especially to the lack of dependent variable. Furthermore, the overall size of the data means that a lot of variables are potentially interesting, and they relate to each other in many

ways that are not causal. Thus cluster analysis is a good candidate to explore this data on a deeper level than tables and crosstabs.

How to facilitate impact in priority domains such as income, health, and education remains a relevant topic. In an attempt to shed additional light on this issue, this paper explores the relationship between frequency of use and users' perceived impact. As part of this question, do high levels of usage in relevant areas (such as word processing, using email, etc) contribute to such impact? At the same time, do high levels of use for recreational activities such as social networking and gaming have the opposite effect? The challenge faced with using crosstabs here is that it is not just pairs of variables that are connected. High use in word processing, for example, may not by itself be indicative of positive impact in education. Rather it may be that high use in a combination of categories is a better indicator of the level of impact in some (or multiple) areas.

Two-variable crosstabs suggested significant connections between usage patterns and perceived impacts, but initial analysis was hampered by the complexity of possible interactions between these two groups of variables. The 24 variables covering usage and perceived impact[1] still required over 160 tables to cover all possible pairs of variables. Cluster analysis successfully provided additional insight into this complex interplay between usage patterns and users' perceived impact.

What is cluster analysis? Why does it help?

Cluster analysis is a tool to take in a data set of N observations over some number of variables, and return a grouping of those N observations based on how similar their responses are to the variables. To illustrate more visually what this means, consider a simple example of just 3 data points. For each data point there is information from two numeric variables, x and y. The goal here is to use cluster analysis on these three points, and highlight how the results match with intuition.

First, a simple plot of the data is shown in Figure 1(a). This shows how points 2 and 3 are fairly close together, but point 1 is much farther from both of them. This is further illustrated in Figure 1(b), where circles centered at point 3 are drawn out to the other two points: point 1 is much further away than point 2. Using this distance as the basis for an intuitive understanding of "similarity", points 2 and 3 are much more similar to each other than to point 1.

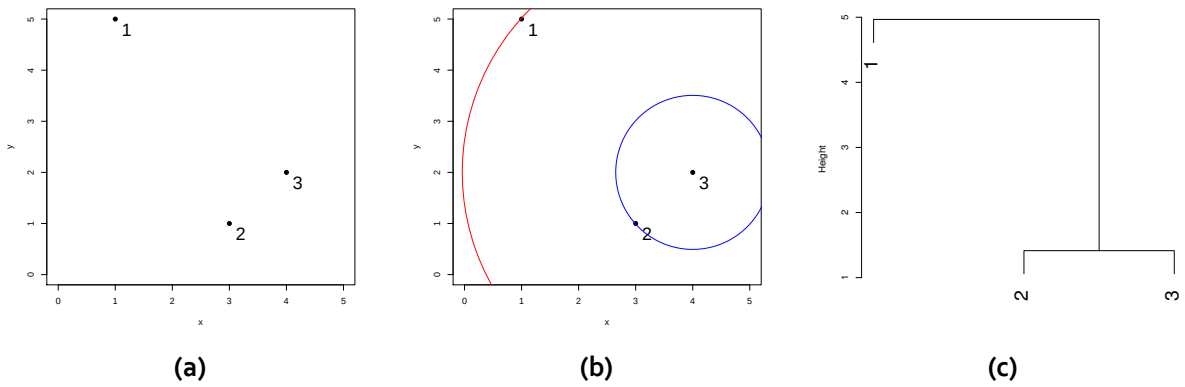
A hierarchical cluster analysis on these three points will show how the clustering results are based on this concept of distance. The hierarchical method used in this paper is an agglomerative ("bottom-up") approach. This starts by taking the data points, and at each step joining the two most similar units (whether two points, two clusters, or a point and a cluster). The process stops when all units are in one large cluster. The output of this is visualized with a "dendrogram" plot, showing which units join and when the join occurred.

The two points closest together in this example are clearly points 2 and 3, and they are joined first by the clustering, Figure 1(c). This is evident from the plot by the low value of "Height" where the horizontal line connecting points 2 and 3 is drawn. To connect point 3 with this new cluster takes a much larger increase in "Height", which matches with the increased actual distance between the points.

Using the information shown in (c), points 2 and 3 form one cluster, leaving point 1 in its own cluster. Building up from this basic example, adding variables will contribute additional components to the overall

[1] 11 covering the frequency of use for specific ICT related tasks (email, social networking, word processing, etc.) and 13 covering the perceived impact from using public access venues for different categories (income, education, health, etc.).

Figure 1: *Overall, this shows two points that are close together (point numbers 2 and 3), and one that is far from both (point 1), shown in (a). In this example it is easy to see these distances, illustrated further in (b) by drawing circles centered at point 3. After clustering these three data points, the resulting dendrogram is shown in (c). This nicely illustrates the basic principles of "height", and how dis-similarity between points (in this case distance) is translated into the "height" in the dendrogram. Points 2 and 3 join quickly at a low height, precisely because they are closer to each other than to point 1.*



level of similarity for different data points. The key message here is that in more complex situations, with perhaps thousands of data points and tens or hundreds of variables, clustering can still produce a useful grouping of the data, and one that we would never be able to see by hand.

3 Results of cluster analysis

Cluster analysis on the 24 usage and perceived impact variables began with data cleaning. After removing observations with missing values in any of the 24 variables, the final sample size was 3585. Furthermore, these 24 were strategically cut to just 14, using the procedure outlined in Section 7. The variables used are listed in Table 1. The cluster analysis procedure detailed in Section 7 produced **six clusters** with a reasonable level of distinction between them.

The clustering itself only produces a new variable showing which respondents are in which cluster. This is then used like any other categorical variable to create crosstabs. Creating tables of key variables within each cluster is useful both for validation (showing that noticeable differences were found) and also to form a concise description of each cluster. As demonstrated shortly, naming the clusters (provided they are interesting to begin with) greatly simplifies subsequent discussion. It allows focus to be directed more at the high-level results and findings, rather than wading through all the relevant details. As such, a short overview of each cluster is presented now, with further discussion to follow.

- **Cluster 1:** *Normal users.* Size = 354
  This cluster shows patterns most similar to the overall population levels, including the general pattern of frequent usage of social networking, surfing, and chatting. Overall, they report high proportions of positive impact in the communications & leisure domain, as well as education. They tend to be younger, but are not dominated by a specific country. The most common venue type is cybercafés.

Table 1: *Variables considered for clustering, from Section 4 of the survey, the frequency of use for technology related tasks, and from Section 5, the perceived impact from using public access venue. Variables used in clustering are marked.*

Variable	Description	Used
Q4_4a	For the venue you usually go to how often do you - email	
Q4_4b	For the venue you usually go to how often do you - chat using IM, VOIP	x
Q4_4c	For the venue you usually go to how often do you - browse, surf internet	
Q4_4d	For the venue you usually go to how often do you - blog	
Q4_4e	For the venue you usually go to how often do you - social network, Facebook, Myspace	x
Q4_4f	For the venue you usually go to how often do you - watch movies or tv online	
Q4_4g	For the venue you usually go to how often do you - play games	x
Q4_4h	For the venue you usually go to how often do you - listen or download music	
Q4_4i	For the venue you usually go to how often do you - read news	x
Q4_4j	For the venue you usually go to how often do you - buy goods online	x
Q4_4k	For the venue you usually go to how often do you - do word processing	x
Q5_1a	Overall impact from using public access venue on - your income	x
Q5_1b	Overall impact from using public access venue on - your access to employability resources	x
Q5_1c	Overall impact from using public access venue on - your education	x
Q5_1d	Overall impact from using public access venue on - your health	x
Q5_1e	Overall impact from using public access venue on - your access to info and services from local and central government	
Q5_1f	Overall impact from using public access venue on - local language & culture	x
Q5_1g	Overall impact from using public access venue on - your time savings	
Q5_1h	Overall impact from using public access venue on - financial savings	x
Q5_1i	Overall impact from using public access venue on - meeting new people	x
Q5_1j	Overall impact from using public access venue on - you maintaining communication with family and friends	
Q5_1k	Overall impact from using public access venue on - you sending or receiving remittances	
Q5_1l	Overall impact from using public access venue on - you pursuing interests and hobbies	x
Q5_1m	Overall impact from using public access venue on - you pursuing other leisure activities	

- **Cluster 2:** *Highly social users*. Size = 514
  This cluster is even more socially active than Cluster 1 (and thus the general population), with a high proportion chatting online frequently, in addition to emailing, surfing, and social networking. They also report even higher proportions of positive impact in communications & leisure. Brazil is barely represented in this cluster.
- **Cluster 3:** *Low-frequency users*. Size = 648
  These users show almost universally low frequency of use in almost all 11 categories. Not even 20% use email most or every time they visit the venue. This group also shows the lowest proportions of positive impact in almost every category, but especially in the communication & leisure domain. Within that domain, only around 35% report a positive impact on maintaining communication, far lower than any other cluster or the overall population. This cluster is noticeably dominated by Bangladesh, at nearly 70% of the 648 users, and a corresponding high proportion of telecenter

users.

- **Cluster 4:** *Social gamers*. Size = 1002
  This group reports the highest frequency of playing games, with around 60% saying they do so most or every time they visit the venue, and almost all the rest saying they sometimes play. However, they also email, chat, surf, and social network with high frequency. Interestingly, there is only a slight decrease in the proportion reporting positive impacts, although there may be a slight increase in proportions reporting negative impacts in some categories (such as time/financial savings and income).
- **Cluster 5:** *Constructive and social users*. Size = 834
  Although users in this group use email, chat, surf, and social network frequently, they also engage in other ways. Over 40% read the news most or every time they visit, and over 50% do word processing with the same frequency. Usage differs from Cluster 6 most noticeably in the areas of gaming, watching movies/TV, and buying goods online, with Cluster 5 showing much lower frequency of use. Increased proportions (compared with Clusters 1-4) report positive impact in key areas such as access to government information & services, local language and culture, and transferring money. However, the numbers are still lower than for Cluster 6. Additionally, 43% of this cluster is female, compared with just 30-35% of the other clusters.
- **Cluster 6:** *Power users*. Size = 233
  These users show, by far, the highest frequency of use in the most categories. From emailing, chatting, and surfing through blogging, gaming, and buying goods online, these are heavy users. They also have high proportions reporting positive impacts in most categories, not just the typical communication & leisure ones, but also less common areas such as culture, transferring money, and health. Although cybercafés still dominate, libraries have a larger showing than telecenters in this cluster (the only cluster to show this). Bangladesh has almost no hold in this group, but the other countries are all represented.

Charts for each cluster are given at the end of this report on pages 24 - 29. These detail the response patterns within each cluster for all the 24 usage and perceived impact variables, and also break down each cluster in terms of country, venue type, gender, and age.

Not only do these clusters show interesting features of the data, but choosing appropriate names for each cluster helps significantly to produce a simple and compelling storyline. Terms like "power users" or "social gamers" immediately provide information about that group just from the name. Once a description of each cluster has been given, it provides a simple and intuitive way of discussing the groups. For example, which of the following is easier to discuss?

1. The group of users that show high usage frequency in most of the usage categories (email, social networking, word processing, etc.) is comprised of a higher proportion of female users than other groups.
2. The "constructive and social users" cluster is comprised of a higher proportion of female users than other groups.

Not only is option 2 a simpler sentence, but the term "constructive and social users" is an easy reminder of the previous cluster definition, and avoids the unnecessary repeated description of the group. This leaves focus on the key point, which is the higher proportion of female users. In a lengthy report, this can be a valuable addition to the narrative.

4 Discussion of findings

The real power of cluster analysis for the User survey is in **connecting** information about each user from numerous variables. In this sense, new findings might take the form of connecting different kinds of usage and/or perceived impacts in ways not yet examined. Perhaps frequent word processing is not directly linked to increased proportions reporting positive impacts in business or education, but when combined with frequent use to obtain news or information the link is strengthened.

While new findings are exciting, obtaining results contradictory to previous work is not desired. In this analysis, however, the top-level findings from these six clusters directly corroborate previous results, with the high overall level of use for communication related activities clearly coming through. Additionally, the cluster analysis successfully separates out a group of users with low frequency of use and a corresponding low proportion reporting positive impacts. This showed up in previous analysis specifically as related to users in Bangladesh. However, building on these previous results, the clusters show a more nuanced picture of these features of the data. It appears that although communications usage/impact is high for almost all users, the activities and impacts paired with it do in fact differ between groups of users.

These are just some of the high-level results obtained via cluster analysis. The rest of this section examines how the clusters might be interpreted, how they relate to broad findings from previous analysis, and then moves into more specific discussions around specific features of the data, and how insight is added in these areas.

Example finding: social networking

The use of resources at public access venues for social networking and gaming is often considered less desirable than other activities, such as learning computer skills or accessing information. But are these activities really less beneficial to the users? Recall that there is both a high overall level of use for email, social networking, and more, and also a very high proportion of users reporting positive impacts to communication related activities.

Several clusters are indeed dominated by social activities, and most clusters show a high level of positive impact in the communication & leisure domain. The one cluster where this is not the case is the "Low-frequency users", where the proportion of users reporting positive impacts are at depressed levels. There is no cluster that **only** uses the venue for communication and social networking, and **only** reports positive impact to these areas. This is a key finding. If such a cluster definitively appeared, then it would support the conclusion that for some class of users, the only use and benefit of the public access is social. Although this lack of evidence is not conclusive by any means, and should not be taken as such, it does provide at least some additional information on which to base future analysis.

Additionally, the combination of social and constructive uses is illustrated with the "constructive and social users" cluster. This group shows how a number of users merge their social uses with unexpectedly high amounts of use in key areas such as word processing and reading news. This cluster has the highest proportion of female users out of all six clusters, and also reports high proportions of positive impact in most of the categories.

Example finding: gaming

As an activity frequently restricted or discouraged at public access venues, gaming is often considered a waste of time and resources. In the main Global Impact Study report, users who frequently played games did not show a large reduction in terms of reporting positive impacts in other categories. However, going back to the initial hypothetical example in the introduction, what is actually the interesting question here is not whether gaming *by itself* determines this outcome, but whether gaming **with** or **without** other activities does.

The six clusters here do not present evidence of a group of users focused on gaming to the exclusion of other activities. The clusters that show a higher level of gaming also show plenty of use in a variety of other categories. Gaming, even if done frequently, is not their sole use of the technology. Secondly in terms of perceived impact, gaming does not directly correspond to fewer users reporting positive impact. There may be a slight increase in the proportions reporting negative impacts in some categories, but it is not a dominating effect. This is not to say that a small group of users who solely play games does not exist, rather that the presence of a large group of such users was not observed in the current analysis.

New features of the data

Additionally (and perhaps most importantly) by using cluster analysis, features of the data emerge that would otherwise remain hard to see. Take as an example a finding from the main Global Impact Study report, showing that telecenter users in Bangladesh had generally lower usage frequencies and a lower proportion reported positive impact compared with other users. Thus it is not surprising that the "Low-frequency users" cluster is largely comprised of users from Bangladesh. What **is** surprising is the key additional insight gained by cluster analysis showing a small contingent of users **in each country** that fit this same pattern of usage and impact. When an entire country is treated as a unit this small group is overwhelmed by the other users. The cluster analysis pulls out the small percentage of users in a given country that are the Low-power users, leaving the rest to a separate, more focused and informative analysis.

Another key finding relates to the "constructive and social users", and is worth mentioning again here. Within most of the clusters the percentage of female users ranged between 30 and 35, but for this group it jumped to 43%. Although this is not a dramatic (or maybe even significant) increase, and does not surpass the 50% mark, it is not small either. This finding is extra interesting because gender was not used as an input variable to the clustering. Rather, this difference is solely due to the responses given on the usage and impact questions.

5 Conclusion

To summarize, cluster analysis has the potential to offer additional insight into a complex, primarily categorical, data set such as the User survey. The results discussed above expand on many of the findings from the Global Impact Study report as related to usage and perceived impact. Additionally, some features are discovered that would not easily result from analysis using crosstabs.

There are a huge variety of ways this research could be continued. A few examples:

1. Only a small, specific subset of variables was used. Other choices could be explored to provide alternate perspectives on the data. Specifically,
   - the domain-specific frequency of use variables from section 5 of the User survey could be used in place of those from section 4 (which were used here).
   - supplement with variables from other sections of the survey, such as seeking information at the venue (from section 3).
2. Add in data on how frequently the users actually visit a public access venue. The usage frequency questions from section 4 of the survey describe how often the tasks are done when the user is at the venue. Thus a user could report using email every time they visit the venue, but only visit the venue a few times a year. This is very different than a user visiting daily or almost daily and also reporting using email every time they visit.
3. Explore visualizations to graphically represent the findings. This is difficult due to the amount of information potentially needed to construct each plot.

These are just a few examples of future work that could be carried out on the User survey. Different subsets of variables that describe other features of the data lead to clusters different than the ones described here. This is not a failing of cluster analysis, in that it does not always find the same groups. Rather, when different variables are used, different users will be similar to each other. This illustrates both the benefit and the difficulty with cluster analysis. Just as there are groups of users based on usage frequency and perceived impacts, there could be interesting groups based on kinds of information searches, etc. These different possibilities highlight the vast potential for analysis present in this data.

The benefits of cluster analysis as a **complement** to other techniques should be apparent. It can allow for a compelling story to be built around complex features of the data, focusing the reader on key findings instead of discussing the dozens of crosstabs needed to support each point. In some cases, such as in the analysis above, the extra effort to cluster the data pays off with some very interesting findings.

Additional sections

The rest of this paper is organized into three sections expanding and supporting the above results. First there is detailed information on the methods, followed by the specific analysis techniques used to obtain the above results. This is followed by supplementary tables and figures. Interested readers, and researchers ready to try these techniques, may find some or all of these sections useful:

Section 6 starting on page 12 outlines in detail the methods researched. This includes coverage of both latent class and cluster analysis, providing technical details and references for further information. Within this section, section 6 starting on page 15 gives an overview of the topic of variable selection. This is not a trivial question, and thus deserves additional focus.

Section 7 starting on page 17 gives a detailed, step-by-step write-up of the analysis used to get the above results. This covers variable selection, choosing the cluster analysis algorithm, interpreting the results of the clustering, and summarizing the findings.

Section 8 starting on page 23 holds supplementary tables and figures. These include a table of the variables used, as well as a graphical representation of each of the 6 clusters.

6 Detailed methods

The challenge with simpler analysis methods, such as crosstabs, center around the extra work needed to connect the many disparate pieces of information into a complete picture. To describe a single feature of the User survey data may require hundreds of crosstabs, each only providing a tiny piece of the bigger picture. The researcher is then left with the tedious task of assembling these bits into a cohesive story.

To assist with this analysis, two data reduction methods were explored: cluster analysis and latent class (LC) analysis. Both of these form groups of observations within the data set, incorporating information from a set of input variables in doing so. Thus, for example, a feature of interest could be described by some subset of the variables. After either cluster or LC analysis, the resulting groups encapsulate the information from the input variables, and the groups are a simpler way of describing the feature, instead of the individual variables.

It may be clear already that cluster analysis gave better results for the User survey data. This section describes each of the two methods in more detail, and in so doing highlights the reasons for this choice. For a detailed description of the specific analysis steps, turn to Section 7.

Latent Class analysis

LC analysis uses Maximum Likelihood and the EM algorithm to fit an appropriate model to the data that estimates the group assignments ([Vermunt and Magidson, 2004], [McCutcheon, 1987]). This model-based approach has advantages, but requires assumptions on the structure of the data. Specifically, it assumes that underlying the observed data is a mixture of known probability distributions. These are either Gaussian (for continuous data) or Multinomial (for categorical data). When this assumption is satisfied, powerful model fit and diagnostic tools are available. For the multi-level categorical data in the User survey, a mixture of Multinomial distributions is the correct choice.[2]

Formally, the LC model can be written as

$$f(\mathbf{y}_i|\theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk})$$

where $\mathbf{y}_i$ denotes the vector of responses from the i^{th} observation, K is the total number of classes[3], and π_k denotes the prior probability of belonging to class k. θ is the vector of model parameters, and J is the total number of variables. Since a higher number of classes usually leads to a better fit, even if the resulting model is useless, the BIC is used to compare the fit for varying numbers of classes. The BIC is a measure of model fit that penalizes the increase in the likelihood from adding variables by a factor accounting for the increased complexity from these same variables [Schwarz, 1978].

Although the model and the analysis are more complicated than ordinary regression, software packages are available for use. For the User survey, LC analysis was done using the `fpc` package in R, specifically the

[2] The astute reader will notice that some of the variables in the User survey are ordinal. This was an additional reason why cluster analysis was chosen over LC analysis, to better take advantage of this fact. Nonetheless, LC analysis is still valid on the ordinal data, the disadvantage is that a small amount of information is thrown away by treating the ordinal data as categorical.

[3] LC analysis denotes the groups of users as "classes", whereas the term "clusters" is used for the results of cluster analysis, even though they are the same object.

`flexmixedruns` function [Hennig, 2010]. This function naturally handles the categorical data, making it easy to fit the model no matter the data type.

There are several advantages to the LC analysis approach over cluster analysis. By fitting a model to the data, a wide range of tools are available to describe model fit, determine the optimal number of clusters, and more. Additionally, LC analysis gives a "soft" classification, in that for each observation, it returns the probability of belonging in each of the K classes.[4] While each observation is generally assigned to the group for which it has the highest probability of belonging, the soft classification can give a sense of how distinct the groups are.

Cluster analysis

Cluster analysis has a very different approach, using numerical optimization methods instead of a statistical model. There are two components needed: a "similarity" measure and an algorithm. The similarity measure is some way of quantifying how "close" two observations are to each other. The algorithm then takes this similarity information and uses it to group the observations into clusters. The goal is to form clusters such that observations within a cluster are more similar to each other than to observations in other clusters, no matter what definition of similarity is used.

One challenge to using this approach is that there are many choices for both the similarity measure and the algorithm. Some have a specific data type or clustering structure in mind, or there are varying levels of complexity in how the similarity is determined. This introduces a large degree of difficulty (but also flexibility) to the clustering problem.

SIMILARITY MEASURES

The similarity measure is the first step in cluster analysis, used to quantify how similar subjects are to each other. Intuitively, similarity can be thought of in terms of distance: subjects close to each other are more similar than those far apart. For example, if two people report incomes of $\$49$K and $\$50$K, they are clearly similar in terms of income. But if a third person reported $\$250$K, they are not at all similar to the first two. Although this example uses a continuous variable (income), the extension of this idea to categorical data will be used on the User survey.

For categorical data distance doesn't work in the same way as for continuous data. The two responses "No" and "Yes" do not have an intrinsic distance between them (e.g. No $-$ Yes $=$???). To get around this, there are a number of similarity measures specifically designed for categorical data. These methods work by assigning a numerical distance between observations, based on pre-defined criteria. One basic approach is called "simple matching" distance. It is based on taking each subjects response to a question, and comparing it to the response from every other subject. If the response matches, those two subjects are "similar" on that question. Other methods add complexity, for example matching on a rare response carries more weight than matching on a common response. A detailed comparison of similarity measures for categorical data is found in [Boriah et al., 2008].

[4] Formally, the posterior probability of observation y_i belonging in class k, $\pi_{k|y_i}$, is given by

$$\pi_{k|y_i} = \frac{\pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk})}{\sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} f_k(y_{ij}|\theta_{jk})}$$

The computed similarities are typically stored in a square matrix of size equal to the number of observations. The matrix cell $n_{i,j}$ contains the similarity between observation i and observation j. With multiple input variables, a separate matrix is constructed for each one and then combined to form the final similarity matrix, either by adding all the matrices together or using a weighting scheme. This combined matrix can then be used as input for the chosen clustering algorithm.

ALGORITHMS

A wide variety of algorithms are available to process the similarity matrix into clusters. These vary, for example, in how the clusters are initialized, what target shape the clusters take, and more. This analysis focused mainly on the **hierarchical** family of algorithms, which are widely implemented in software.

Given a data set, there are two extremes in terms of clusters: each observation in its own cluster, or all observations in one cluster. In between these two extremes, a "hierarchy" is computed showing when clusters are joined or split (see Figure 2). There are two methods within the hierarchical family: **divisive** (starts with one cluster) and **agglomerative** (starts with each observation in its own cluster).

In divisive clustering, all the observations start in a single cluster. At each step a cluster is split, such that the overall level of similarity within the clusters is increased. The last step leaves all observations in their own cluster.

In agglomerative clustering, the reverse path is followed. All observations start as individual clusters. At the first step, the two most similar clusters (at this point they are just observations) are merged into a new cluster. The next step is to merge the next two most similar clusters together (this could be either merging two singletons, or merging a singleton into the previously formed cluster of two). This merging (agglomeration) continues until all the observations are in a single cluster.

Figure 2: **Splitting/merging of observations into clusters***: Toy example of 4 observations and how they might split (divisive clustering) or join (agglomerative clustering).*

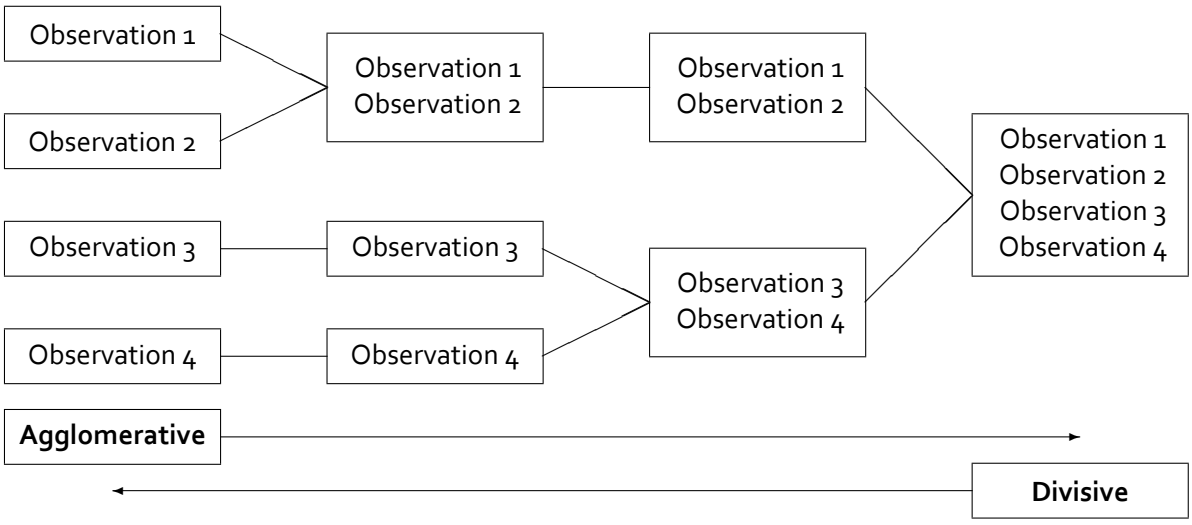


Figure 2 illustrates these two methods with a small example. In this figure, divisive clustering moves from right-to-left, breaking up larger clusters into smaller ones until all the observations are in single clusters.

Agglomerative clustering moves from left-to-right, joining clusters at each step until all the observations are in a single cluster. In neither of these two methods are observations allowed to switch clusters. For divisive clustering, this means that once a sub-group is split off, those observations can't re-join with observations in another group at a later step. Similarly for agglomerative clustering, once two sub-groups join, those observations can't leave that cluster.

An additional parameter in the agglomerative clustering is the method of determining which two units to join at a given step. These vary from simple (join the two units closest together) to more complex (join units such that the overall variance of all the clusters has a minimal increase). This latter method, Ward's minimum variance approach [Ward Jr, 1963], has an advantage over other joining methods in that it keeps the clusters as compact as possible.

DETERMINE OPTIMAL NUMBER OF CLUSTERS

The final step in using a hierarchical method is to determine the "best" number of clusters. This is actually not a trivial problem, because the whole range of clusters, from one cluster to many clusters, is given, and unlike LC analysis there is no built-in measure of the best fit. A review of some methods is given in [Milligan and Cooper, 1985], and two options are described here.

The first is a numerical approach, which uses a cost function to optimize the fit of the clusters. This cost function could, for example, compare the distance between elements of a cluster to the distance between clusters. If the clusters are both far apart but compact, this suggests the clustering successfully differentiated observations. However, the cost function is an additional layer on top of the similarity measure and the clustering algorithm, and was thus avoided in this analysis.

The approach used in this paper is made possible by the hierarchical method of clustering. The "height" at which each join is made is plotted against the number of clusters. An "elbow" becomes apparent where the height change to the next join starts increasing dramatically. The number of clusters at the elbow then strikes a balance between describing the data adequately and finding distinct clusters. See Figure 5 for an example.

ADDITIONAL (NON-HIERARCHICAL) CLUSTERING METHODS

There are other possible clustering algorithms for categorical data. One such possibility tested on the User survey is the **k-modes** algorithm [Huang, 1997]. Based on the popular k-means algorithm for continuous data, it is adapted for categorical data by using the mode (the most common value) in the calculations. Although this method is relatively fast, there are a couple of issues that make the k-modes algorithm less suited to the User survey. First, a required input is the desired number of clusters to make, which doesn't work here because of the exploratory nature of the analysis. Second, the results can be dependent on the order in which the observations are processed, an undesired constraint. Thus this method was not pursued further.

Variable selection

Variable selection is an important step prior to either cluster or LC analysis. Both of these methods will try and group the data based on information in the input variables. If these variables are poorly selected,

interpretation of the results may be more difficult. For example, the "Country" variable in the User survey would be a poor choice, because it is clear that this already strongly defines groups in the data. When used as input to cluster analysis, this variable dominated the results such that clusters spanning multiple countries (but similar on other variables) were much less distinct.

Faced with nearly 500 variables in the User survey, it is not a trivial task to select a subset for analysis. Other large data sets present similar problems, because many variables are of interest and directly or indirectly apply to even a simple research question. Unfortunately, it is impossible to simply take a large number of variables and use a software algorithm to pull out the best subset to use in analysis. Although such algorithms exist, it is actually important to eliminate some variables (like "Country") that would almost certainly be chosen by any such method. A short discussion of algorithms for both LC and cluster analysis follows, and the resulting recommendation is in Section 7.

VARIABLE SELECTION ALGORITHMS IN LC ANALYSIS

Since LC analysis uses the BIC as a measure of model fit, a variable selection algorithm based on the BIC was proposed by [Dean and Raftery, 2010]. This is a forward/backward algorithm that alternately proposes adding a candidate variable to the model, calculating the BIC, and adding the variable in if it improves the fit. Once a variable is added, each variable in the model is tested to see if removing it also increases the fit. This method robustly discerns the best variables in the shortest time.

Unfortunately as applied to the User survey, there were several fatal issues with this method. First, the large number of observations meant that every added variable showed a significant increase in model fit. Secondly, using the `fpc` package in **R** to fit the LC model, the variable selection algorithm (when presented with an input set of 22 variables) would take 3-8 days to run[5], and return all 22 as significant. These two difficulties essentially mean that for the User survey, the LC variable selection algorithm was not helpful.

VARIABLE SELECTION ALGORITHMS IN CLUSTER ANALYSIS

For cluster analysis, variable selection is more commonly referred to as "feature selection" in the machine learning literature. Feature selection is often based on choosing those variables that improve a specified cost function, where the cost function is some measure of how good the clusters are (see [Dash and Liu, 1997] for more information). For example, if using a distance-based similarity measure, the cost function could compare the distances between subjects *within* a cluster versus the distance *between* clusters. If there is a lot of separation between clusters, then the between-cluster distances will be much larger than the within-cluster distances, and the clusters will be more distinct than if the two distances were similar. The algorithm thus chooses variables which improve this metric.

HEURISTIC VARIABLE SELECTION

Although relying on an algorithm for variable selection is appealing, we experienced significant difficulties with actually using any of the above methods on the User survey. Our approach was to use knowledge of the research area, the research questions, and other aspects of the data to form a focused initial set of

[5]Using an Intel i7-2600K quad-core processor overclocked to 4.4 GHz, with 16 GB of DDR3-1600 memory.

variables. This was a good start, but in certain sections of the User survey, response patterns were very similar between some questions. In our experience, keeping highly correlated variables in the clustering led to more, but less distinct, clusters. Thus two approaches were considered to select variables with the most unique information about the users:

1. Compute the correlation between variables. Exclude one variable from those pairs that are highly correlated.
2. Use cluster analysis on the **variables** (not the subjects), and pick variables that are from separate clusters to use in the analysis.

With the categorical data, the second option was easy to use and interpret.

7 Example application of cluster analysis to the user survey

This section illustrates the actual approach used with success on the User survey, and produced the results already discussed. The goal of this section is to provide a concrete, worked through example of the steps in analysis and interpretation. There are 5 steps:

1. Variable Selection
2. Constructing the Similarity Matrix (if needed)
3. Clustering
4. Determine best number of clusters
5. Interpretation

These steps are discussed in detail below, and move through the analysis from start to finish.

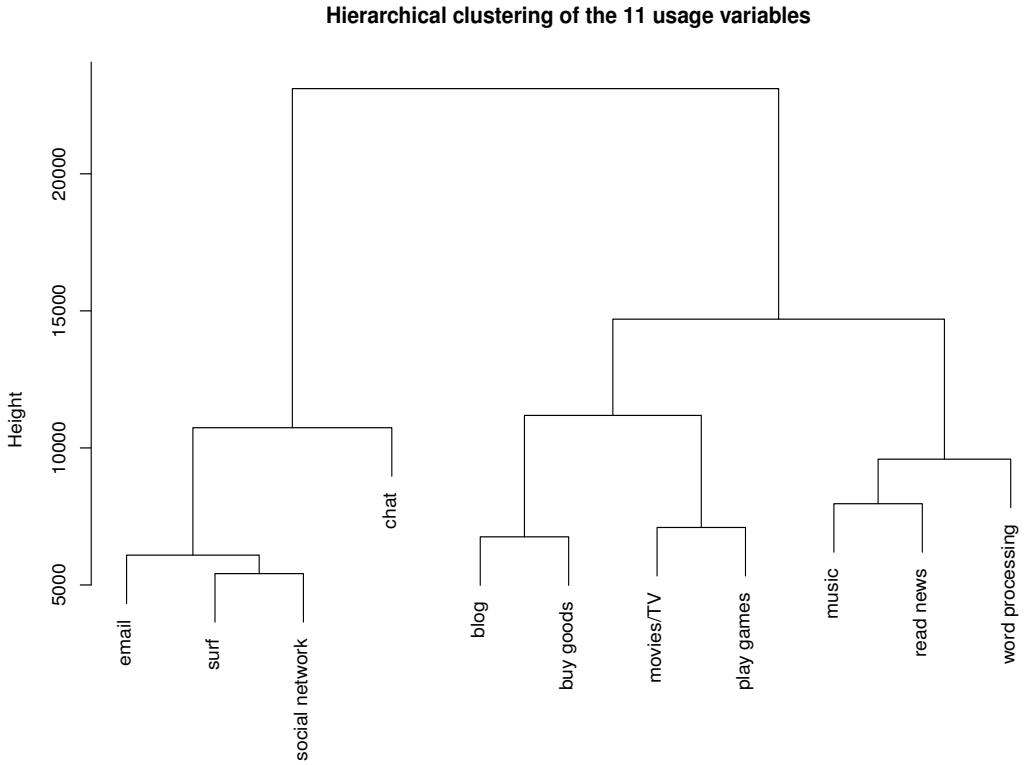
Step 1: Variable selection

The first step in analysis is to choose a set of candidate variables for the clustering algorithm. This began by restricting the focus to just the interplay between usage patterns of public access venues, and the perceived impact from public access venues. This left 11 usage variables from section 4 of the survey, and 13 perceived impact variables from section 5. Cluster analysis on the usage variables and the impact variables separately suggested that a number of variables were redundant, and could be dropped.

This step (clustering the variables) is illustrated with the dendrogram in Figure 3. The height (vertical axis) is a measure of the dis-similarity between groups. The horizontal lines are connections between clusters (with one or more elements) drawn at the height where the level of dis-similarity allows them to join. To interpret this, the relative level of dis-similarity between two groups (or elements) is how high one must go in the graph in order to connect them. The vertical distance necessary gives a general sense of how different the groups are.

For example, take several cases from Figure 3: the frequency of use for surfing, social networking, email, and playing games. Surfing and social networking joined at the lowest level of dis-similarity (a height of

Figure 3: *Hierarchical clustering of* **variables** *for 11 questions relating to the frequency of use of technology related tasks. Taken from section 4 of the User survey, (Q4_4a - k) the questions are phrased as "For the venue that you usually go to, how frequently do you -- { email, blog, ... }"*



just over 5000), implying that users responded most similarly to these two questions. The frequency of use for email is joined to this cluster (surfing and social networking), and without much of an increase in height (possibly just under 6000). This difference (less than 1000) is a sense of how dis-similar email is from surfing and social networking. Since the scale goes to well over 20,000 this implies that these three activities see similar usage patterns among the users.

Compare this with the variable on frequency of playing games. Recall that a comparison of the similarity of two clusters is the height of the horizontal line needed to connect them. For gaming to connect with the email/surfing/social networking cluster requires traversing the highest horizontal line, at a height of nearly 24,000. This means that gaming is not associated as strongly with communication as with (for example) watching movies/TV.

To select variables to use in the analysis, it is now possible to eliminate some that are potentially unnecessary. Since email, surfing, and social networking are the most similar in the figure, only one was selected for clustering the users. In a similar fashion, playing games is used since that variable is of continuous interest, but watching movies/TV is dropped. In the end six of the 11 variables were selected as part of the clustering: frequency of use in social networking, chatting, buying goods online, playing games, reading news, and word processing.

The perceived impact variables can also be trimmed (see Figure 6 on page 23 for the dendrogram). This figure suggested the use of 8 of the 13 variables: perceived impact to income, health, culture, financial savings, access to employability resources, education, meeting people, and pursuing interests & hobbies. A summary of all 14 variables used is in Table 1 on page 7.[6]

Step 2: Constructing the similarity matrix (if needed)

For simplicity in this analysis, and because the variables are ordinal, they were recoded using an integer scale: $\{1,2,3\}$ for the levels of perceived impact, and $\{1,2,3,4,5\}$ for the frequency of use. This preserved the information about some levels of a variable being more different than other levels, while at the same time eliminating the need to explicitly compute the similarity matrix. If the data were categorical, then the built-in distance functions may not work, and thus the similarity must be defined and computed before clustering. The **R** functions used here, **agnes()** and **diana()**, do not require a similarity matrix as input, as long as the data is numeric.

Step 3: Clustering

Using agglomerative hierarchical clustering with **Ward's minimum variance** method gave the best results. The dendrogram for this clustering on the 14 variables chosen is shown in Figure 4. The bottom of the figure is dense and messy, as most of the singletons and small clusters quickly join together. This is good, because it indicates that a lot of the users share similarities with at least some other group of users. As the height increases the clusters grow larger, and the increase in height before the next join keeps increasing, implying that the clusters are more strongly differentiated.

Step 4: Determine best number of clusters

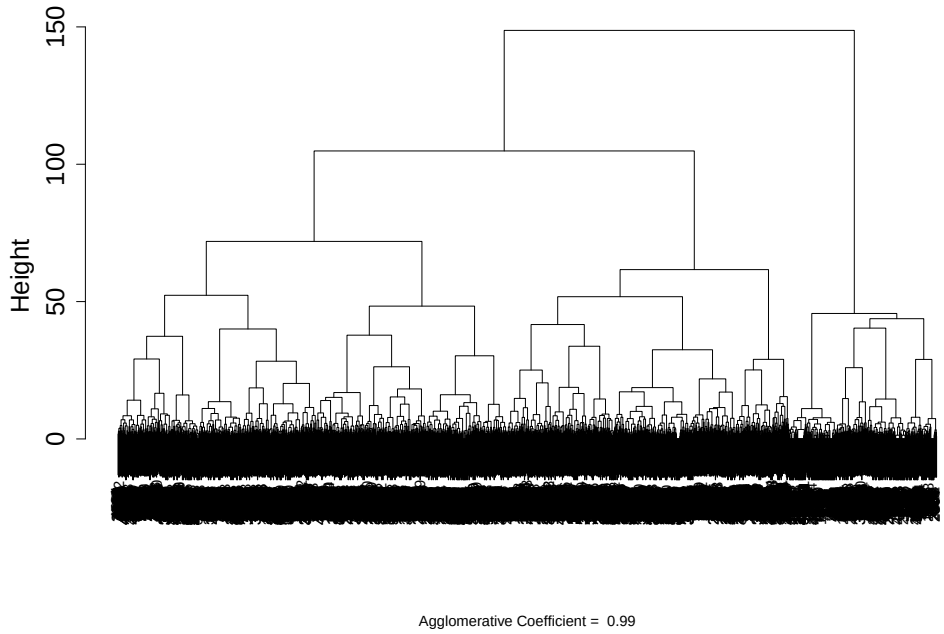
Plotting the height at which each merge occurs versus the number of clusters, as discussed in the methods section, simplified the decision on how many clusters were appropriate. Using Figure 4 as an example, the final merge (from 2 clusters to 1) occurred at a height of just under 150, while the merge from 3 clusters to 2 clusters was at a height of just over 100. This is a difference of close to 50, whereas the difference going from 3 to 4 clusters is closer to 30. These differences should become smaller as the number of clusters increases, and this is shown in Figure 5. The vertical line (drawn at 6 clusters) locates this distinct change, as the height difference to the next merge is much greater than to the previous one.

Step 5: Interpretation

The last step in this analysis was turning the clusters into usable information. The began by comparing clusters across variables that were actually used in the clustering, for example usage frequencies. Table 2

[6] Using all 11 usage variables and 13 perceived impact variables is also a viable approach. However, using the same analysis techniques as outlined below highlighted both positives and negatives to the outcome. There may be slightly more information retained in this analysis, which did lead to an increase in the optimal number of clusters, but using the correlated variables showed in that some of the clusters were barely distinct. If these nuances are acceptable (or even desired), then there may be advantages to this approach. For simplicity and interpretability, trimming the variables is definitely something to at least consider in the initial analysis.

Figure 4: *Dendrogram from* **agnes**, *using Ward's minimum variance method for joining clusters. This figure shows good clustering: the bottom of the figure is quite messy and dense, as there are a lot of merges happening. As the number of clusters decreases, the jump in height between merges increases dramatically.*



shows the crosstab created from the cluster assignment versus frequency of gaming. This clearly shows the dramatic differences between clusters in terms of their responses to this question.

Table 2: *Crosstab of the cluster assignment with frequency of use for gaming.*

Cluster	Never	Rarely	Sometimes	Most times	Every time	Total
Normal users	51	19	12	11	7	100%
Highly social users	69	26	0	5	0	100%
Low-frequency users	70	5	18	3	5	100%
Social gamers	0	2	37	12	49	100%
Constructive and social users	60	30	1	9	0	100%
Power users	21	15	14	19	30	100%

The next step in this paper was to cross the cluster assignment with other variables of interest that were not included in the cluster analysis. This shed additional light onto the different groups of users. Take gender, for example, shown crossed with the cluster assignment in Table 3. Most clusters hover around 33% female, but the "constructive and social users" cluster is markedly different with 43% female. Although not necessarily a significant difference, it is certainly noticeable, and could suggest future analysis.

The summary description of each cluster has already been presented on page 8, but figures for each cluster, showing usage frequency and perceived impact for all categories (as well as age, country, venue type, and gender breakdowns), are found in Figures 7 to 12 on pages 24 to 29.

Figure 5: *Plot comparing the number of clusters versus the height (required dis-similarity) for them to merge. The vertical line shows where the required dis-similarity takes a sharp increase. This is a likely candidate for the appropriate number of clusters.*

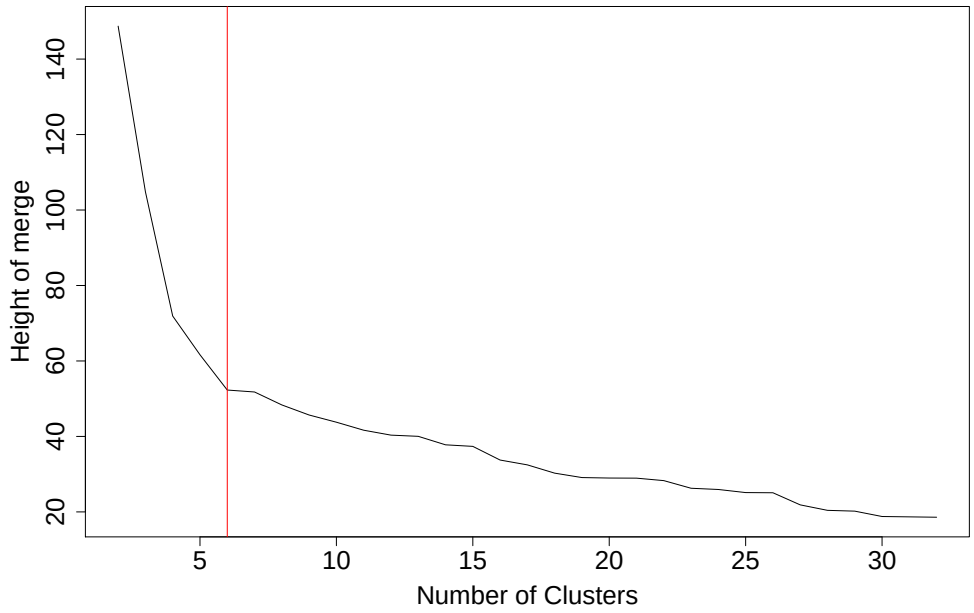


Table 3: *Crosstab of the cluster assignment with gender variable. Shows the breakdown of gender within each cluster.*

Cluster	Female	Male	Total
Normal users	33	67	100%
Highly social users	35	65	100%
Low-frequency users	32	68	100%
Social gamers	31	69	100%
Constructive and social users	43	57	100%
Power users	31	69	100%

References

S. Boriah, V. Chandola, and V. Kumar. Similarity measures for categorical data: A comparative evaluation. *red*, 30(2):3, 2008.

M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131--156, 1997.

N. Dean and A.E. Raftery. Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62(1):11--35, 2010.

Christian Hennig. *FPC: Flexible procedures for clustering*, 2010. URL `http://cran.r-project.org/web/packages/fpc/index.html`.

Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Workshop on Research Issues on Data Mining and Knowledge Discovery*. Citeseer, 1997.

A.L. McCutcheon. *Latent class analysis*. Sage, 1987.

Glenn W. Milligan and Martha C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159--179, 1985.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461--464, 1978.

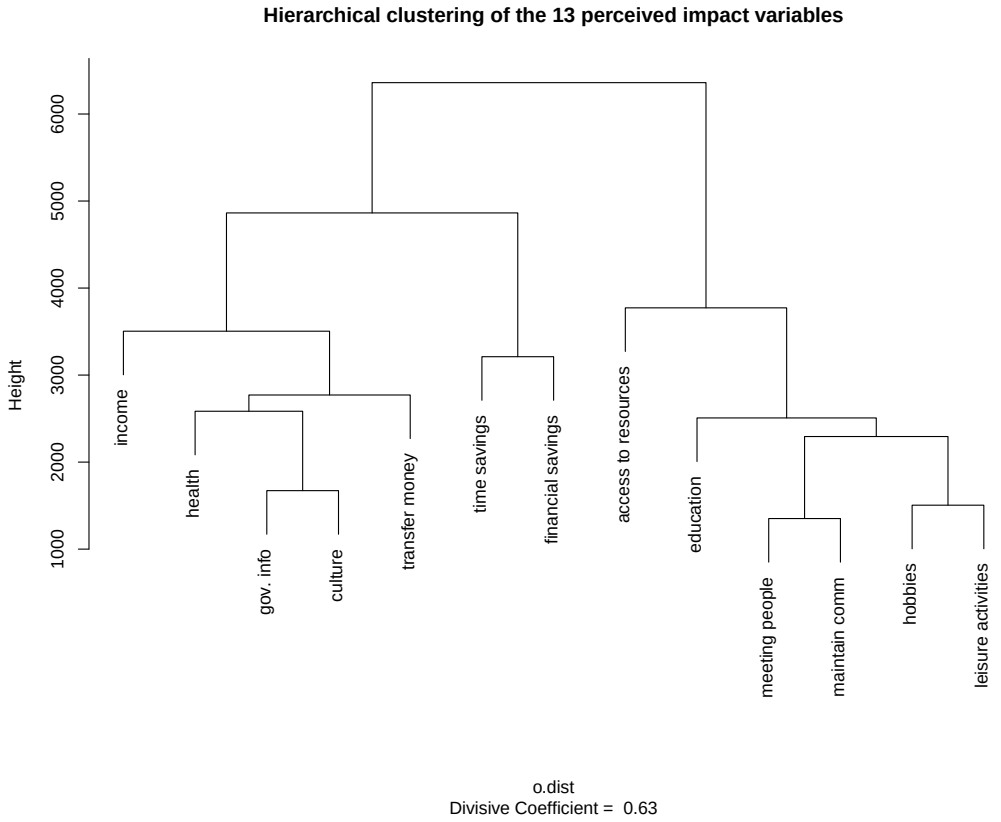
A. Sey, C. Coward, F. Bar, G. Sciadas, C. Rothschild, and L. Koepke. *Connecting people for development: Why public access ICTs matter*. Technology & Social Change Group, 2013.

J.K. Vermunt and J. Magidson. Latent class analysis. *The sage encyclopedia of social sciences research methods*, pages 549--553, 2004.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236--244, 1963.

8 Additional figures and tables

Figure 6: *Hierarchical clustering of* **variables** *for 13 questions relating to the perceived impact to the user. Taken from section 5 of the User survey, (Q5_1a - m) the questions are phrased as "Overall impact from your use of public access venues -- { income, health, education, ... }"*



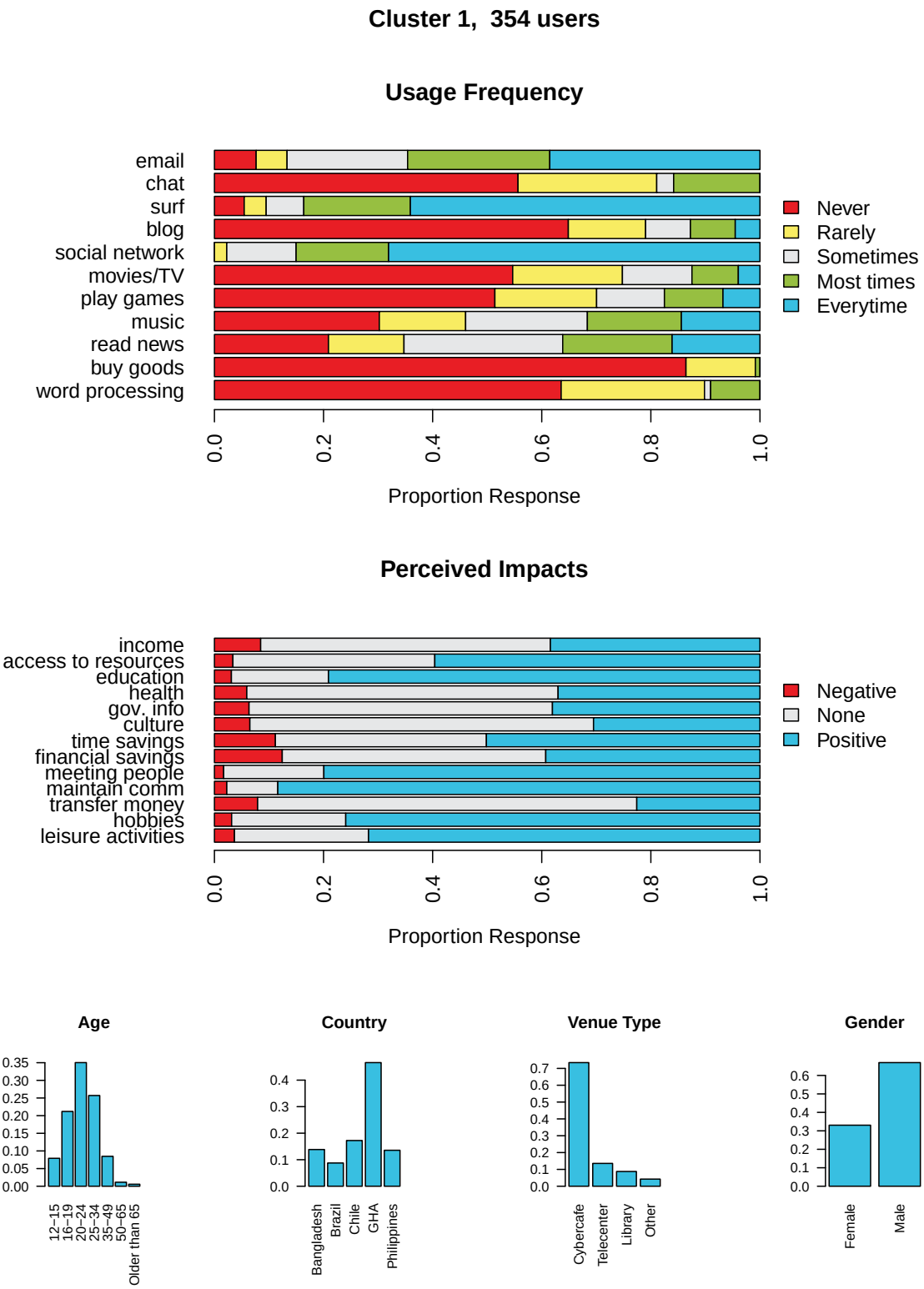


Figure 7: *Typical, largely social users*

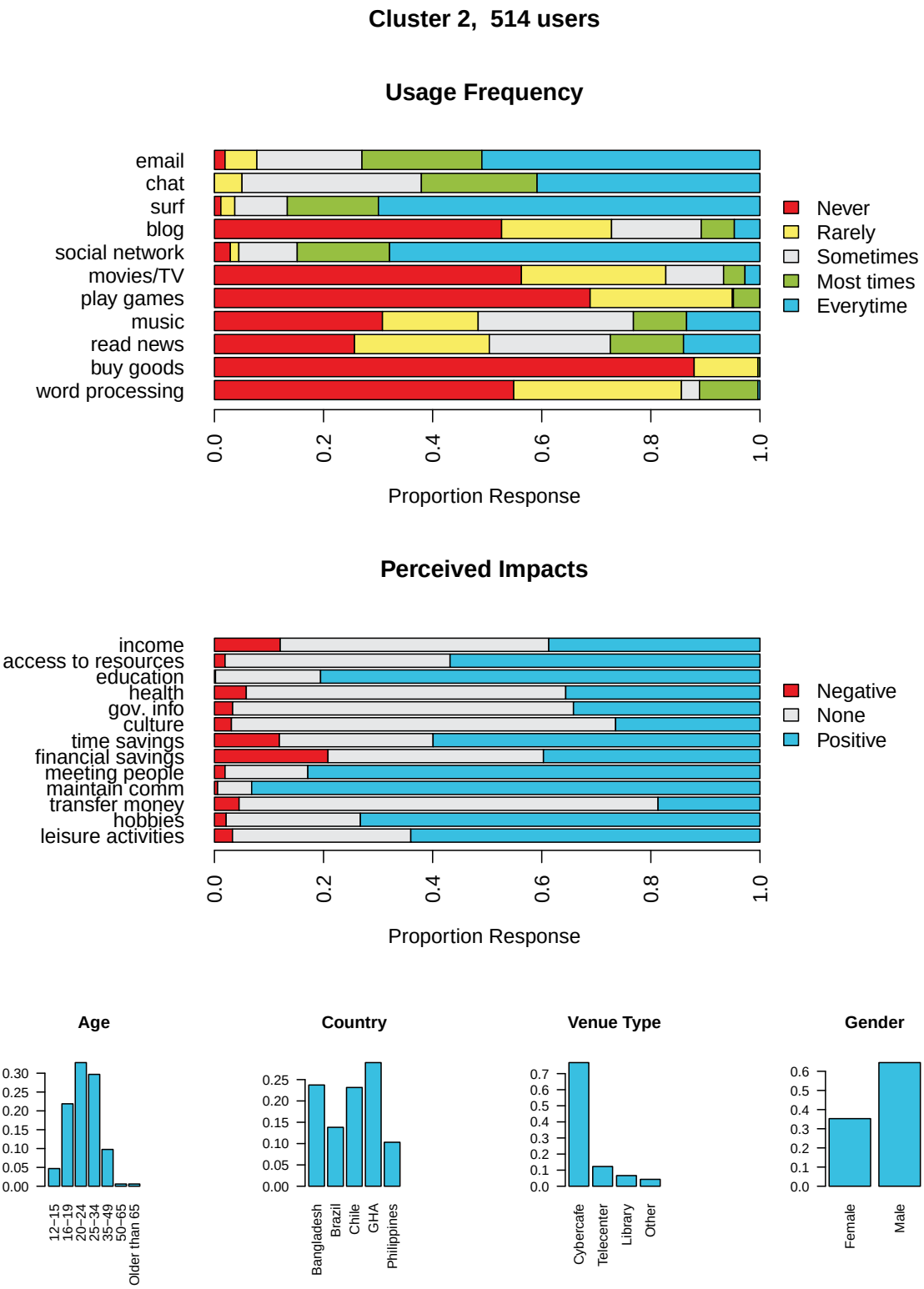


Figure 8: *Highly social users*

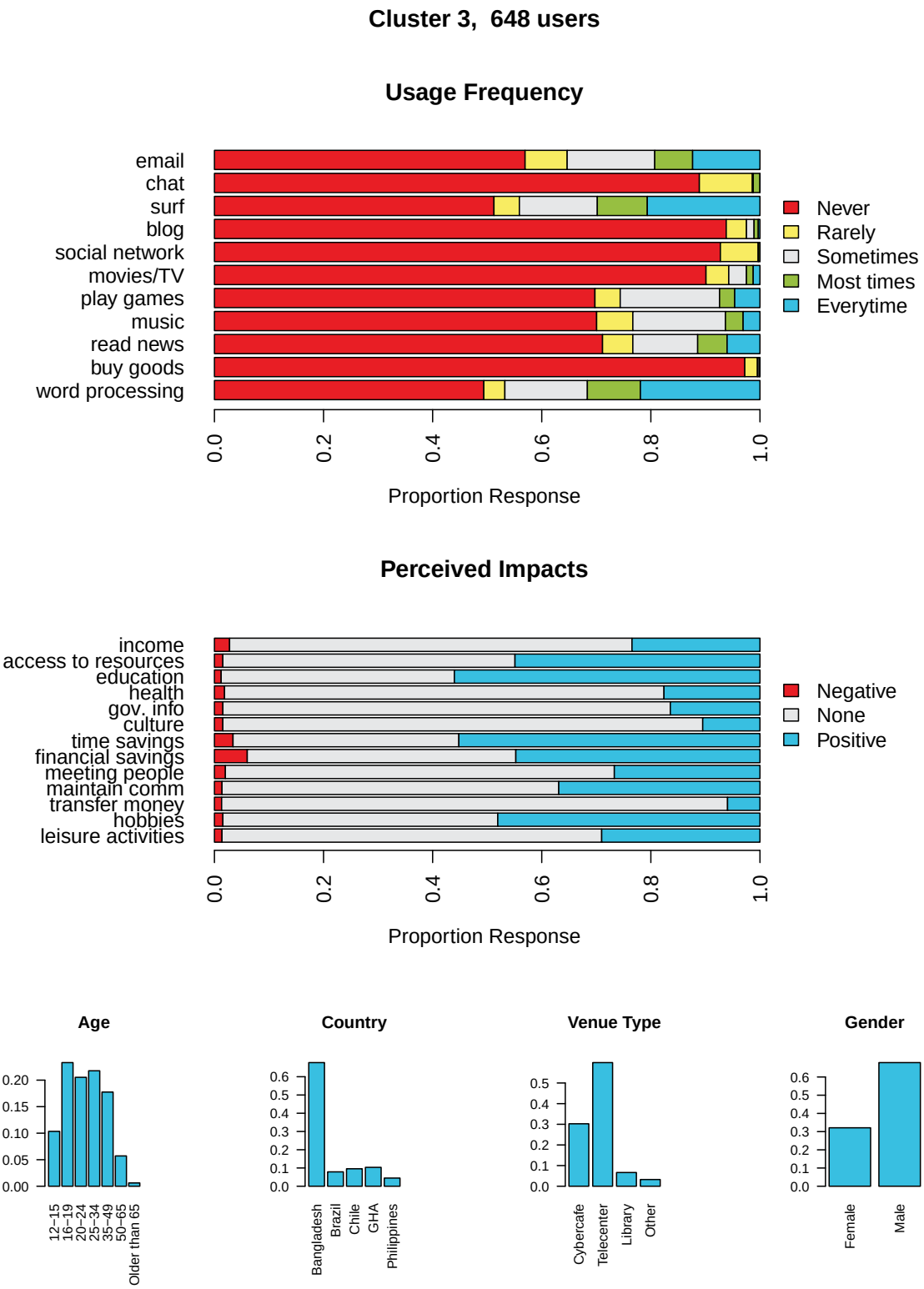


Figure 9: Low-frequency users

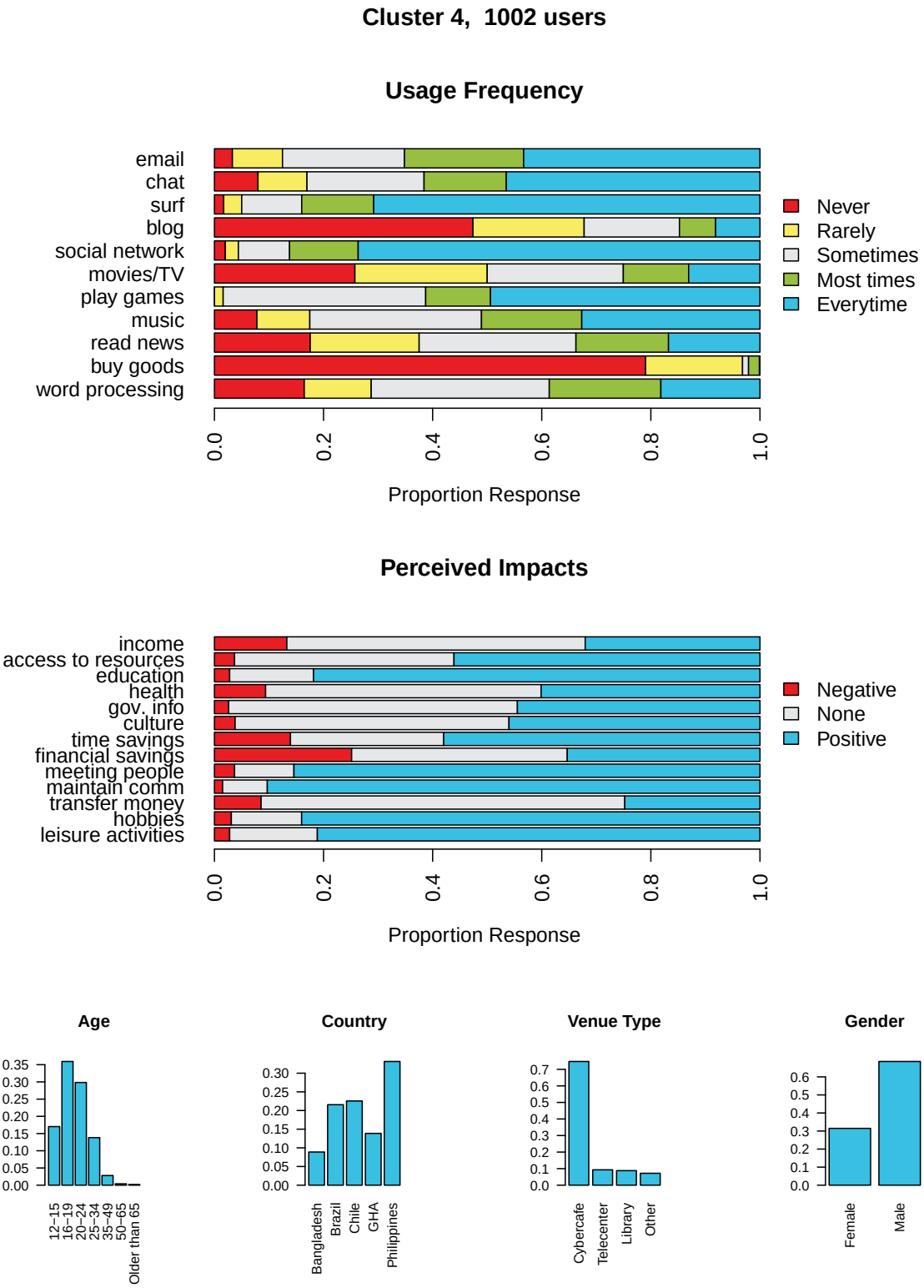


Figure 10: *Social gamers*

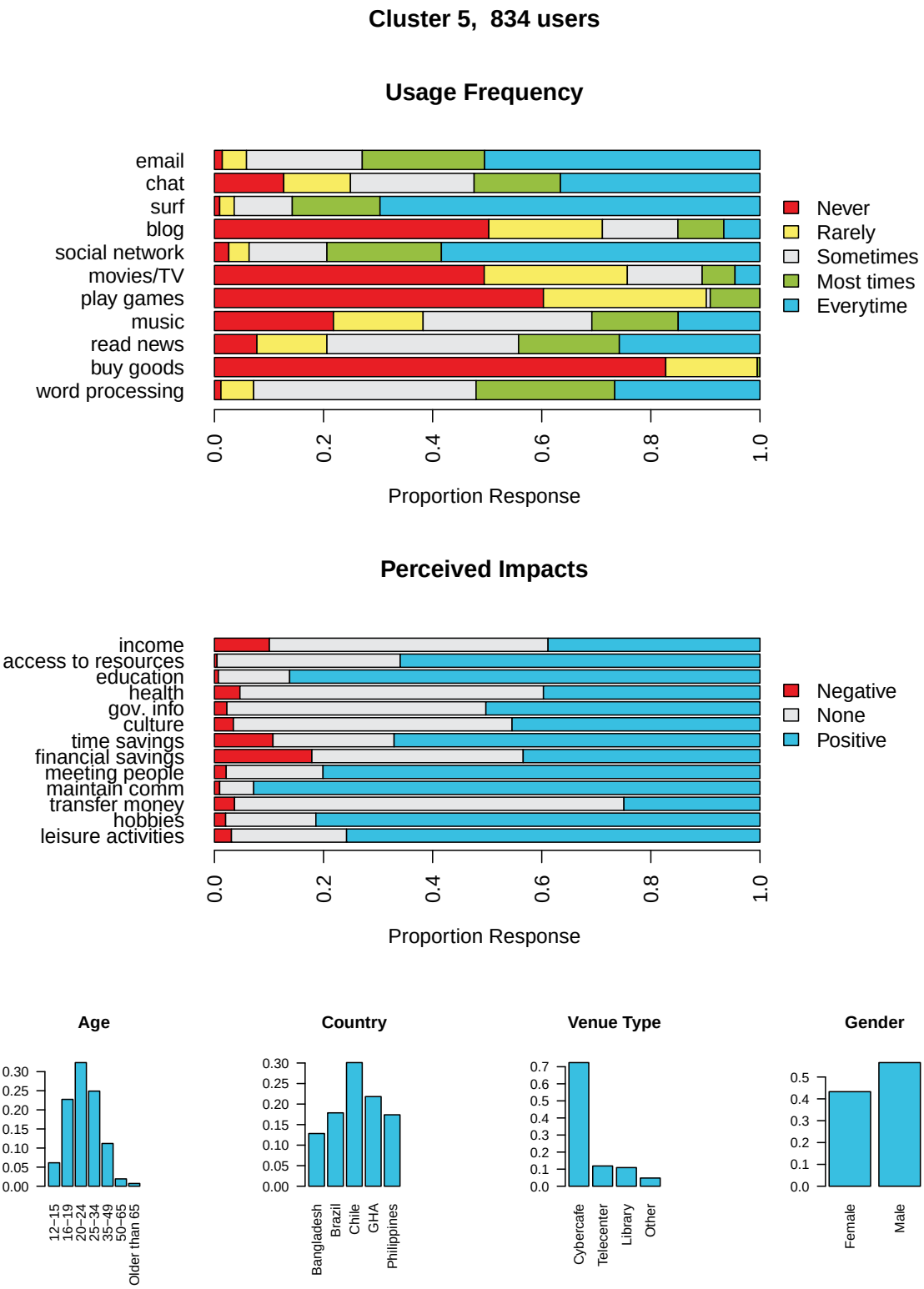


Figure 11: *Constructive and social users*

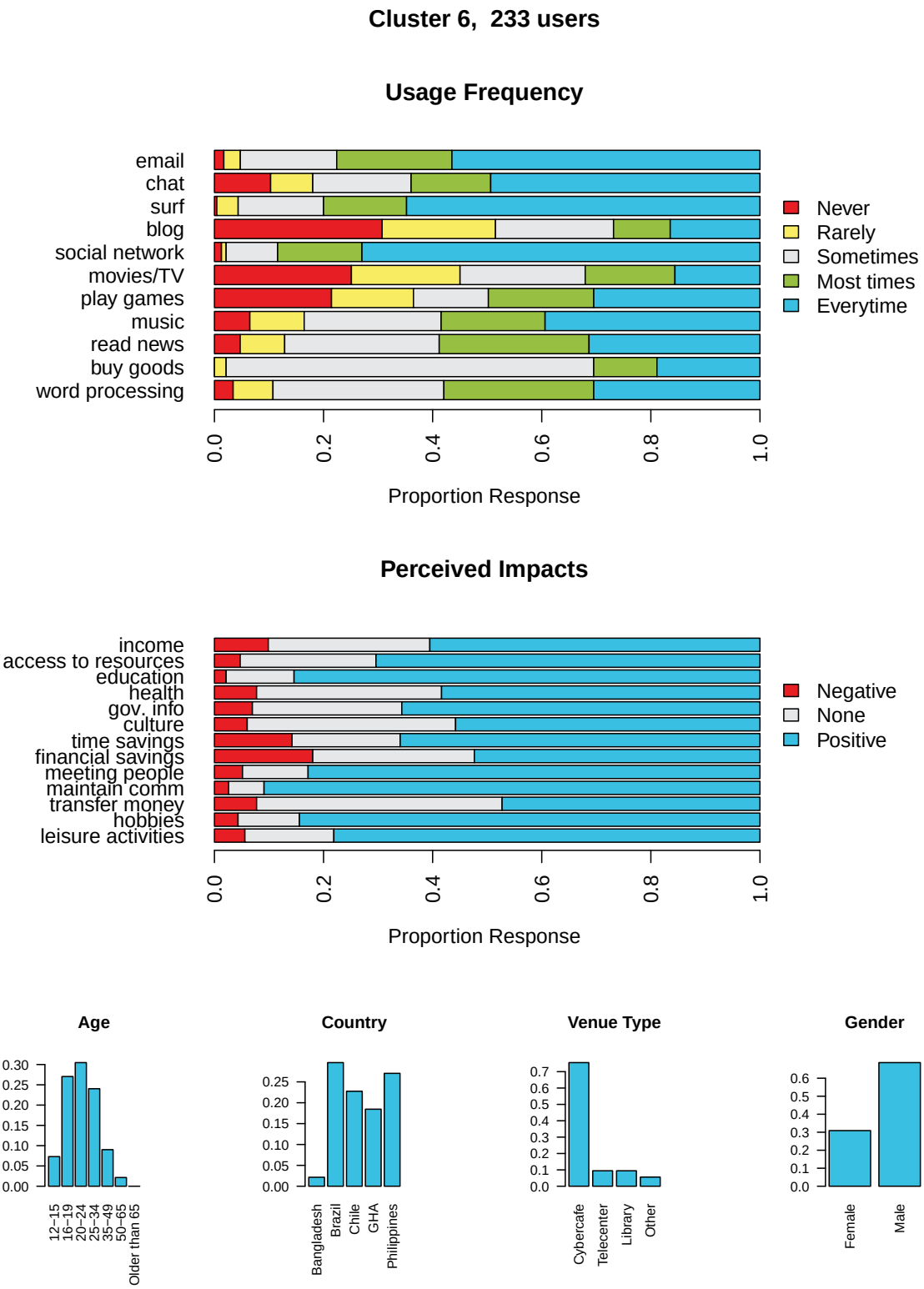


Figure 12: Power users