# The six business models for copyright infringement

A data-driven study of websites considered to be infringing copyright



A Google & PRS for Music commissioned report
with research conducted by BAE Systems Detica.

27th June 2012

Acknowledging contributions of data from:





THE PUBLISHERS ASSOCIATION

with the assistance of:

# Executive summary

*The Six Business Models for Copyright Infringement* is a segmentation driven investigation of sites that are thought by major rights holders to be significantly facilitating copyright infringement. In this study, we investigate the operation of a sample of these sites to determine their characteristics. Among other things, we investigate how they function, how they are funded, where they are hosted, what kinds of content they offer, and how large their user bases are.

The aim of this study is to provide quantitative data to inform debate around infringement and enforcement. Although a large amount of quantitative and qualitative data has been collected in the past through consumer surveys into why people use these sites, there is insufficient data-driven analysis of the sites that are considered to facilitate copyright infringement.

## How the data was collected

For this study, BAE Systems Detica collected from rights holders lists of sites that they believed to be significantly infringing copyright. These lists provided more than one thousand sites. A systematic sample of 153 sites, together with publicly available information, was used to build a segmentation model. The resulting segments were analysed, and their characteristics were confirmed in a subsequent analysis of 104 additional sites. In contrast to previous research this analysis of the market for copyright infringement is based on a statistically significant representation of these sites.

This research provides industry and policymakers with information about the business of copyright infringement. The segmentation of the results revealed six major business models, which are shown in Figure 1-1:


*Figure 1-1: Six major copyright infringement business models identified in this study*

Each of the segments identified in this study are characterised by the type and operation of the sites found within them. Below we describe the differences between the segments in terms of the way they are financed, the content and formats provided, how users arrived at sites and where the segments are predominantly located. See Figure 1-2 for more details.

## Key Segment Characteristics

### Financing

This study provides data-driven insight into how copyright infringement operates as a business across a range of business models. It shows that websites are most commonly funded in part or in combination by either advertising or payments (including subscriptions, donations, and transactions).

For each segment, this study helps to identify which are the significant economic drivers. This data is likely to prove useful and insightful to industry and policymakers who seek to tackle infringement by 'following the money'.

### Advertising

Advertising plays a key role in at least three of the segments. To understand where these adverts were coming from, we examined the advertisements found on each site by checking for the presence of the "Ad Choices" logo. The "Ad Choices" scheme is administered by the Internet Advertising Bureau (IAB) in the UK, and ad agencies must sign up to be included. For all the sites we segmented, 86% of advertisements did not display the Ad Choices logo suggesting that the advertisers do not associate themselves with the online advertising self-regulation scheme.

Each segment has different proportions of advertising or payments. For example, two-thirds (67%) of the 'Live TV Gateway' segment, the fastest-growing segment, which consists of sites that provide live-streams of free-to-air and pay TV content as well as other content, are funded by advertisers. These sites also solicit donations as a part of their business model.

'P2P Communities', the second fastest growing segment, are even more dependent on advertising income (86%) than the Live TV Gateway segment and more likely than all five other segments to solicit donations from their community members.

### Payment and card processors

The study also examined in an objective way the presence and influence of payment processors and card processors. In at least three of the segments, the existence of the logos for credit card and/or electronic payment processor logos were significant. Whilst the presence of these logos does not give us certainty that card processors or payment processors actually facilitate payment, it does suggest the strong likelihood that these payment facilities are used for payment collection.

Two of these segments include sites which collect subscriptions via their payment pages: we called these 'Subscription Community' and 'Rewarded Freemium'. A third segment, which we called 'Music Transaction', contained sites that appeared to collect payment for the content that they sell.

Overall, 36% of the segmented sites had payment pages; credit card company logos were present on 69% of them. However, that is not to say that the remaining 64% were not taking payment, only that a payment page was not visible to us, for example if a site was closed and we could not obtain membership.

The visibility of card and payment processor logos suggests a critical relationship between those sites and the subscription and transaction services that they may rely on. More specifically, those engaged in these transaction services appear to be clustered in particular countries.

### Content and format

In addition to insight on financing, this study also provides data on which kinds of sites favour certain kinds of content.

A broad range of content including music, films, software, games and ebooks appears on many sites. However, it is the Live TV Gateway segment, containing a significant number of sites offering live free-to-air and pay TV in addition to other content, which is growing the fastest.

The largest individual site is one in the P2P Community segment. Sites in this segment generally make all forms of content, except live TV, available to download. Downloads allow the user to obtain a full copy of the file which they can then view offline or copy for each of their various gadgets. Unlike streaming, downloads can be obtained independent of the speed of the user's internet access, enabling the highest quality of experience.

Many sites also offer streamed content for the user to consume. This is obviously required for live TV but can support other types of content such as music or video.

We investigated where and how the content was hosted and found that both Live TV Gateway and P2P Community sites, the two largest and fastest growing segments, tended to link to content on other sites or services rather than host the content.

These two segments use quite different architectures to achieve this: Live TV Gateway sites deliver the content from one central server to which they link, whereas P2P Community sites offer links to the files which are served from a distributed array of servers or other users within the community.

**Arriving on the sites**

This study also examined referral data on how users arrive at sites considered to be infringing. It shows that different kinds of sites are reached in quite different ways.

Users of sites in the Live TV Gateway, P2P Community and Music Transaction segments were all more likely to have arrived directly without first visiting any other internet sites than was the case with the other three segments.

Users were more likely to have visited a search engine prior to arriving on a Music Transaction site than was the case with the other five segments.

Live TV Gateway users were most likely to have visited a social network prior to their visit to the site we examined. These sites were also the most likely to have a social networking presence, in the form of a social networking 'action' icon, for example Facebook 'like' buttons, Twitter 'tweet' button or similar.

Prior to their visit, users of Embedded Streaming and Rewarded Freemium sites were more likely to have visited other sites that don't fall into the social or search categories than was the case with the other 4 segments.

**Location**

We examined the geographical location of the sites IP addresses and found two notable facts: sites in the 'Music Transaction' segment were far more likely to be hosted in Russia than any other segment, and a disproportionate number of sites in the 'Rewarded Freemium' and the 'Embedded Streaming' segments were hosted in the Netherlands. The UK is a significant home to only a relatively small proportion of one segment: P2P Community, but these types of site appear to have high numbers of users and are growing.

This report provides a snapshot of the market taken in April/May 2012 and is intended to inform debate about how to address online copyright infringement. More can be done in terms of data: while we have analysed the growth and decline in user numbers, as a snapshot, the report is unable to evaluate other changes in the market.

This report provides a baseline from which to monitor the market. Detica believes that with the addition of time-series data, a full picture of the market and the segments respective trajectories can be realised.

**6: Embedded Streaming** 6.4K – 470K

- Provides hosting where users can upload content , and where others can stream the content from. Allows users to embed content on their own sites and on 3rd party sites and on forums.
- Contributors that upload content can receive a financial reward.
- Predominately advertisement funded, with some sites providing the option to donate.
- Advertising is generally provided by organisations not affiliated with the Ad Choices scheme.
- Comparatively high level of hosting in the Netherlands.

**1: Live TV Gateway** 9.5K – 1.1M

-The sites predominantly offer links to streams of live free-to-air and pay TV. These sites offer above average levels of games and eBooks, as well as other content in lower proportions.
- The sites typically provide links to downloads or streams. The content is centrally hosted (as opposed to using P2P) in a different location from the site.
- Predominately advertisement funded with some donations. Typically free to the user.
- Rapid growth in last year.
- Most likely to have a mobile site and a social networking presence.
- Users often arrive after typing the address into the browser.

**5: Rewarded Freemium** 20K – 1.5M

- Freemium service, meaning users can get basic access for free, funded by advertising, and pay a subscription for more.
- Users who provide content for others are rewarded financially for their contribution.
- This content is uploaded to the sites own servers where it is downloaded by others, unlike P2P where content goes user to user.
- Typically music content.
- Users likely to arrive through links from other sites.
- Large, but declining user base.

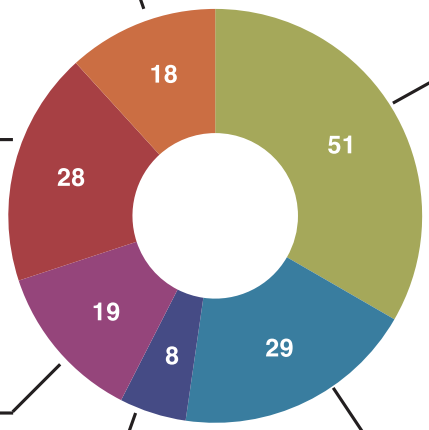



Chart labels are the **number of websites** in each segment

Values in title boxes are the **range of unique UK visitors per month** for sites in that segment. **Growth arrows** indicate whether global segment PageViews are increasing (red) or decreasing (green) based on data for February 2011-2012 (Source: Google Ad Planner).

**4: Music Transaction** 6.5K – 33K

- User is able to buy music to download from the site's own servers. Also offer some games and eBooks.
- Likely to have social networking presence and discovery via search is relatively high. Returning users often type the address directly into the browser.
- Content hosted on sites on servers. Relatively large proportion hosted in Russia.
- All have card processor logos on payment page.
- Small, declining user base.

**2: P2P Community** 10K – 2.1M

- Well organised range of content types with the exception of live free-to-air and pay TV, offered free to the user.
- Engages user with Forums and ability to comment on content.
- Facilitates downloading of content via P2P or distributed servers.
- Heaviest dependency on advertisement and donation funding
- The advertising is largely provided by organisations not affiliated with the Ad Choices scheme.
- Sustained growth over five years.
- Direct access levels very high
- Europe appears to be the main home of these sites

**3: Subscription Community** ~7.8K

- User pays a subscription fee for a well organised range of content types with the exception of free-to-air and pay TV.
- Tired community, where users can gain increased levels of access/content via payment or contribution
- Engages user with Forums and ability to comment on content.
- Facilitates downloading of content via P2P or distributed servers.
- Only 12% of sites carry adverts.
- All have card processor logos on payment page.

*Figure 1-2: The six business models for copyright infringement*

*The numbers of websites identified in each segment in the donut chart presented in Figure 1-2 above describe only volumes of websites that fell in each segment after a systematic sample of websites had been taken for the segmentation. This can be used as a proxy for the presence of total numbers of different websites available to the user. However, no inference can be drawn on the size of the market for each segment in terms of users, importance, market value or loss to rights holders. A small segment, above, might have a lot of business but be limited to a few websites, where a much larger segment in terms of the numbers of websites may undertake less business.*

# Contents

# 1 Context and terms of reference

BAE Systems Detica (Detica) was commissioned by PRS for Music and Google UK (Google) to investigate the characteristics of websites that are alleged to infringe copyright.

There have been many studies and surveys of online copyright infringement but this report is the first to provide a purely data-driven description and analysis of the online copyright infringement industry.

Detica was provided with a list of websites by The Federation against Copyright Theft (FACT), The British Phonographic Industry (BPI), The Football Association Premier League (FAPL), UK Interactive Entertainment (UKIE), PRS for Music and the Publishers Association. The rights holders believed the sites contained in these lists to be significantly facilitating copyright infringement. The lists formed the basis for the subsequent data-driven analysis. The lists themselves were provided confidentially and are not detailed in this report. Detica does not confirm or deny the claims made by the rights holders as to whether these sites can be said to facilitate copyright infringement.

The aim of the study was to measure and analyse these websites in a way that was objective, evidence-based and determined by the data. The goal was to create a map of the alleged copyright infringing market, based on evidence, that could provide industry and policymakers with insight into how these sites operate.

# 2 Results

Detica's data-driven segmentation identified six clear segments within the 'copyright infringement industry'. Each of these segments contain sites with business models similar to other sites within their segment but significantly different from sites in other segments.

In the same way that collecting data about furniture retailers might show that there are a range of quite different business models in that industry (Swedish flat-pack giants, sofa superstores, antique shops, hi-design boutiques, etc), Detica's data-driven analysis of the sites identified by rights holders shows that they cluster into six segments; in effect six types of business model for copyright infringement. In this chapter we describe those segments and the metrics collected in the analysis.

Detica used over 100 different metrics in this study. These metrics gathered information on the size and growth of each site, the type of content offered, how users navigated to them, their network arrangements, their sources of revenue, their community and their social features. A full list of metrics can be found in Appendices G and H.

The majority of the metrics were collected on a yes/no basis e.g. Does a site offer music content? Does a site have a social networking presence? etc. In addition, a number of non-numeric metrics were also used to aid the description of our segments. These categorical metrics include:

- IP Address Location – The country location of 'A record' (IP address).
- Top Level Domain Location – The country location of the Top Level Domain.
- Ad Provider Type – Is advertising present? If so, is it provided by Ad Choices?
- Card Processor Logo – Does a payment page exist? If so, are the logos of Visa, MasterCard or American Express present?
- Electronic Payment Provider Logo – Does a payment page exist? If so, is the PayPal logo present?

Six segments were identified using a statistical method, effectively grouping sites with similar characteristics. Examining these characteristics enabled Detica to provide a clear profile of each segment.

The following section of this report sets out the profiles for each of the six segments, in the following manner:

1. Segment name – based on discussion between Detica, PRS for Music and Google.
2. Description of operating drivers and characteristics – based on the underlying metrics.
3. Key metrics for the segment:
   - Standard – Size of the cluster, range of unique UK visitors per month and a growth indicator. The growth indicator is based on the global change in activity on the websites in terms of page views. It cannot be compared directly with unique UK visitors but it does provide a relative view of change.
   - Numeric – Selected significant metrics displayed in a chart showing the segment average compared to the population average. It should be noted that some metrics are relative values, and that all the metrics displayed have been normalised for comparison between different segments.
   - Categorical – The two most significant non-numeric metrics.

## 2.1 The six segments

Detica analysed the six segments and identified the following operating drivers for each segment (see Appendices A and B for comparisons of all metrics):

### Segment 1: Live TV Gateway

This segment contains 33% of the sites examined and is the fastest growing segment, with an average increase in global page views of around 61% (in the twelve month period studied). The segment is mid-high in terms of volume when compared to the other segments with up to 1.1M unique UK users per month on one site alone.

- The sites offer links to streams of live free-to-air and pay TV.
- These sites offer above average levels of games and eBooks, as well as other content in lower proportions, but their stand out feature is live TV.
- The sites typically provide links to downloads or streams. The content is centrally hosted (as opposed to using P2P) in a different location from the site.
- Predominately advertisement funded with some donations. 67% have adverts with 86% of those ads served by networks not affiliated with the Ad Choices scheme.
- Typically free to the user.
- Rapid growth in last year.
- Most likely to have a mobile site and a social networking presence.
- Compared to the other segments Live TV Gateway has very high levels of direct access and referrals from social networks. It also has the highest level of social network presence. Search referral, albeit to a lesser degree, is also above average in this segment.
- More of these sites are in the US than any other single country.

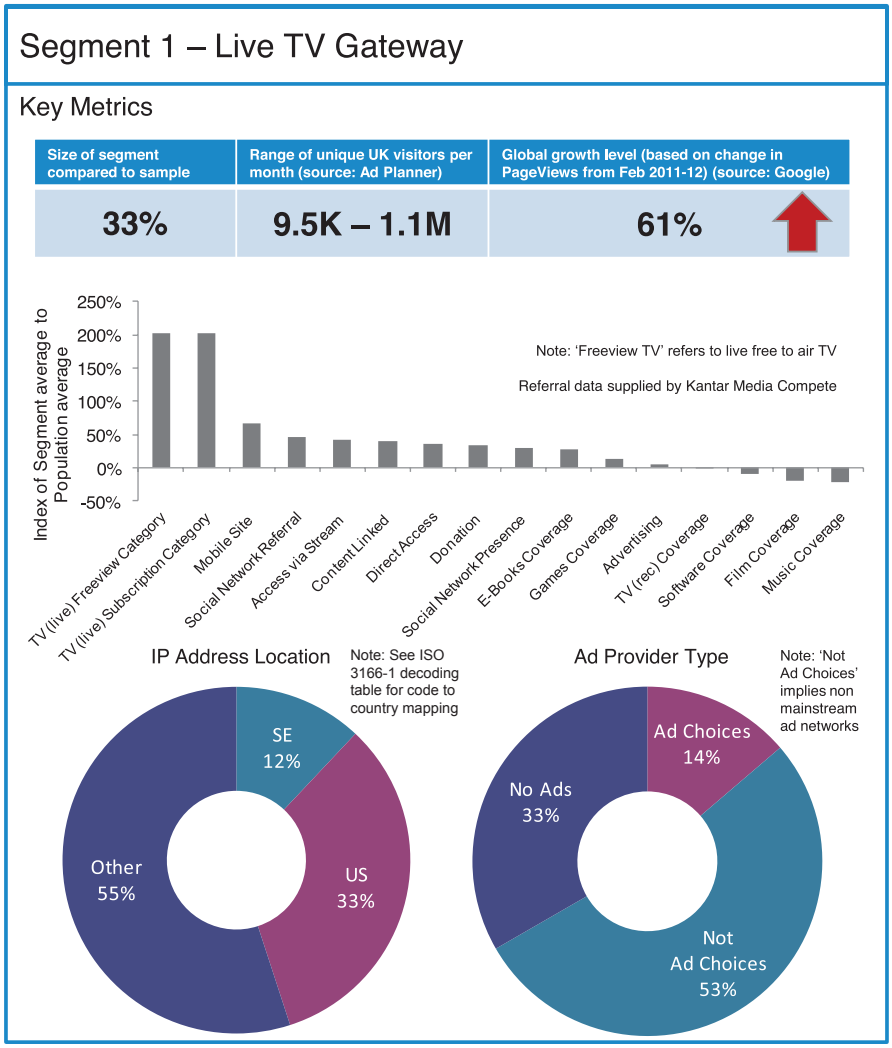



*Figure 2-1 : Graphical representation of Segment 1 – Live TV Gateway*

## Segment 2: P2P Community

The second fastest growing segment, with an average increase in global page views of around 17% per year. The segment contains 19% of the sites, and at least one site in the segment could be considered high volume, containing around 2.1M unique UK users per month.

- Well-organised range of content types offered free to the user. Content available does not include live free-to-air and pay TV.
- Engages user with forums and ability to comment on content but they have relatively low levels of user login or user ratings.
- Facilitates downloading of content via P2P or distributed servers.
- Heavy dependency on advertisement and donation funding. 86% have adverts and 84% of adverts are served by Ad networks not affiliated to the Ad Choices scheme.
- Sustained growth over past five years.
- Direct access levels very high compared to other access methods.
- Europe appears to be the main home of these sites, including the United Kingdom.

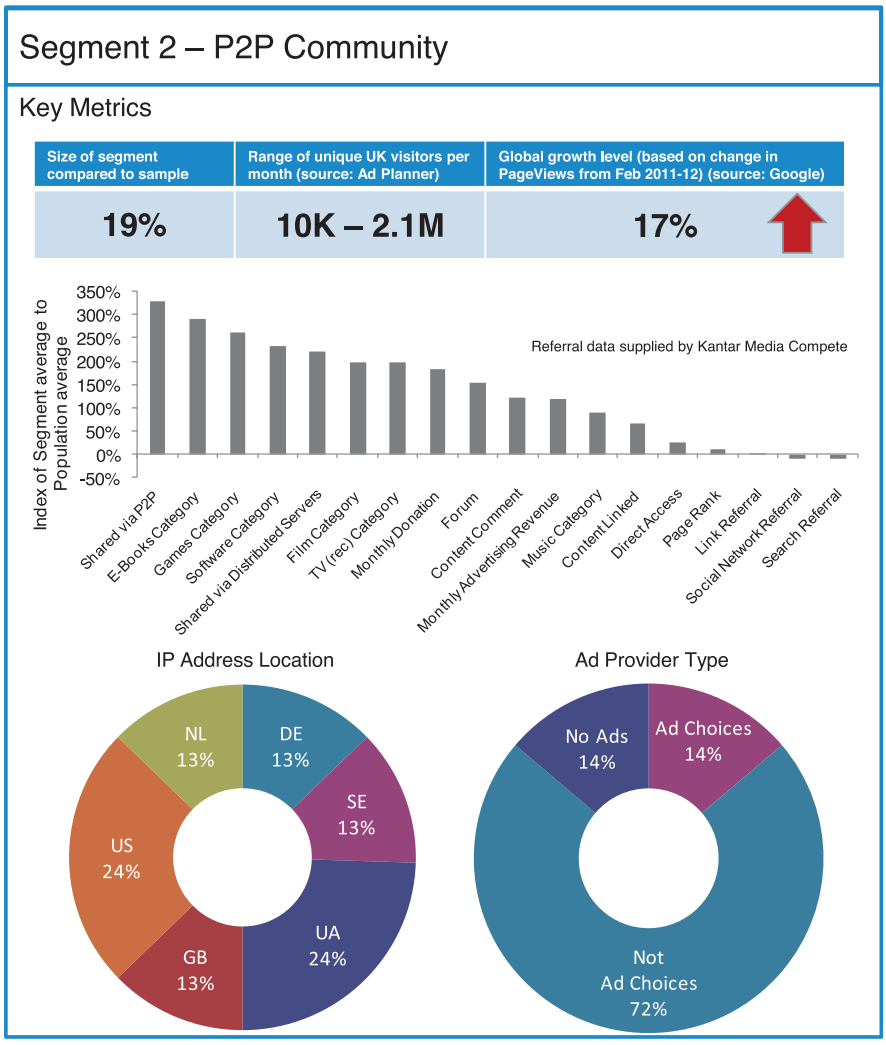



*Figure 2-2 : Graphical representation of Segment 2 – P2P Community*

## Segment 3: Subscription Community

This segment contains the fewest sites, with only 5% of the sites sampled. The usage volumes and level of growth present for sites across this segment were unclear due to a lack of available data.

- User pays a subscription fee for a well-organised range of content types. This content does not include live free-to-air and pay TV.
- These sites offer a tiered community model, in which users can earn different levels of access and content quality through payment or uploading of content to the site.
- Users are much more engaged than in other segments – with relatively high levels of user login, user rating systems and ability to comment on content.
- These sites have the highest levels of donation and the second highest of level of monthly subscription of any segment.
- Only 12% of sites carry adverts.
- Facilitates downloading of content via P2P or distributed servers.
- These sites do not have any dominant forms of referral but they do have a high number of other sites linking in to them (Alexa ranking).
- Around two-thirds of the sites contained clearly visible payment pages, and debit/credit card payment logos were clearly present on all of them. The sites that did not have clearly visible payment pages and logos, may have payment mechanisms but they were not visible.

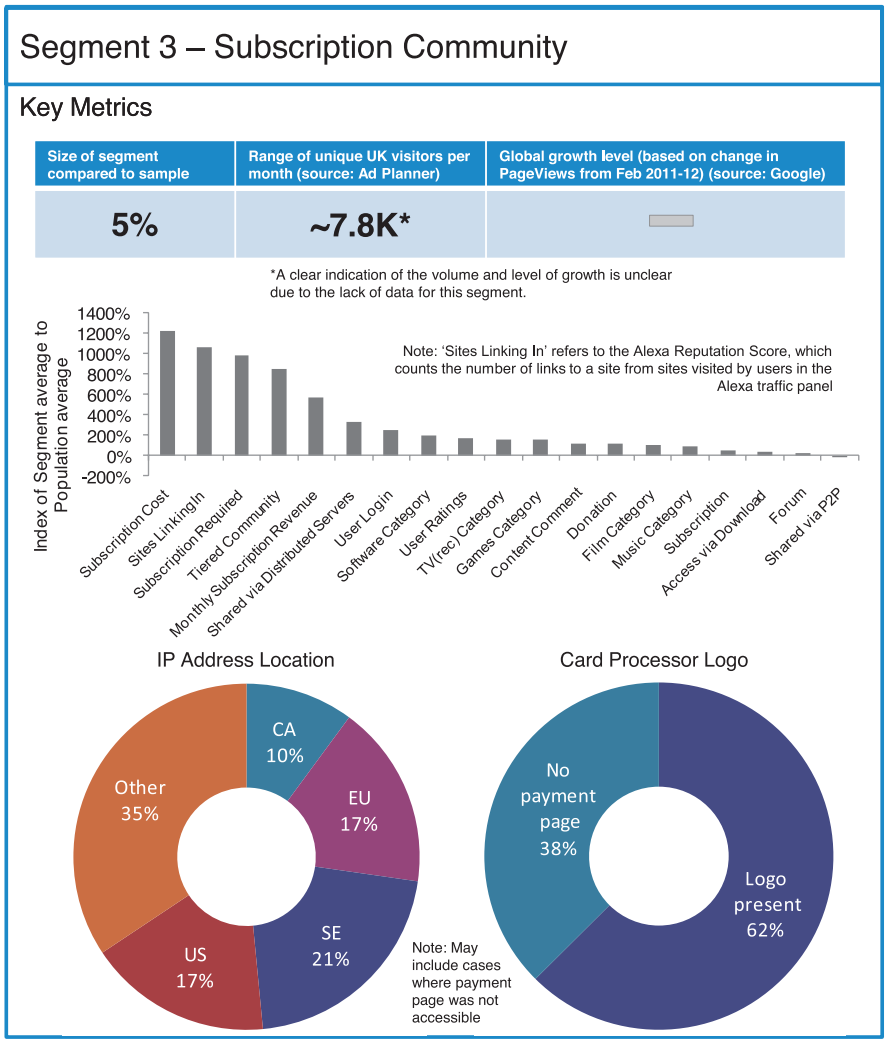



*Figure 2-3 : Graphical representation of Segment 3 – Subscription Community*

## Segment 4: Music Transaction

The fourth segment contains around 13% of the sites examined. On average these sites are marginally declining, with an average decline in global page views of 19% per year. Excluding Segment 3 due to the lack of available data, these sites contain the lowest average UK user volume, only up to 33K per month.

- The standout feature of these sites is that they are transaction-based: users buy content.
- There are some games and ebooks available but music is by far the most significant type of content on offer.
- Content hosted on site's own servers. Relatively large proportion hosted in Russia.
- All have debit/credit card payment logos on any visible payment page.
- Higher than average likelihood of users arriving directly or after visiting search engines.
- Second highest levels of social networking presence and referral (after Live TV Gateway).
- Small, declining user base.

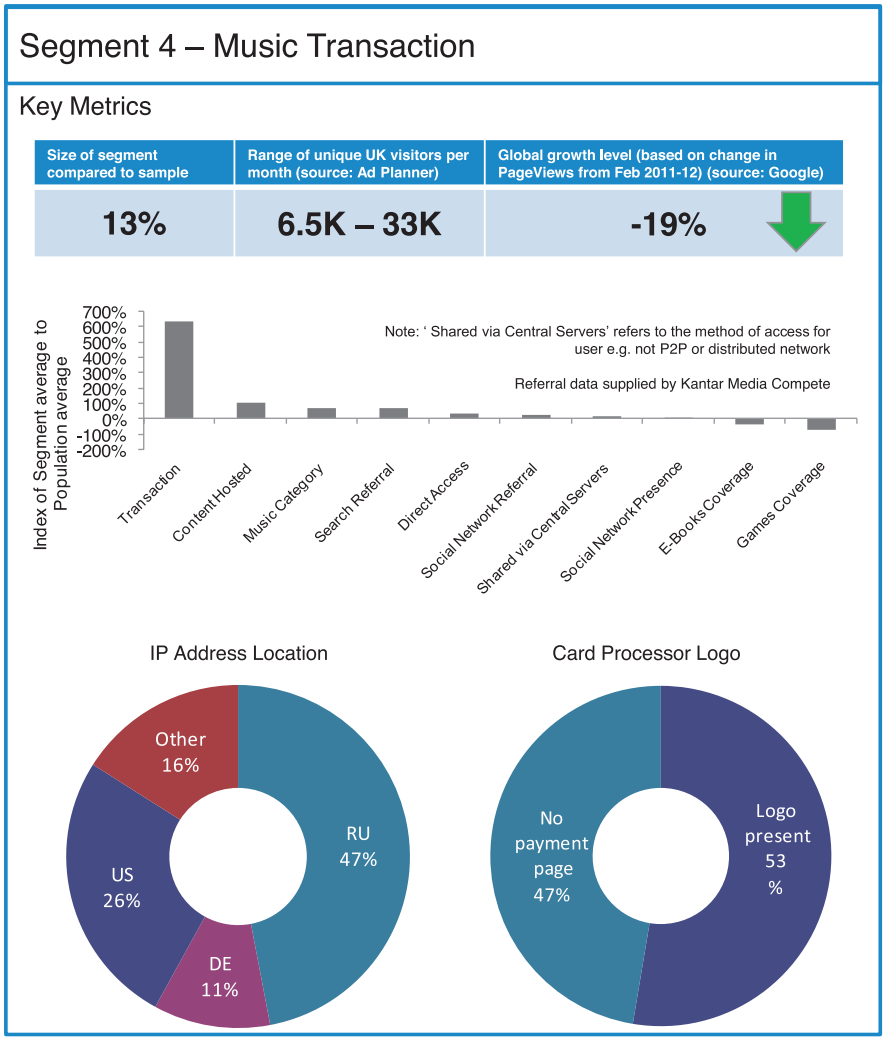



*Figure 2-4 : Graphical representation of Segment 4 – Music Transaction*

## Segment 5: Rewarded Freemium

This segment contains 18% of the sites. These sites have both free and paid options for accessing content. The segment is on the decline, with an average drop in yearly global page views of around 4%. These sites have a high volume of UK users, up to 1.5M per month on one site, and a number of these sites provide financial rewards to contributors (e.g. for users who have content supplied by them downloaded by others).

- The standout features of these sites are that they offer financial rewards to uploaders and operate a freemium model.
- These sites offer Freemium services funded through advertising, meaning users can get basic access for free, and a paid subscription options for enhanced services.
- Electronic payment provider logos were present on 61% of sites, with debit/credit card payment options present on 46% of sites.
- Users who provide content for others are rewarded financially for their contribution.
- This content is uploaded to the sites' own servers where it is downloaded by others, unlike P2P where content is transferred from user to user.
- These sites typically offer music content.
- The user more likely to arrive through links from other websites. Lower than average levels of search referral, social networking and direct access.
- Large, but declining user base.
- The Netherlands and the United States appear to be the main locations of these sites, accounting for a 31% and 29% share respectively.

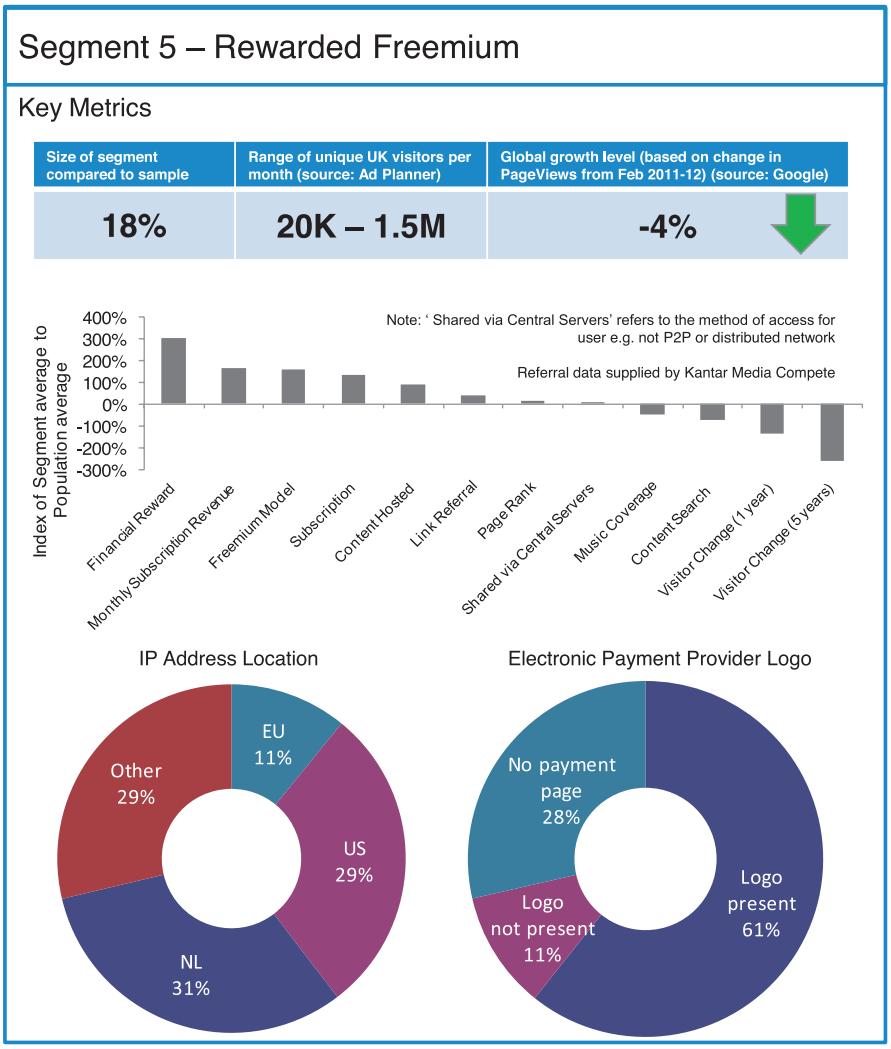



*Figure 2-5 : Graphical representation of Segment 5 – Rewarded Freemium*

## Segment 6: Embedded Streaming

The final segment, containing 12% of sites, is declining the fastest. Sites in this segment are on average mid-volume with a range of 6.4K to 470K unique UK visitors per month. On average, sites in this segment are declining, in terms of global page views, by 33% per year.

- Provides hosting where users can upload content, and where others can stream the content from.
- Allows users to embed content on their own sites, on 3rd party sites and on forums.
- Contributors that upload content can receive a financial reward.
- Advertisement funded, with some sites providing the option to donate. 89% of sites carry ads, with all adverts served by Ad networks not affiliated to the Ad Choices scheme.
- Comparatively high level of hosting in the Netherlands.

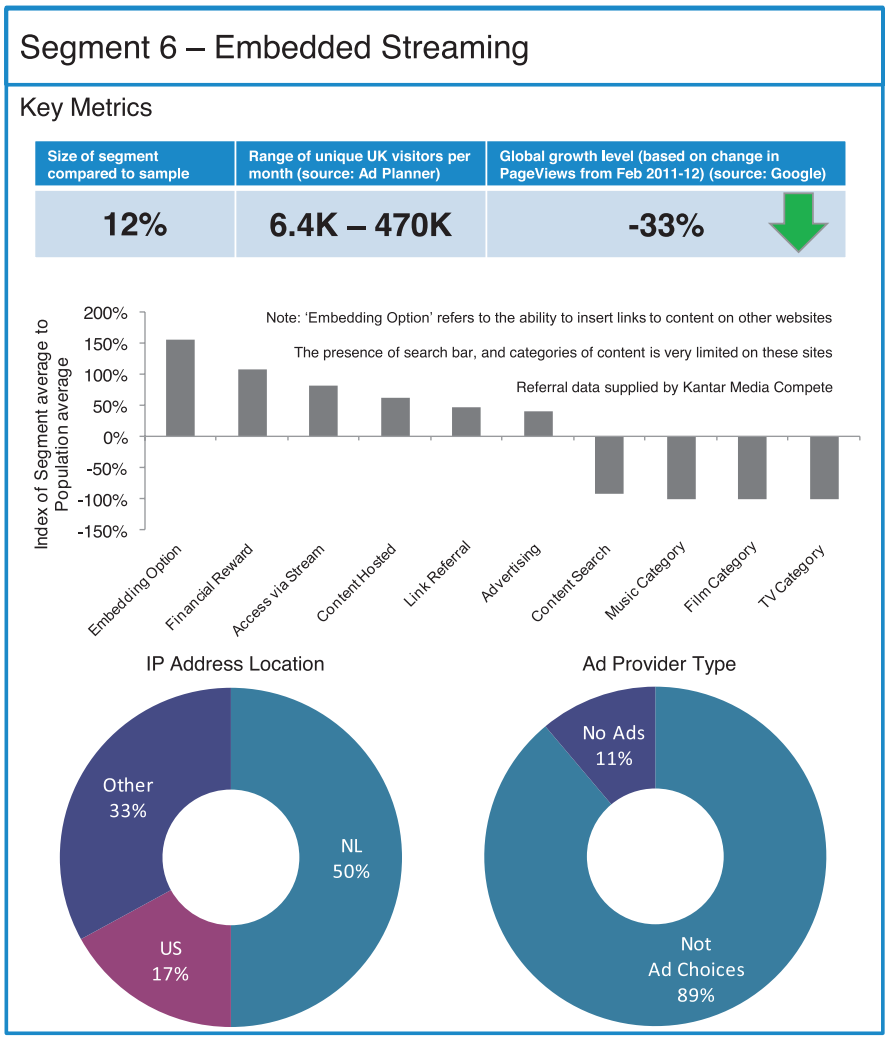



*Figure 2-6 : Graphical representation of Segment 6 – Embedded Streaming*

*The following sections of this report detail the analysis of the metrics, the methodology used to sample and segment the data, and the potential next steps for this research. Detailed appendices are also presented for reference.*

# 3 Analysis

In this section, we present our findings for each of the categories we studied. In all cases, the metric for a given segment is compared to the average for all sites and normalised so that the segment with the highest likelihood of the characteristic scores 1.

Because each category is normalised by a different ratio, it is not possible to compare the normalised values of two metrics: only comparisons between one segment and another within a metric are valid. For example the scale of the music coverage metric is very different to the scale of the software coverage metric, thus cross comparisons cannot be drawn.

## 3.1 Content

We looked for a range of popular content on each site to understand what they offered and the amount of choice the user had available for that content type. In Figure 3-1 we have plotted the type and coverage of content available from each segment.

Figure 3-1 also shows how you will find Live TV content on sites found in the Live TV Gateway segment, with all other segments scoring zero. It also shows how you are more likely to find games and ebooks on sites in the Live TV Gateway segment than anywhere else, with Live TV Gateway scoring 1 for each of these categories. It shows how recorded TV is also quite likely to be found on sites in this segment, with a score of 0.82, although not as often as on P2P Community sites, which scores 1 for this category.

Figure 3-1 shows how you are very likely to find most types of content except Live TV on P2P Community sites and to a slightly lesser depth on Subscription Community sites.

Music Transaction sites seem to focus on music while also having some ebooks and games available to their customers. Rewarded Freemium sites appear to concentrate only on music.

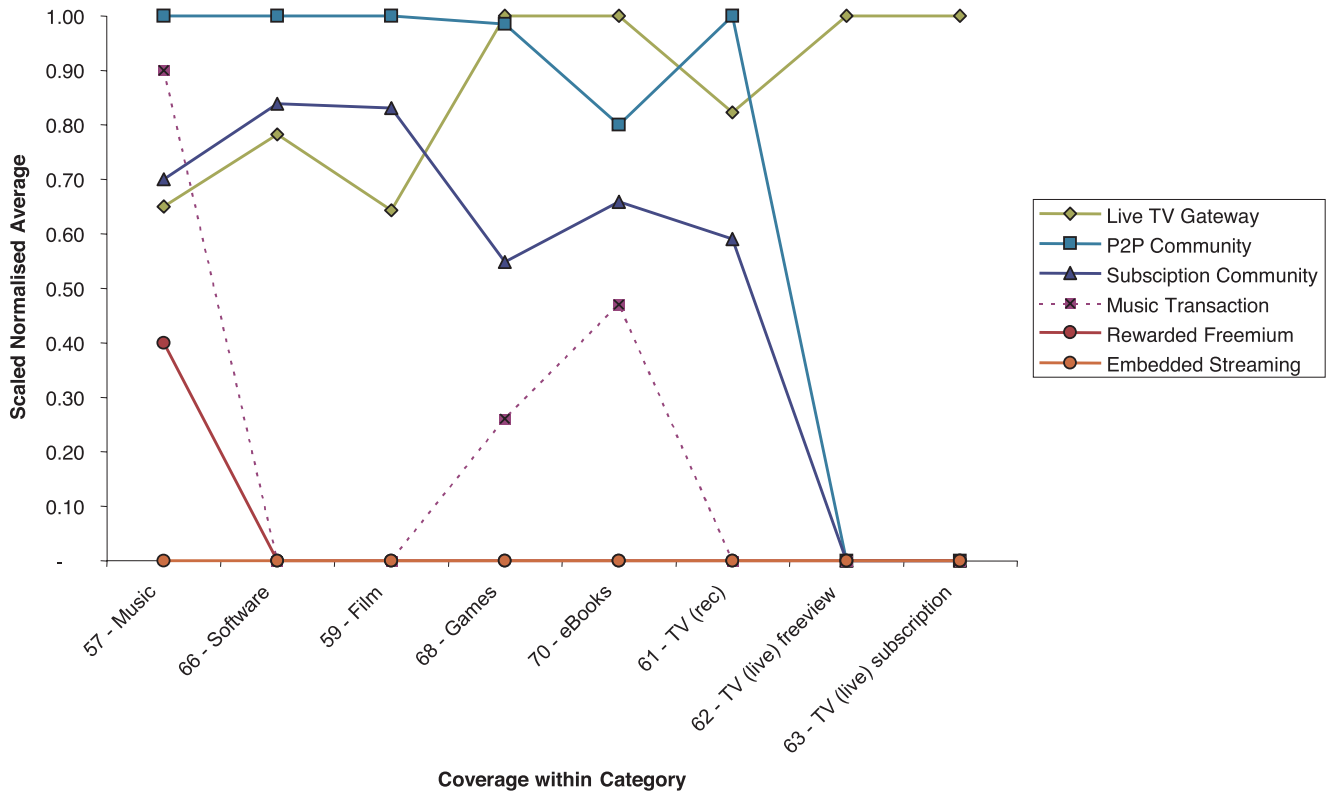



*Figure 3-1: Chart showing scaled normalised averages of content coverage metrics for each segment*

## 3.2 Navigation to the Site

We investigated the user's journey to each site using Kantar Media Compete data. While this does not show us the page that the user linked from directly, it does allow us to see categories of sites that the user had visited in the 30 minutes prior to arriving at one of the sites we were looking at.

We found that users of Live TV Gateway, Music Transaction and P2P Community sites were more likely to arrive directly, after 30 minutes or more of no online activity at all (Direct Access), than they were to sites in other segments.

Social Networking was also more likely to have been accessed prior to users arriving at Live TV Gateways and search more likely for Music Transaction sites.

Embedded Streaming, Rewarded Freemium access was more likely from users who had been browsing other pages than was the case with Music Transaction and Live TV Gateway sites. This suggests that these users were led to the sites we examined by links from the sites that they visited.

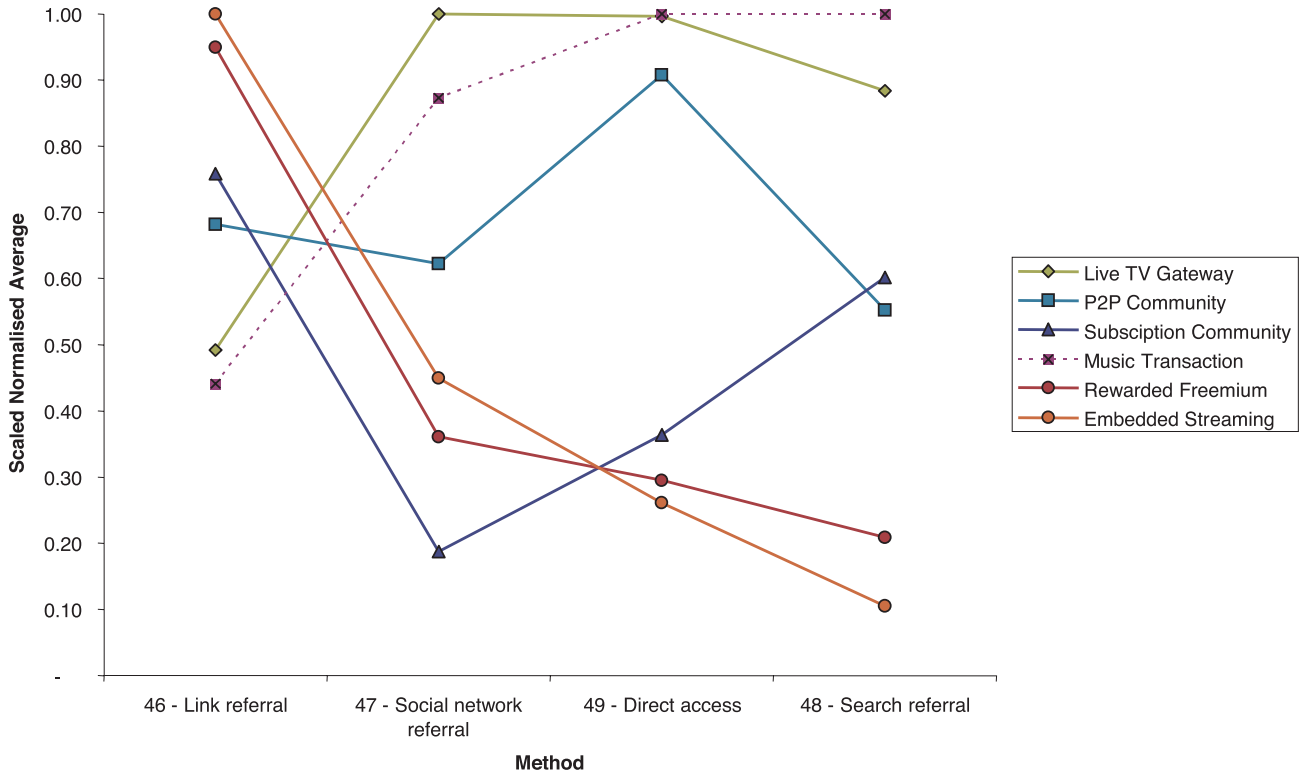



*Figure 3-2: Chart showing scaled normalised averages of site navigation method metrics for each segment*

## 3.3 Network Arrangement

We investigated the network arrangements of the sites we looked at. We categorised sites depending on whether they used P2P or other distributed server configurations that break up the raw file and deliver pieces of from different sources. The alternative arrangement to that is provided by a central server where the whole file resides ready to be accessed.

We found that P2P and Subscription Community sites tended to use the distributed arrangement while the other four segments favoured centralised content hosting.

We examined who owned the domain names where the content was found and whether the content was hosted by the site itself or stored on a linked site somewhere else.

We found that although Live TV Gateways favoured one Central Server, this was not a server that they appeared to own. Their users follow links to content that is stored on a central server elsewhere.

Music Transaction, Rewarded Freemium and Embedded Streaming hosted content on their own central servers.

P2P and Subscription Communities rely on links, and client software, to find the various pieces of the file that the user is downloading, from these distributed locations.

The figure also shows whether the content is available to download or stream or both.

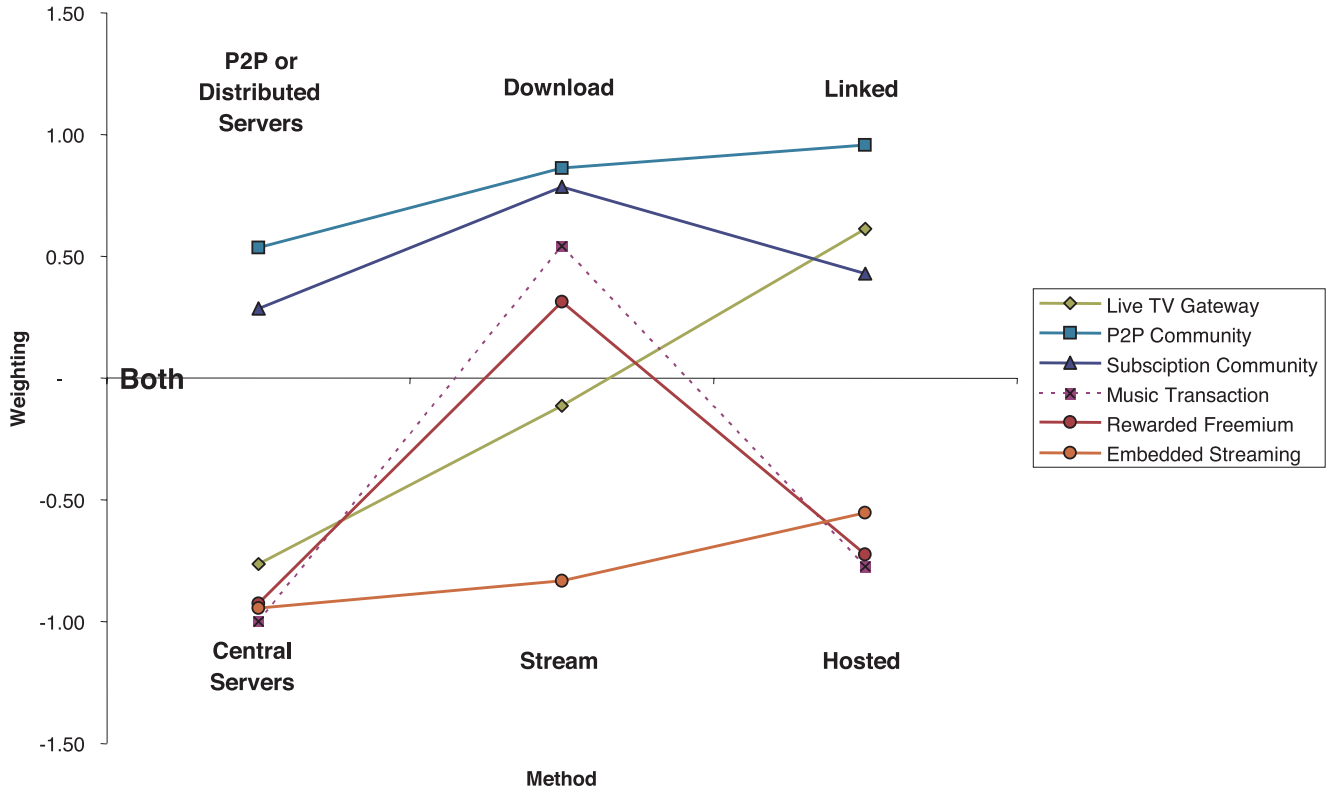



*Figure 3-3: Chart showing weighting of site network arrangement metrics for each segment*

## 3.4 Sources of Revenue

We looked for evidence to indicate how the sites are funded. We looked for payment gateways that supported transactions, subscriptions or donations and whether advertising was present.

Figure 3-4 shows the relative likelihood of each source being used by each segment. For example, Music Transaction sites were characterised by their use of transaction based pricing which was not present on other sites.

Community sites (Subscription and P2P) were the most likely to solicit donations.

Advertising is an important source of funding for many sites as described elsewhere, with Embedded Streaming and P2P Communities depending even more on ads than other segments.

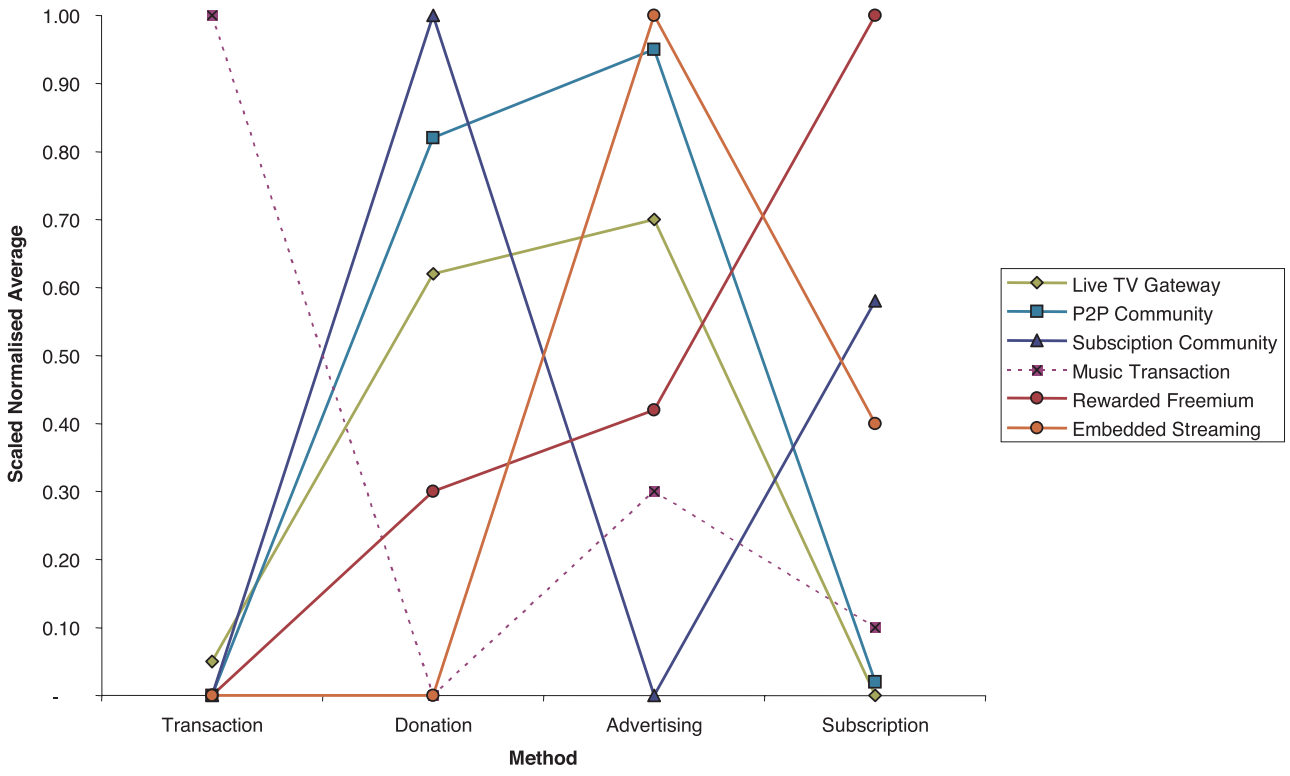



*Figure 3-4: Chart showing scaled normalised averages of site source of revenue metrics for each segment*

## 3.5 Community and Social Features

We looked for various features to understand the level of engagement with users. Subscription Communities were characterised by their use of a tiered structure whereby the more a member contributes, the better their level of access.

We looked for evidence of forums and the ability for a user to comment and interact with other users which helped us to further indentify sites with a strong emphasis on creating a community. We studied whether sites paid contributors for content too, either in cash or in kind.

We found that the Live TV Gateway sites in particular were exploiting social networks and mobile to reach out to their users.

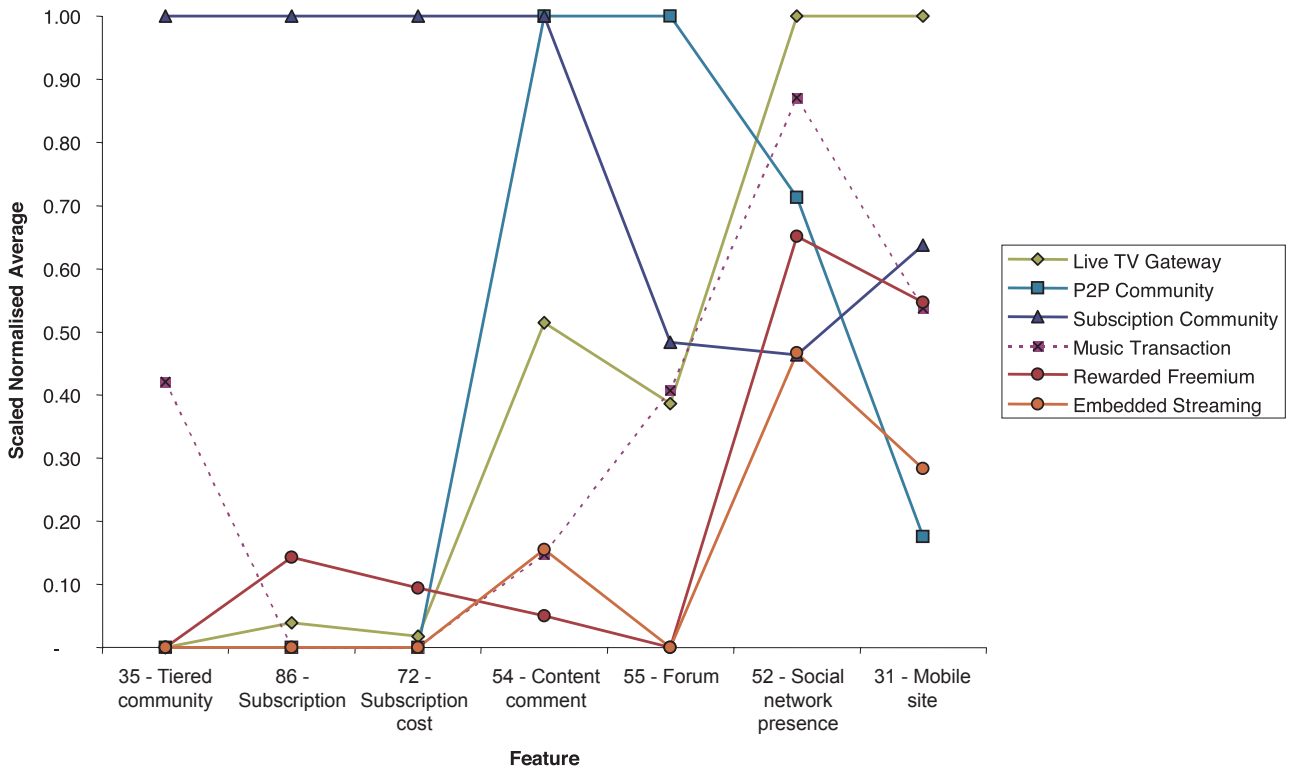



*Figure 3-5: Chart showing scaled normalised averages of community and social feature metrics for each segment*

## 3.6 Cost to User, User Base and Growth

We looked at the service types and whether the user had to pay for access and plotted that against the relative size of the user base and the growth pattern of each segment.

We found strong indications that free sites are collecting the largest user bases and growing the quickest. Subscription services appear to be quite small while freemium services where users can access some services for free, or pay for enhanced features appear to be experiencing the sharpest decline.

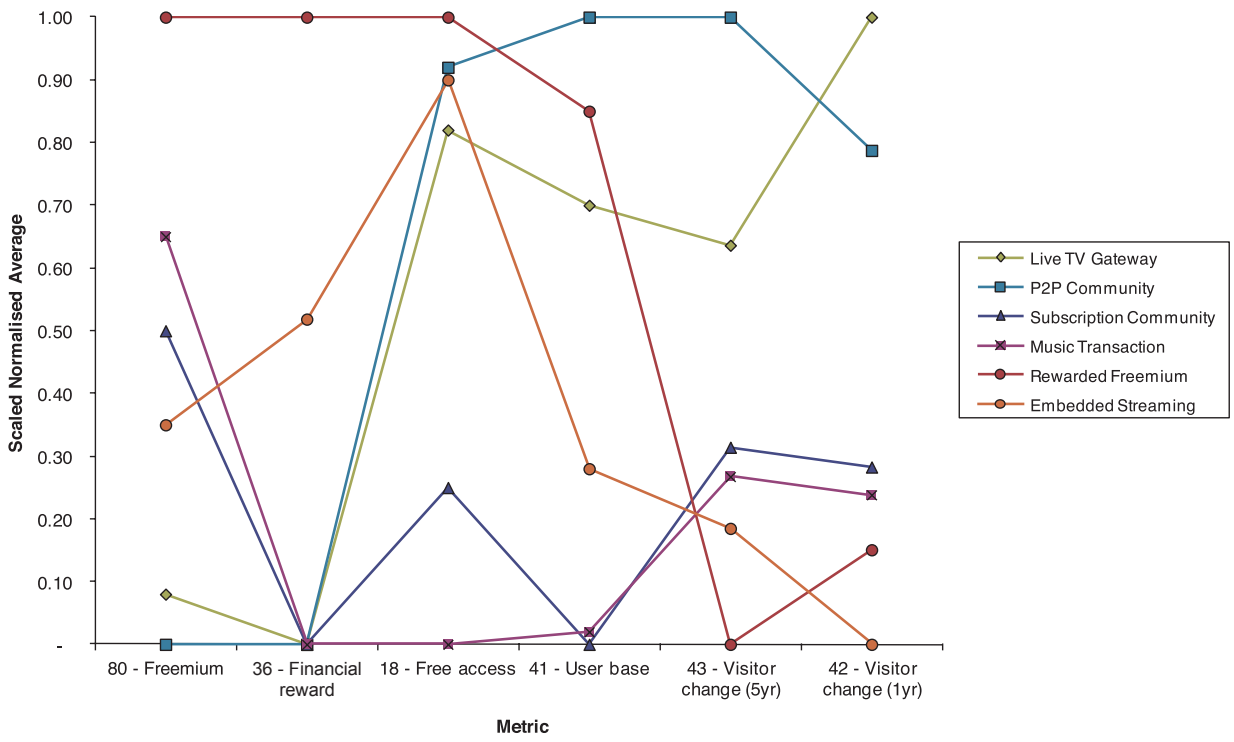



*Figure 3-6: Chart showing scaled normalised averages of metrics associated with cost to user, user base and growth for each segment*

# 4 Methodology

This chapter provides an overview of the methodology that was used to derive the segments, and will provide detail of the process shown in Figure 4-1:

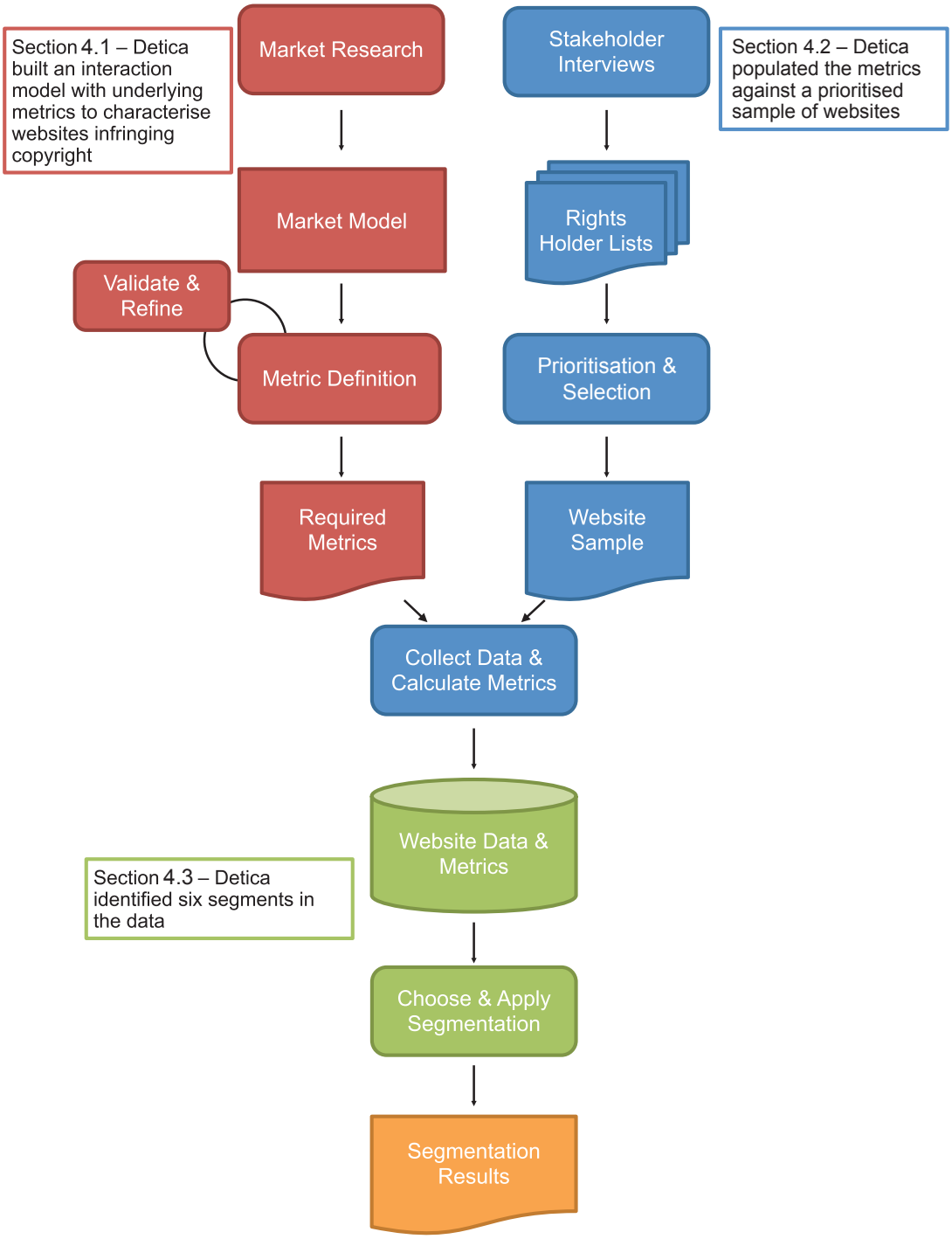



*Figure 4-1: The process used to segment the market into business models*

## 4.1 Copyright infringement market model

We required a way to identify relevant data to collect that could be used to effectively segment websites that are seen to infringe copyright.

We used domain expertise and market research to create a market model allowing us to describe the websites considered to be infringing copyright. This market model looks at the actors in the market, the actors' personas, and the actors' motivations.

Using the motivations we identified a set of attributes that allowed us to measure these motivations, finally resulting in a set of metrics we wanted to calculate for each website. These metrics were calculated and used in segmentation described in Section 4.3. This process is depicted below in Figure 4-2.

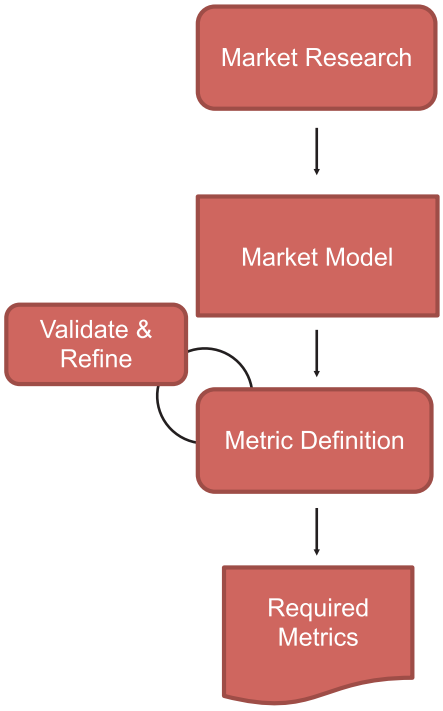



*Figure 4-2: Creating the market model and defining the metrics to enable the segmentation*

The market model forms the foundation of the analysis we undertook. We wanted the model to take into account the ecosystem in which these websites were being used and operated, and therefore the segmentation would be based on an unbiased and fully rounded set of metrics.

### 4.1.1 Relevant case law and pilot websites

To populate the model we reviewed UK legislation with specific interest to this study to understand how the constituent players in the market operated and used four pilot sites to understand the motivations of these players.

We reviewed the Digital Economy Act 2010 and the Copyright, Designs and Patents Act 1988, as well as the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPs) administered by the World Trade Organisation. Within this legislative framework the Newzbin judgement[1], Newzbin 2 judgement[2] and The Pirate Bay judgement[3] are of particular relevancy.

The four pilot sites selected are described in Table 4-1. PRS for Music and Google chose the pilot sites to be representative of a range of technological implementations and content types.

| Pilot site type | Technology | Primary Content |
|---|---|---|
| Torrent index | Index, Torrent P2P | Music, Film, TV, Software, Games, Books, other |
| Usenet reporting | Index, Usenet | Music, Film, TV, Software, Games, Books, other |
| Sports streaming | Index, Streaming | Live Sport |
| Invite forum | Invitation | None |

*Table 4-1: A description of each of the pilot sites used to test the model*

The following sub sections outline the components of the model as shown in Figure 4-3 and are summarised as follows:

- The key **actors** in the market that are involved in and impacted by the websites;
- The **personas** that actors played in the market (extremes of character for each actor);
- The **motivation** that led them to be involved in the market; and
- The **attributes** that allow us to measure the motivations.

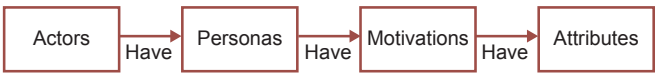



*Figure 4-3: The market components that enabled us to build a robust model*

*1 Twentieth Century Fox Film Corporation and others v Newzbin Limited [2010] EWHC 608 (Ch), [2010] All ER (D) 43 (Apr)*
*2 Twentieth Century Fox Film Corporation and others v British Telecommunications PLC [2011] EWHC 1981 (Ch)*
*3 Dramatico Entertainment Limited & others v British Sky Broadcasting Limited & others [2012] EWHC 268 (Ch)*

### 4.1.2 Actors and personas

We used the pilot sites, listed above, to produce the list of actors and their interactions. We identified five key actors: Consumers; Contributors; Rights Holders; Site Owners, and Service providers. The interactions are described in the model below, Figure 4-4.

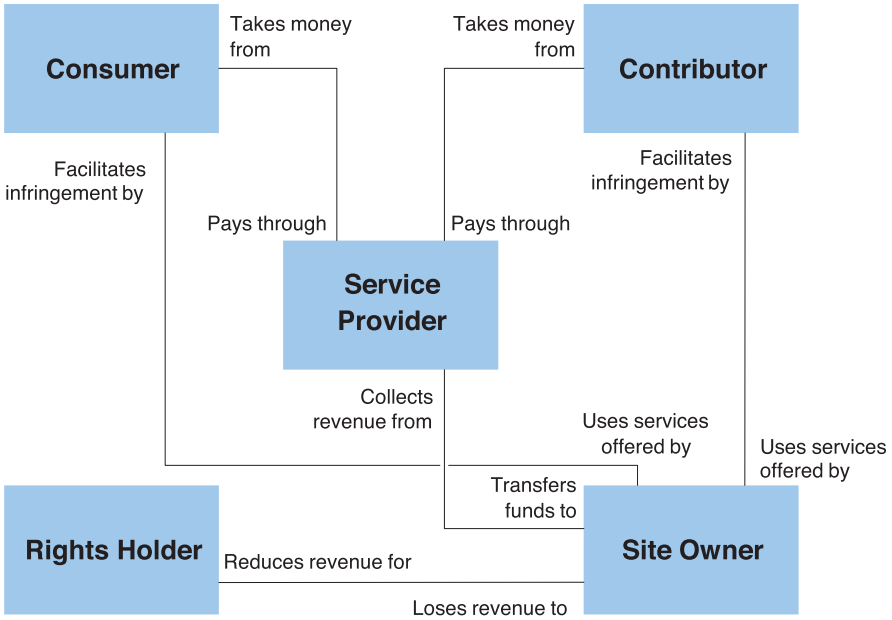



*Figure 4-4: The actors and their relationships who have a role in the websites*

Further researching the actors, the extreme roles of the actors were identified as personas. The personas are defined in Table 4-2.

| Actor | Persona | Description |
|---|---|---|
| Website Owner | Venture capitalist | Interested in making money from operation of the website |
| | Altruist | Believes that facilitating the sharing of unlicensed content is a legitimate activity |
| | Innovator | Wants to bring new technologies and ideas to market without core financial motivation |
| Contributor | Accidental | Doesn't realise they are sharing content with others |
| | Enthusiast | Shares content to impress people and show that they can |
| | Altruist | Believes that facilitating the sharing of unlicensed content is a legitimate activity |
| | Profiteer | Motivated by the incentives offered by being an uploader of popular content |
| Consumer | Unintended | Didn't realise the service was infringing |
| | Casual | Uses infringing services because everyone else does |
| | Regular | Seeks out free services and think they know how to avoid the associated risks |
| Rights Holder | Amateur | Individual artist seeking to release content in a controlled way |
| | Corporate | Rights holder representing a group of artists and other entities seeking to commercialise content |
| | Independent | A manager of 'a' band, but not signed to a label |
| Service Provider | Non-mainstream ad agency | An agency which specialises in placing ads on a website which is not a member of an industry body or regulator |
| | Mainstream ad agency | An agency which places ads on a website which is a member of an industry body or regulators |
| | Payment provider | Companies who facilitate transfer of funds through transactions, donations or subscriptions |

*Table 4-2: The Actors and their Personas*

#### 4.1.3 Motivations and attributes of the model

We used our pilot sites and case law to understand the motivations for each of the personas. The motivations were both positive and negative and were derived from the following questions:

- Website Owner: Why would they set up the website?
- Contributor: Why would they contribute to this service?
- Consumer: Why would they use this service?
- Rights Holder: Why would they want to stop their content being on the website?
- Service Provider: Why would they provide service to this website?

A full list of motivations against each persona can be found in Appendix D.

For each motivation we defined as set of attributes that we would like to measure. The list of these attributes can be found in Appendix E.

#### 4.1.4 Expert review and validation of the metrics selected

We reviewed this model with experts in the copyright infringement market and their review comments and suggestions were incorporated into the approach described in the following pages. Specifically, we gained peer review input from:

- Andrew Clark, Expert Witness in Computer Assisted Crime, Primary Key Associates
- Simon Morrison, Copyright Policy Manager, EMEA, Google
- Theo Bertram, Policy Manager, UK, Google
- Frances Lowe, Head of Regulatory and Corporate Affairs, PRS for Music
- Will Page, Chief Economist, PRS for Music
- Jeremy Penston, Independent Consultant

We mapped these attributes to the available data to create a set of metrics which could be measured for all websites.

We identified three categories of website data:

- Technical – data relating to the websites technical setup, for example, the information contained within the WHOIS record.
- Usage – data pertaining to the usage of the website, for example, visitor figures, demographics or referrals.
- Direct inspection – data captured through visually inspecting the website and its source code, for example, whether adverts are present, whether users have to log in to access content or whether the site has a forum.

Examples of each type of metric are shown below in Table 4-3.

| ID | Title | Type | Calculation / definition |
|---|---|---|---|
| 19 | A record location | Technical | The country location of the IP address of the 'A' record for the website. |
| 42 | Visitor Change (1yr) | Usage | The number of Pageviews for the website in the month preceding this research minus the number of Pageviews 12 months earlier. |
| 96 | Ad Provider Type | Direct inspection | Positive (equal to 1) if the first display advertisement on a website has the Ad Choices logo on or around it and negative (equal to 0) if not. Note that this metric is not applicable to sites without advertising. |

*Table 4-3: A samples of the metrics used in the model*

### 4.2 Populating the metrics against a prioritised list of websites

In this section we describe the construction of a sample list of websites to be segmented, and describe the process of collecting the relevant data to populate the metrics for each site:

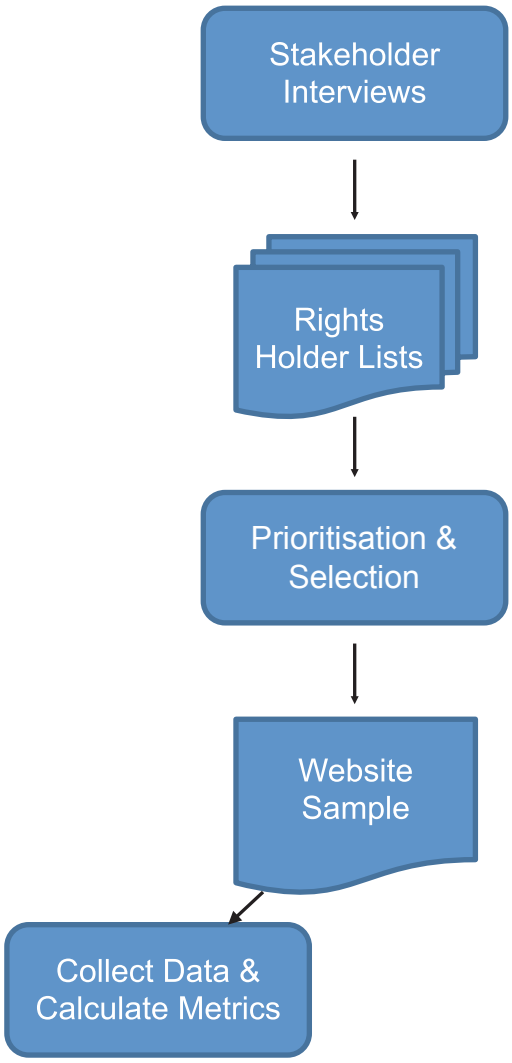



*Figure 4-5: Creating the sample list of websites and collecting the data*

#### 4.2.1 Sample websites to be used as part of the analysis

To establish a list of websites for use in this research, we engaged rights holders representing the creative and content industries. We asked them to provide lists of websites which they considered to be infringing copyright. These lists were an input to the modelling process outlined in this report.

The following representatives of the rights holders were engaged:

- The British Phonographic Industry (BPI)
- The Federation against Copyright Theft (FACT), which was also representing the British Video Association (BVA) and the Motion Picture Association of America (MPAA)
- The Football Association Premier League (FAPL)
- The Publishers Association (PA)
- UK Interactive Entertainment (UKIE)

We would like to thank these representatives for their help and support during this research.

We held interviews with each of the representatives and obtained in addition to their list:

- The methodology for construction of their lists; and
- other research or data sources relevant to the research which they could make available to Detica.

We took the websites obtained and consolidated them, retaining the grouping provided by the representatives of the right holders. We used Alexa Global Traffic Rank[4] to identify the most popular websites in the categories, and then calculated the number of websites required from each category to create a de-duplicated sample list of approximately 150 websites.

This sampling process was designed to ensure that the full range of website types were retained through inclusion of sites from all stakeholder categories, whilst reducing the overall number of websites to a manageable level for data collection purposes.

We formed a sample list containing 153 websites for the 'Training data'. We selected a further 104 websites to be used to validate the segmentation – 'Validation data'.

#### 4.2.2 A process to transform the collected data into the metrics

Obtaining the required data to populate the metrics was always going to be a key challenge for this research. Our strategy was to only use publically available data and automate the collection where possible.

We identified the following data sources to be used to obtain data points and calculate the identified metrics:

- Data obtained directly from the website or inspection of the website;
- Google – Historic page views, Ad Planner data and Brand Rank;
- Kantar Media Compete – Website referral information;
- Alexa – Reputation Score;
- Robtex/DNS/ WHOIS lookup – IP address and Website data;
- Team Cyrmu Community Services – ASN and Country codes;
- IANA – data on the top level domain, for example .com, .uk, or .tv.

Once the data sources were identified we analysed the data points available and our ability to calculate the metrics using them. This resulted in the identification of four groups of metrics:

- **Simple metrics:** Metrics where the data points are available and therefore the metric can be calculated simply.
- **Proxy metrics:** Metrics where data is not available, but where we use other data points as a good proxy for the metric.
- **Excluded metrics:** Metrics that we could not calculate with the available data and therefore had to be excluded from the segmentation.
- **Complex metrics:** Metrics that require a number of data points to allow us to calculate them.

Through this analysis we are confident we obtained a set of metrics that could be used for our segmentation.

A full list the data used and the details of each of the metric calculations can be found in Appendices G and H.

#### 4.2.3 Obtaining the data and calculated the metrics

We completed the data collection and metric calculation for the websites in five stages:

1. User journey URL and search URL capture
2. Automated data capture
3. Manual data capture
4. Third party data capture
5. Metric calculations completed

For full details of each metric, the data points contained within it and the details of each stage of the data capture process please consult Appendices G, H and I.

### 4.3 Identifying six segments in the data

As we have seen in the previous chapters, the 102 data points were collected for 257 websites. The data collected varied in nature and consistency. The chosen method of segmentation needed to be able to manage numerical (e.g. Revenue=2401), categorical (e.g. Country=SE) and missing data.

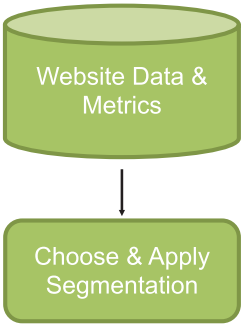



*Figure 4-6: Collating the metrics, choosing and applying the segmentation*

*4 Alexa, (2012) description, [Online], Available: http://www.alexa.com/help/traffic-learn-more [18 May 2012]*

The purpose of this report was to find structure and patterns in the websites considered to be infringing copyright, without recourse to experience, which led us towards a data and algorithmic approach, based on the pros and cons outlined below:

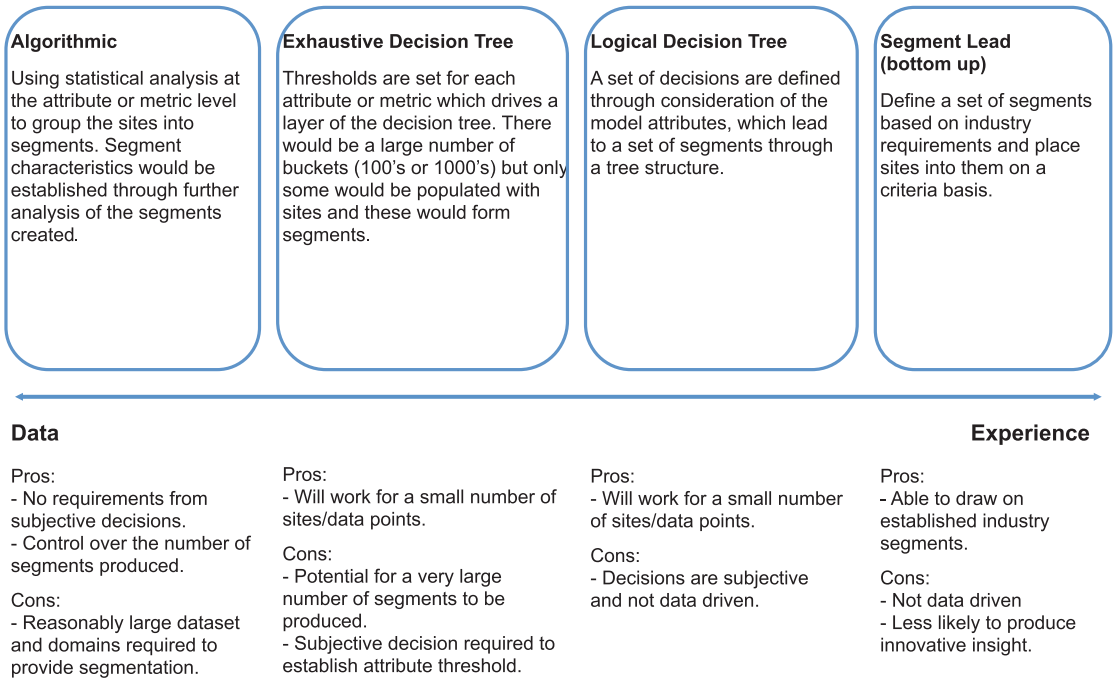



*Figure 4-7: Different segmentation methods that were considered for this report*

### 4.3.1 An algorithmic approach to the segmentation

There are many algorithmic approaches that could be used. We selected the 'Random Forests' algorithm to create a ranking of similarity for various solutions, along with the use of the 'Within Sum of Squares' technique to identify the number of segments. A detailed discussion on how and why we selected this approach can be found in Appendix J.

### 4.3.2 Six segments as the natural solution

From the 257 websites, we used 153 websites as the 'Training' set and the remaining 104 websites as the 'Validation' set. We used the training set of websites to test the optimum number of segments needed to classify the market. After we had determined the similarity matrix for the training set, we started by assuming three segments, or clusters, and used the 'With Sum Squares' to calculate an understanding of 'cluster suitability'. We then repeated this calculation assuming four segments, then five, then six, and so on until fifteen. We looked for a significant change in the 'cluster suitability' which would highlight to us a segmentation that was distinct from other segmentations and therefore likely to be the most representative of the market. We found this significant change when moving from five segments to six, shown in Figure 4-8.

We then reviewed the segmentation into six clusters with the use of 'dendrogram' plots and by validating with the second set of data. These techniques are discussed in more detail in Appendix J.

We analysed the second validation set of 104 sites to test the indicated segmentation solution of six segments by assigning these to the existing segments.

We did this by rerunning the random forests algorithm using all 257 sites. We then used the resulting similarity matrix to match each validation site to the training sites that it was most similar to. We then assigned the validation site to the same cluster as its match. This had the advantage that the validation site could be assigned the entire hierarchy of its most similar training site, so a complete hierarchical clustering of all the new sites was achieved. We found the result shown in Figure 4-9 when comparing the Within Sum of Squares against the number of clusters for the 104 validation web sites.

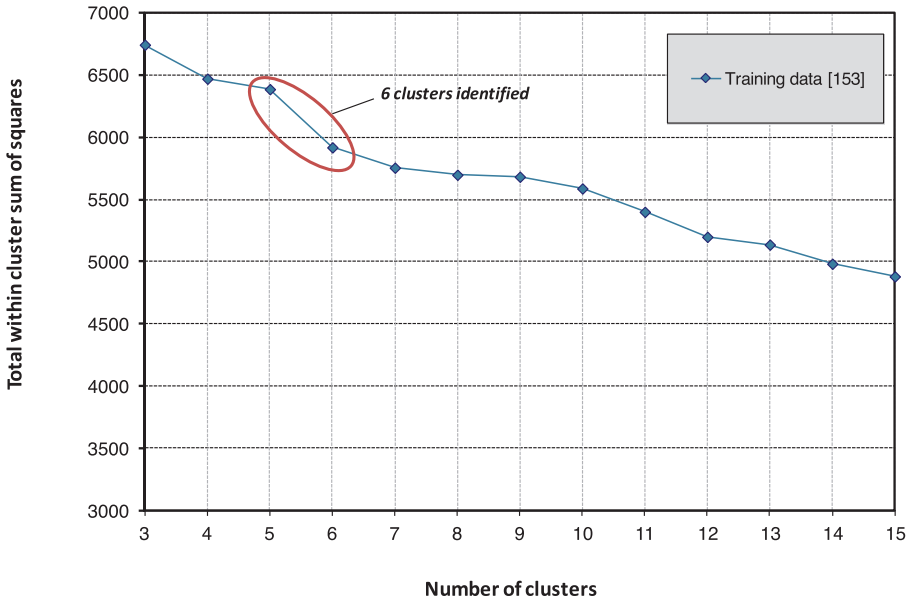



*Figure 4-8: The 'Within Sum Squares' plot on the training data shows a significant decrease between five and six which suggests a six segment presentation of the market*

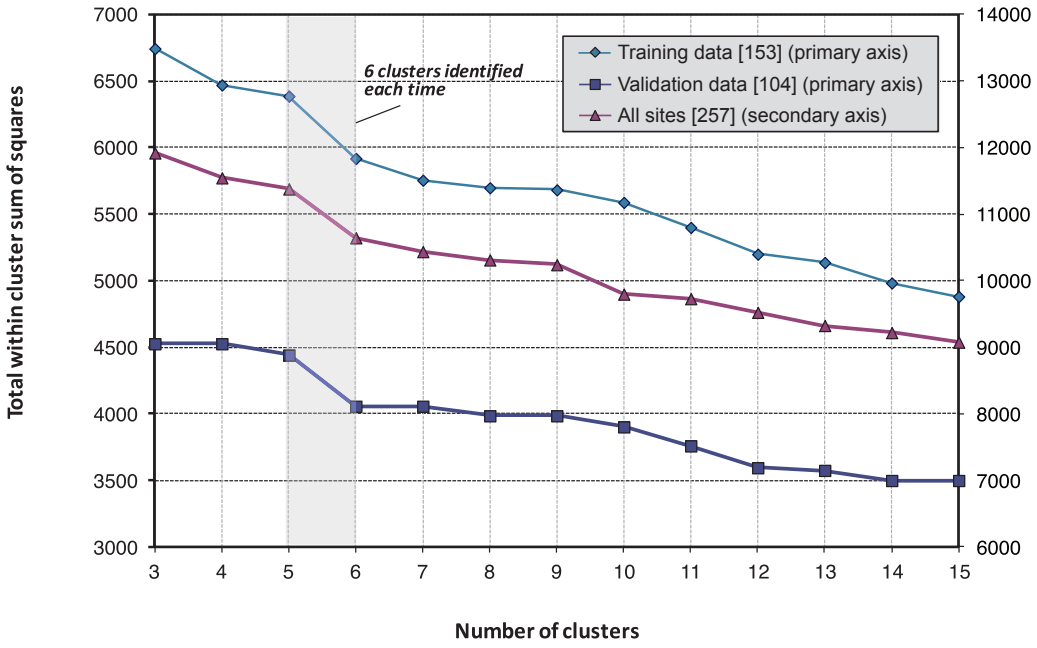



*Figure 4-9: The 'Within Sum Squares' plot on the validation data and repeated for all the data confirms the initial six segment representation*

When the total Within Sum of Squares per cluster for the combined 257 training and validation sites was then computed for 3 up to 15 clusters, the result obtained also showed a sharp drop when going from 5 clusters to 6. This confirmed our initial finding of six segments.

During the clustering, we observed that some of the segments were better defined than others. We confirmed this by undertaking Principal Component Analysis (PCA) on the 153 training data websites; see Appendix K for more details. The PCA provided a view on how close or different the segments were; the 1st two principal components give a rough idea of what is happening within the data. It should be noted that this is a simplification, and the features distinguishing the other segments are probably being expressed in higher dimensions.

# 5 Next steps

Detica recommends that this study is repeated as it is likely to prove valuable for stakeholders and policy makers for two reasons:

1. It will identify trends in a rapidly changing and dynamic market; and
2. It will allow for the impacts of industry changes and mitigating policy actions to be monitored and evaluated.

The benefits of this will be to increase the probability of policy achieving its objectives and to reduce the risk of unintended consequences and unforced errors.

## 5.1 Repeating the study to understand changes to the market conditions over time

Chapters 2 and 3 highlight that, once the market is segmented, various conclusions can be drawn about the dominant behaviour of each of the segments. However, as a result of use of data analytics to define the segments in the first place, no conclusions can be drawn on the inter-dependencies between the segments. These inter-dependencies might include movement of consumers from one segment to another, shifts in the technology-use between segments or shifts in advertising spend, for example.

These inter-dependencies between market segments and changes to market conditions are conclusions that can be drawn over time. In order to understand how the six segments change over time, we recommend this study is repeated at intervals in order to assess the changes from the previous study. This could also provide the basis for any impact assessments that may be required before undertaking market changing actions.

## 5.2 Repeating the study to analyse the cause and effect of events

In the same vein as the section above, this study is unable to report on the effects of the implementation of certain actions undertaken or events that happen in the market for material that infringes copyright.

In order to understand the impact on the six segments found, after the implementation of an action or market changing events, we recommend this study is repeated in a timely manner to assess the actual impact.

## 5.3 Industrialising the study for a wider dataset

In future this report may be provided on a regular basis. This may need the number of websites sampled to be enlarged and the processes by which the datasets are collected to be undertaken with greater automation. We discuss possible enhancements to this study, below, if it were to be undertaken again.

### 5.3.1 Industrialisation of Data Capture

The findings present in this report are based on data collected from 153 websites. A further 104 websites were used to independently validate the presented segmentation result. Whilst we attempted to automate as much of the data capture as possible, via scripted website data collection, a significant number of metrics required either manual collection or verification.

A second key output of this research has been the development of a metric-based segmentation model; looking forward, this could be applied to future studies. However, it is likely that additional research in this space will seek to focus on increasingly large numbers of websites. This presents a number of challenges given the manual effort undertaken during the course of this study to capture the necessary data points. In this section we discuss potential ways to fully automate or industrialise the data capture process for website specific metrics.

### 5.3.2 Automated website data collection

For this study a number of simple Python scripts were written with the intention of automatically collecting model metrics. This was accomplished by fetching Web pages over HTTP using the appropriate python libraries (for example 'urllib'). For each site a number of HTML pages were required, corresponding to user's journey on the site. Parsing each of our collected website's publicly available HTML pages we were then able to search for specific content including keywords and links to other websites relating to specific model metrics.

The above approach contained a number of limitations:

- We required prior knowledge of the website specific user journey necessary to consume content. Thus in all cases, this information had to be manually collected and used as an input.
- We observed that complicated websites, e.g. those with a significant amount JavaScript, were not always reliably collected – requiring manual verification.

### 5.3.3 Alternative methods

A number of alternative approaches to data collection were considered however given the time constraints of this project these were not employed. These are presented below, and may prove more suitable for future studies.

- *Site specific website data collection*

It may be more robust to write scripts that are unique to each website. Whilst this may be more robust in term of collecting metrics, it is certain to prove time consuming to develop

- *Browser automation*

To ensure that any future capture of website data is consistent, a browser automation tool could be employed to collect the required HTML pages on a user journey for a particular site.

For each of the cases listed above, it should be noted that as individual website change over time, a significant amount of effort may be required to ensure that the necessary metrics can still be collected.

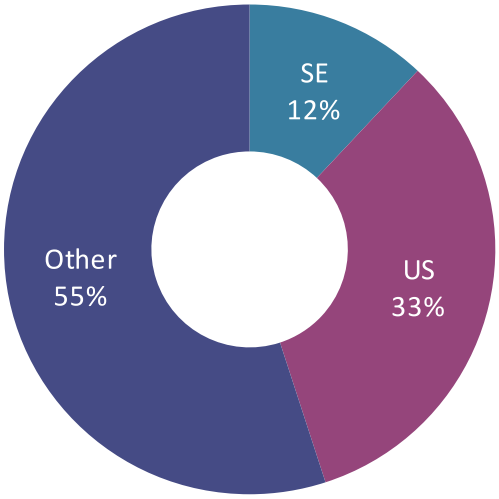
# Appendix Contents

## A. Categorical metric detailed results

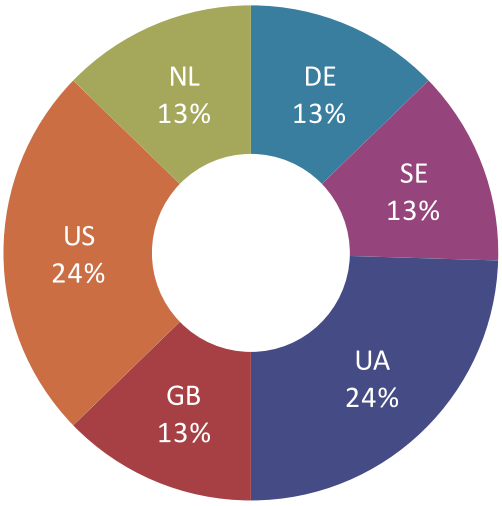
### IP Address Location

This page represents the value of the 'A record' metric (IP address location) for comparison across the six segments.

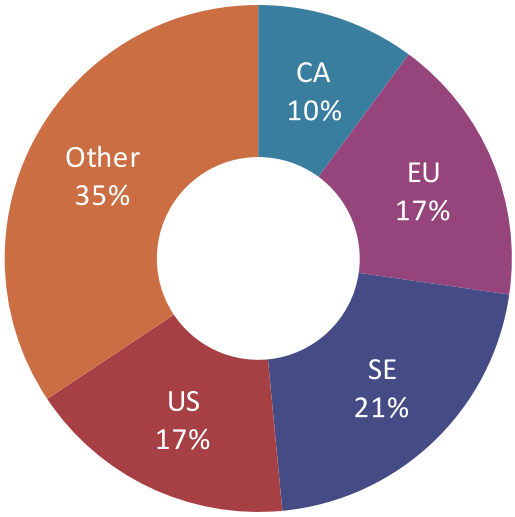
1



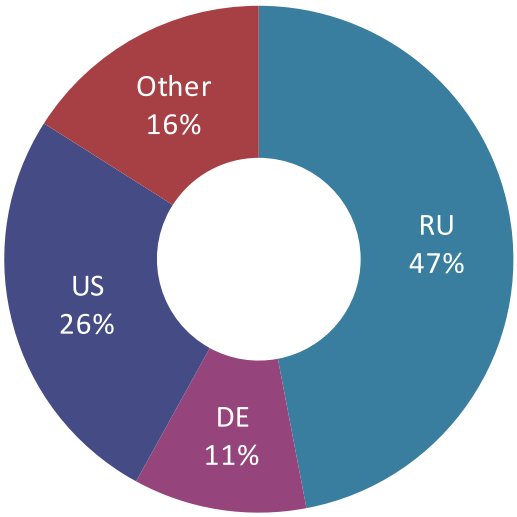

2



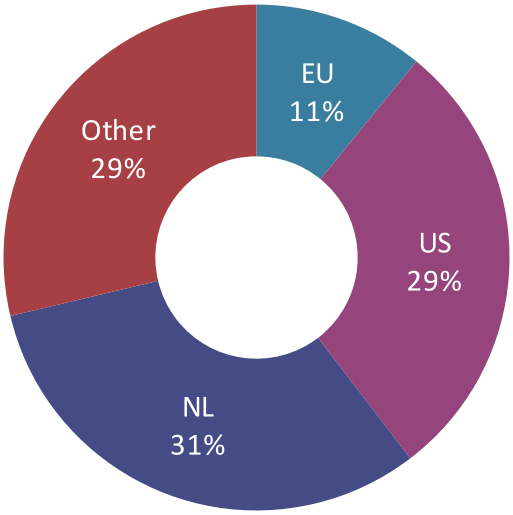

3



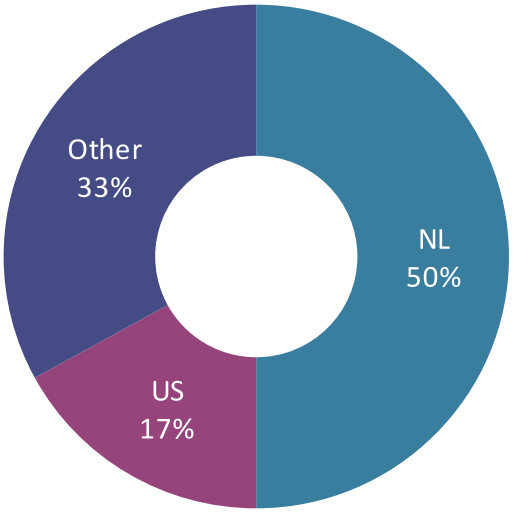

4




5



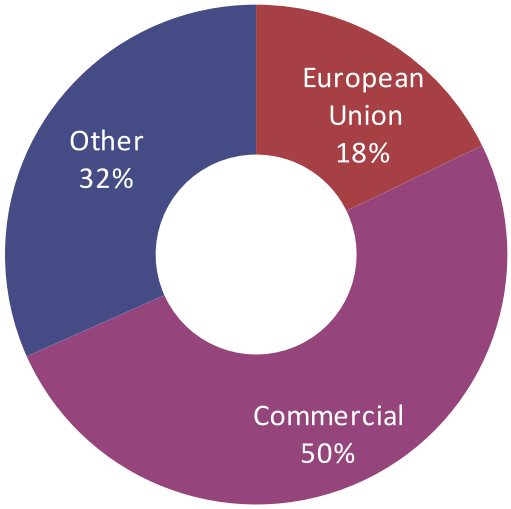

6



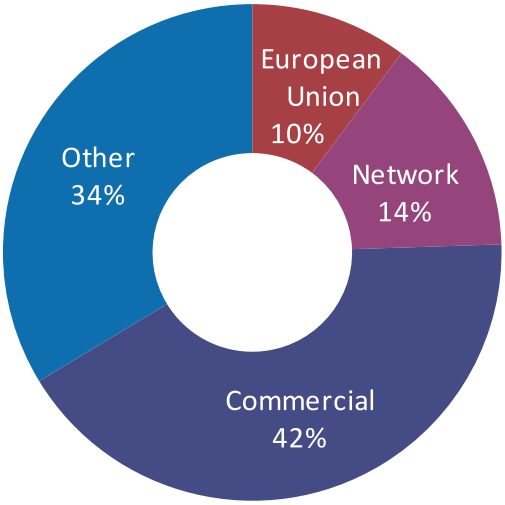
## Top Level Domain Location

This page represents the value of the Top Level Domain Location metric for comparison across the six segments.

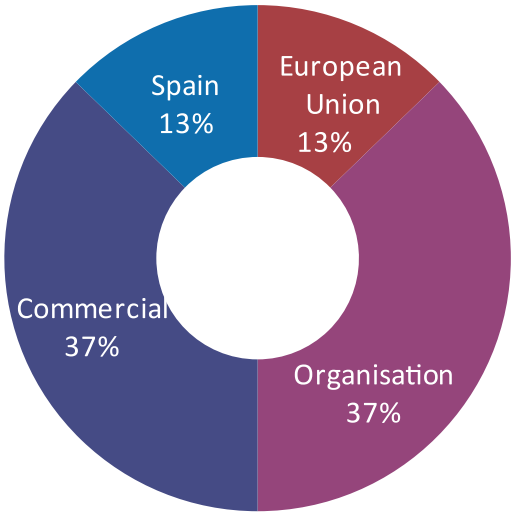
1



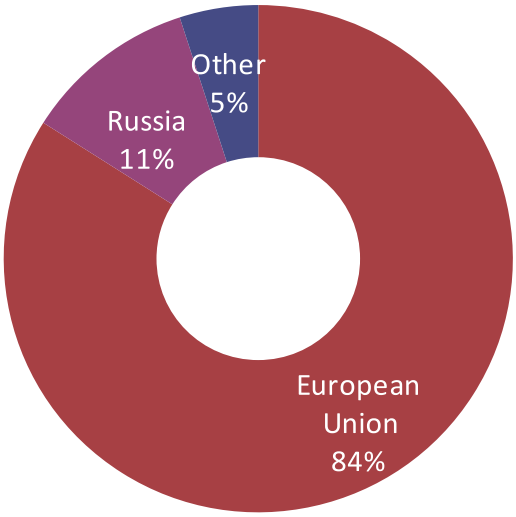

2



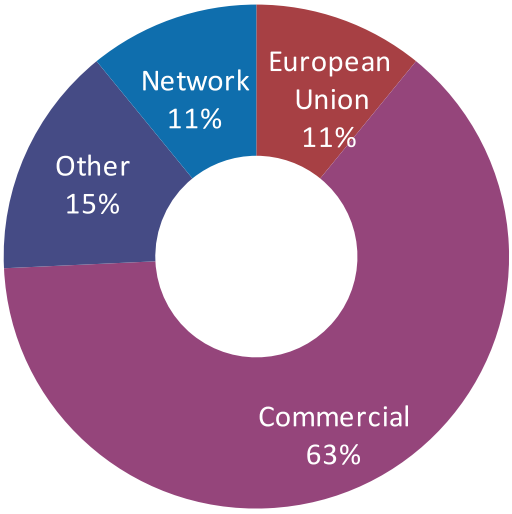

3



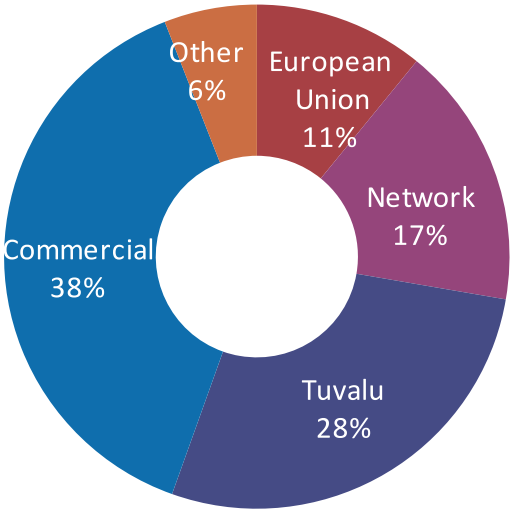

4




5



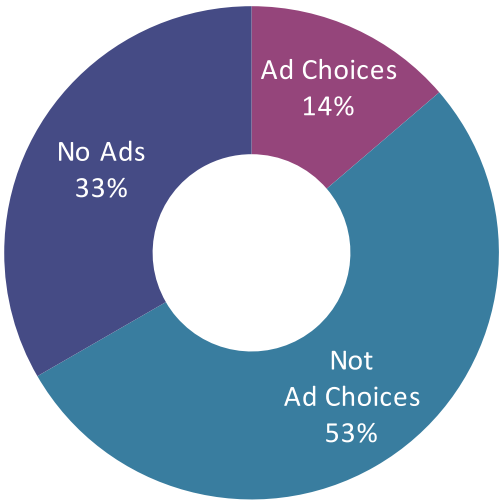

6



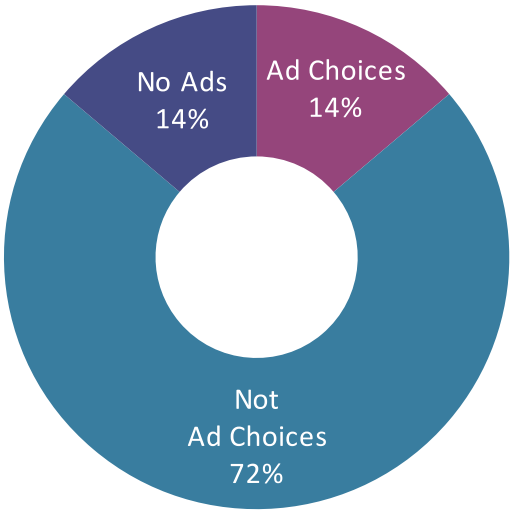
## Ad Provider Type

This page represents the value of the Ad Provider Type metric for comparison across the six segments.

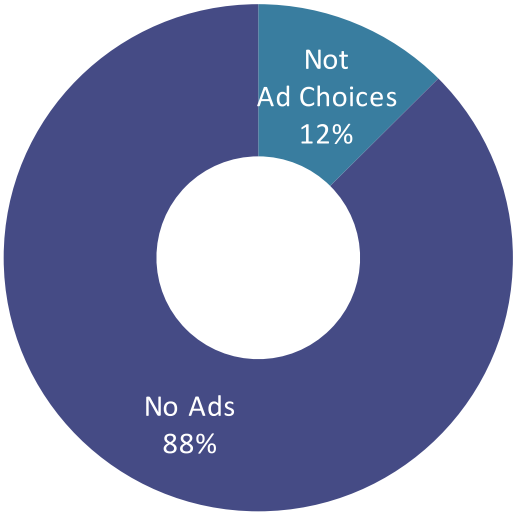
1



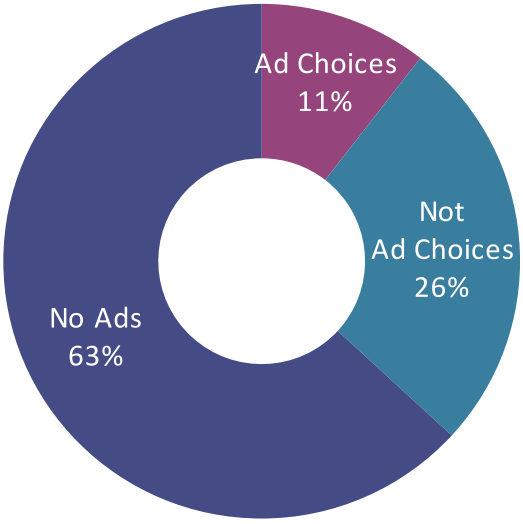

2



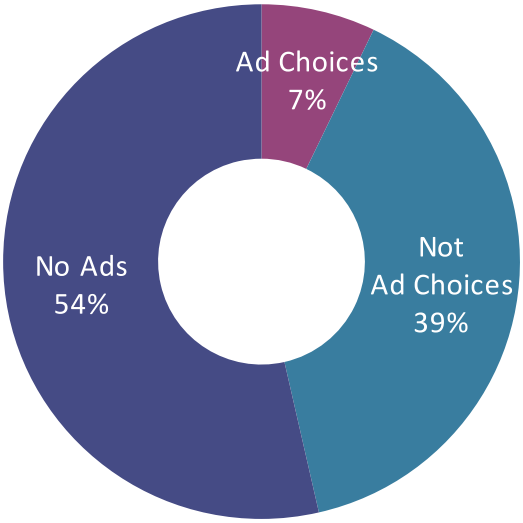

3



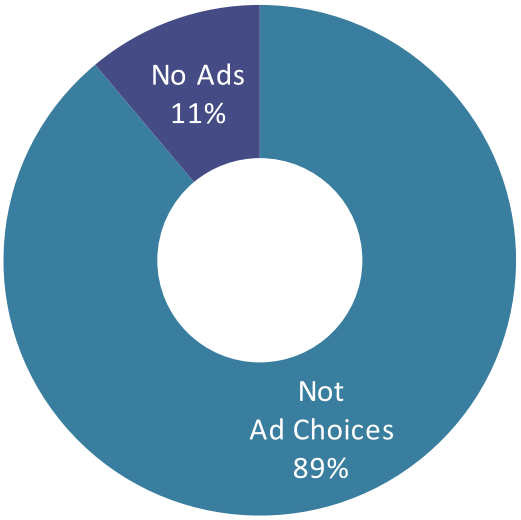

4




5




6

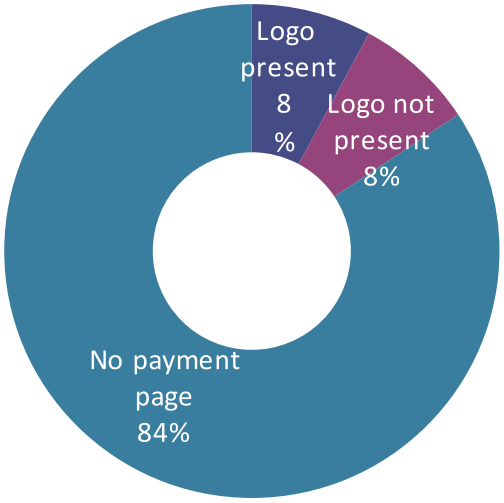



N.b. Presence of an Ad Choices logo on a website was used to determine if the ad provider is a member of the Ad Choices scheme.

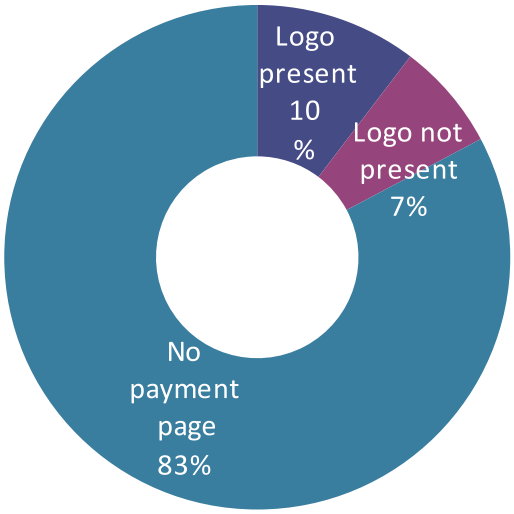
## Card Processor Logo

This page represents the value of the Card Processor Logo metric for comparison across the six segments.

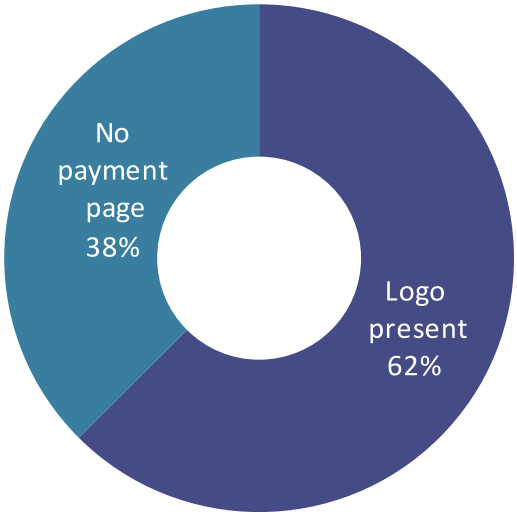
1



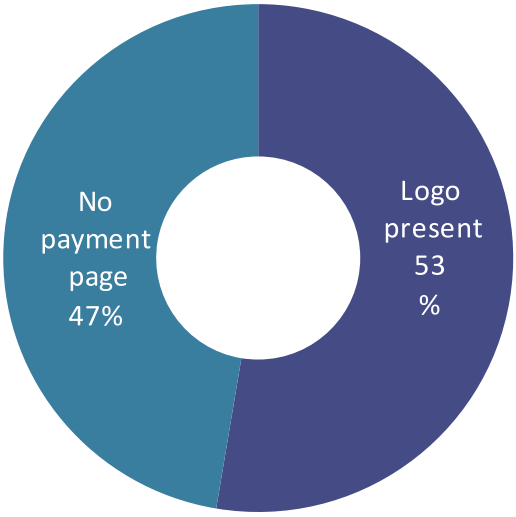

2



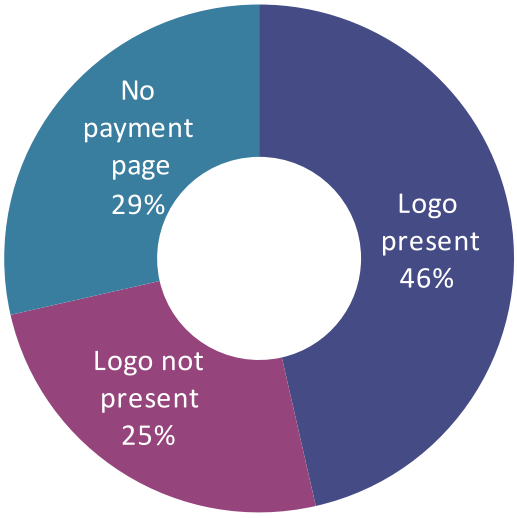

3



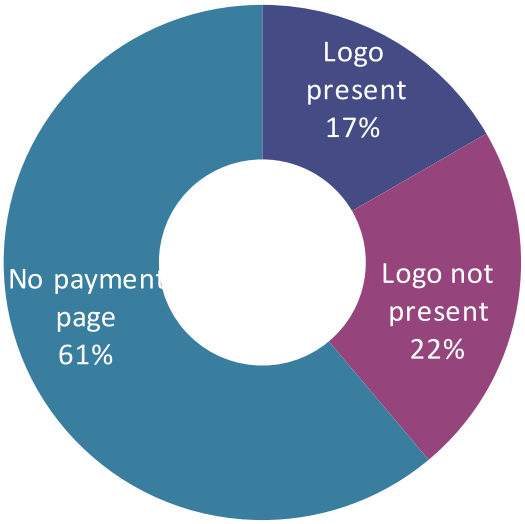

4




5



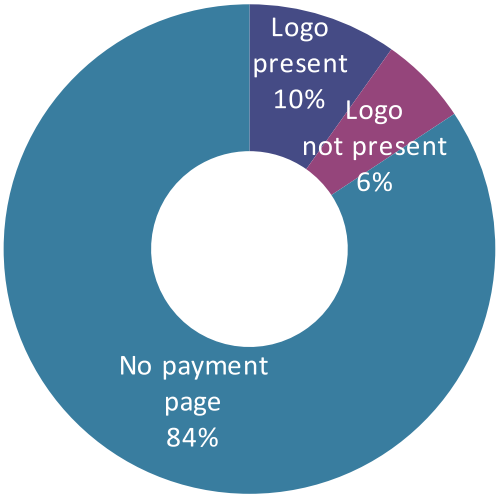

6



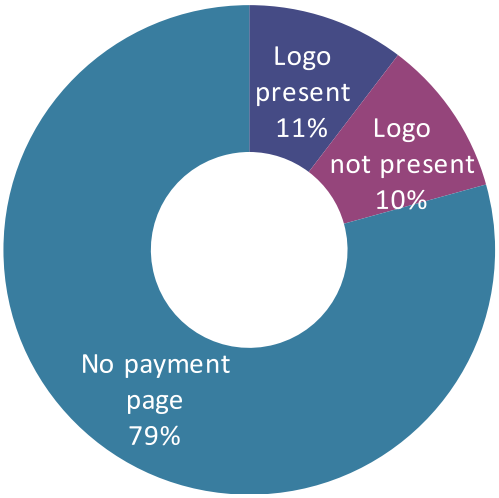
## Electronic Payment Provider Logo

This page represents the value of Electronic Payment Provider Logo metric for comparison across the six segments.

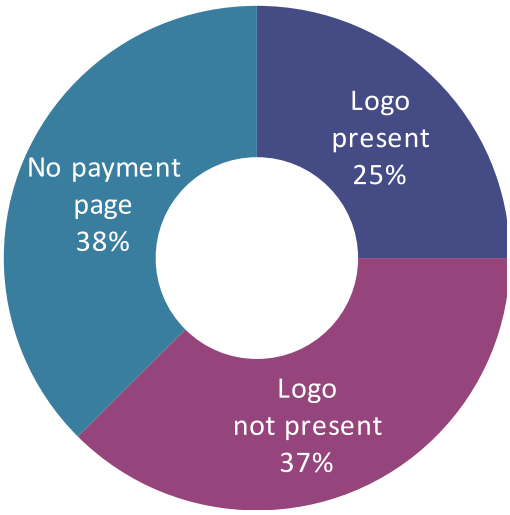
1



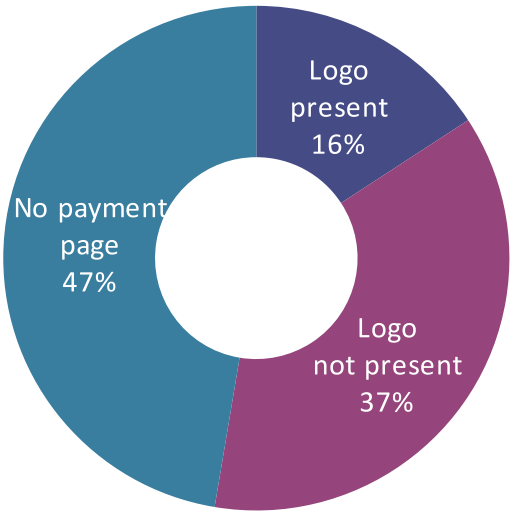

2



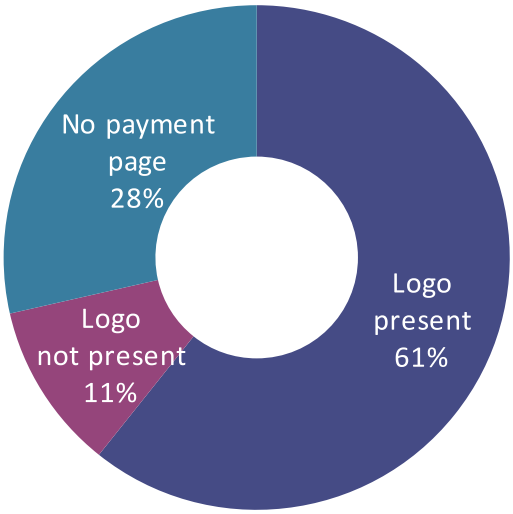

3



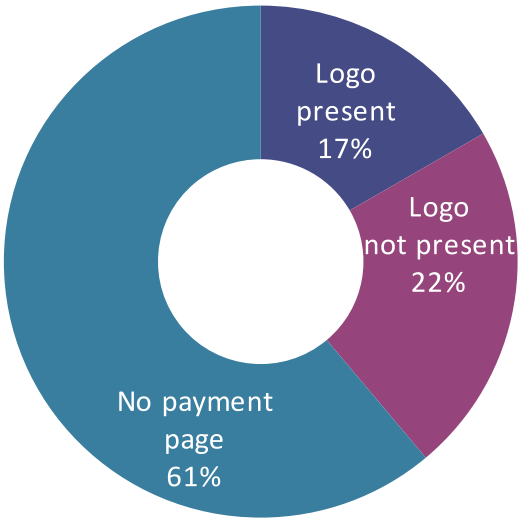

4



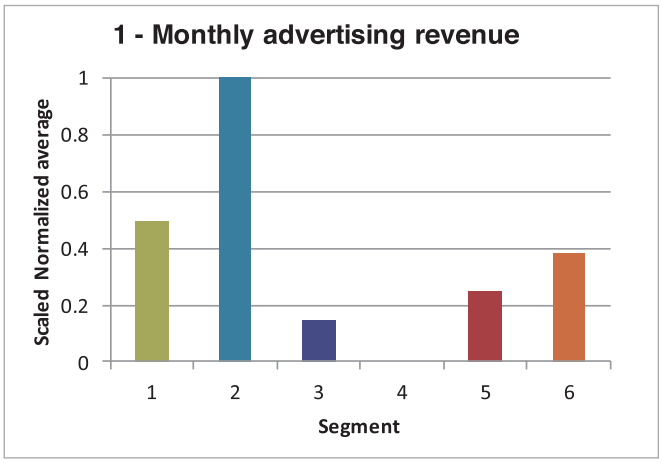

5

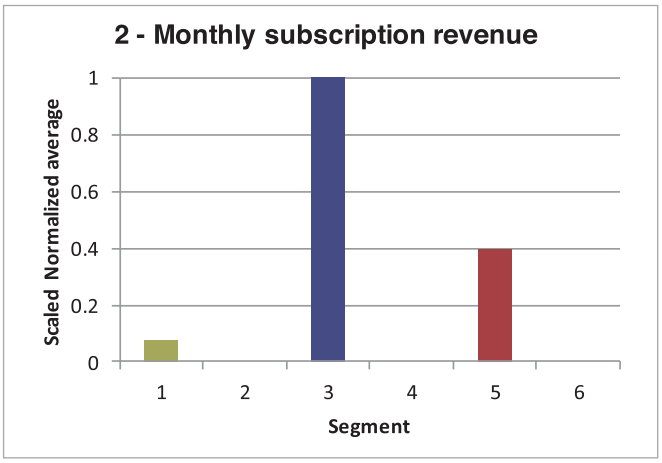


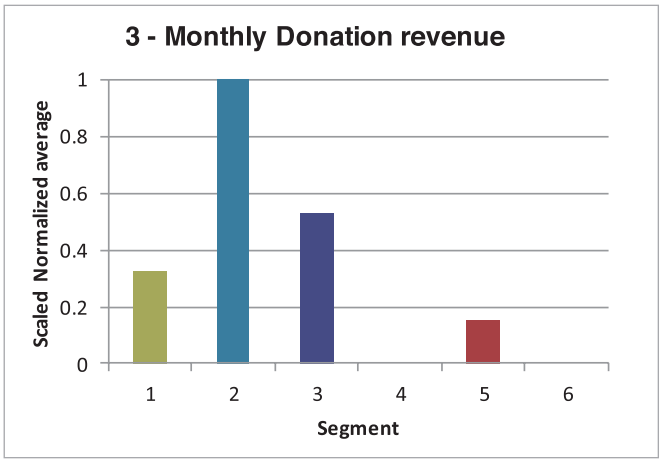

6

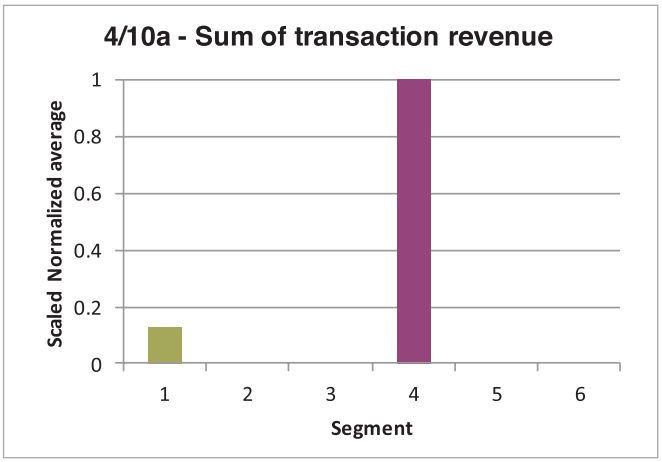


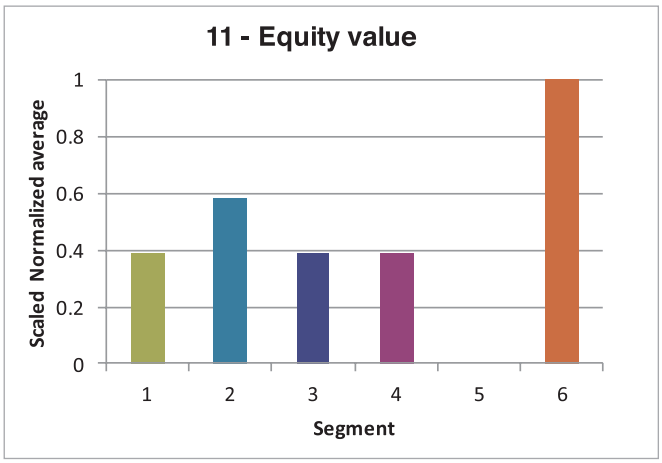
## B. Numeric metric detailed results

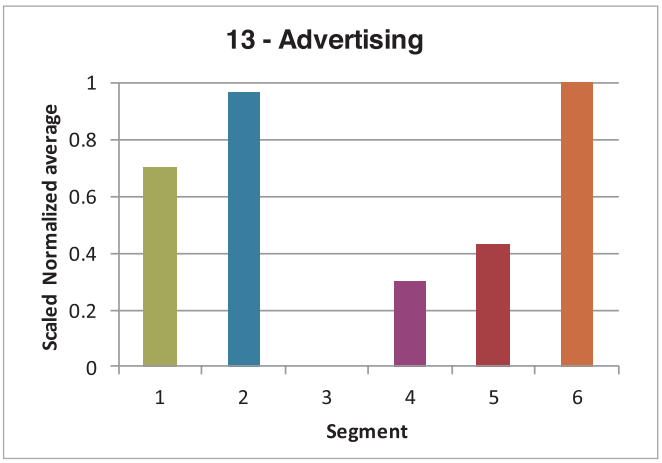


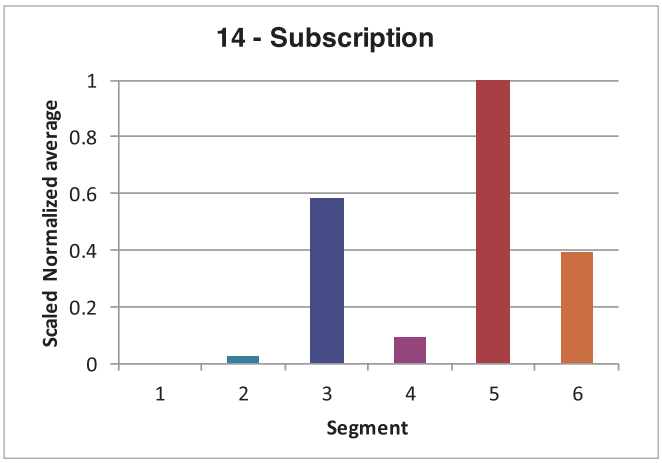



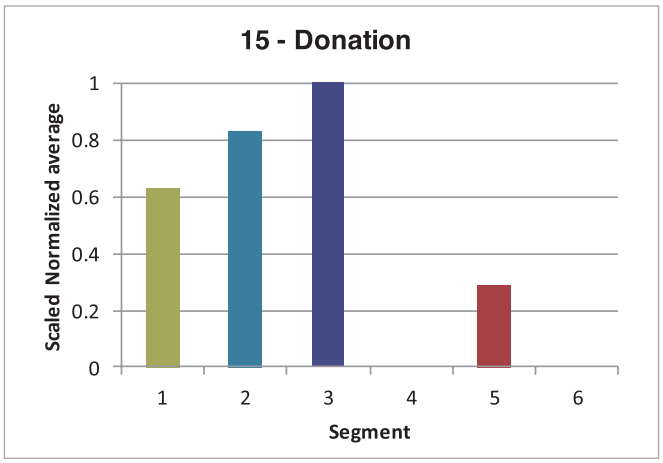



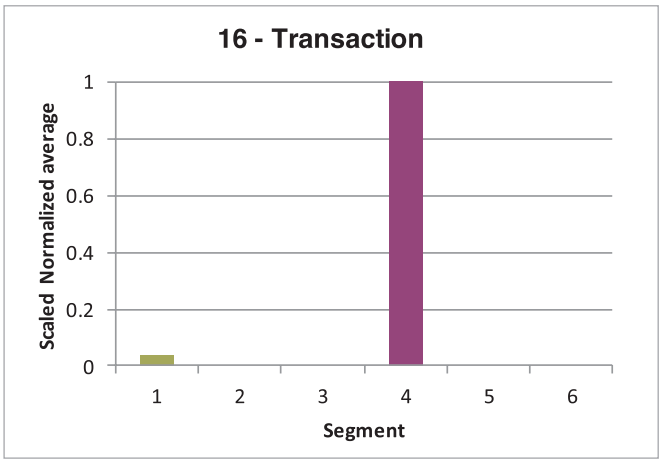



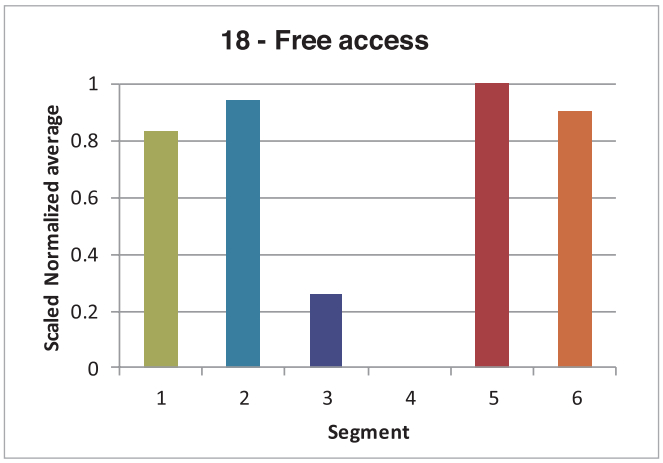



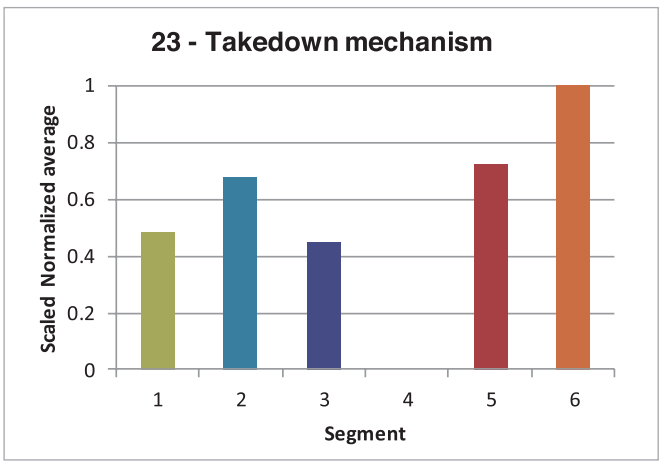



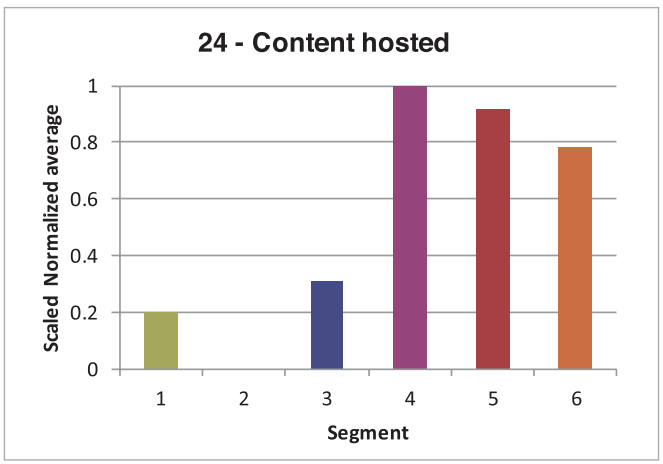

14 - Subscription
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

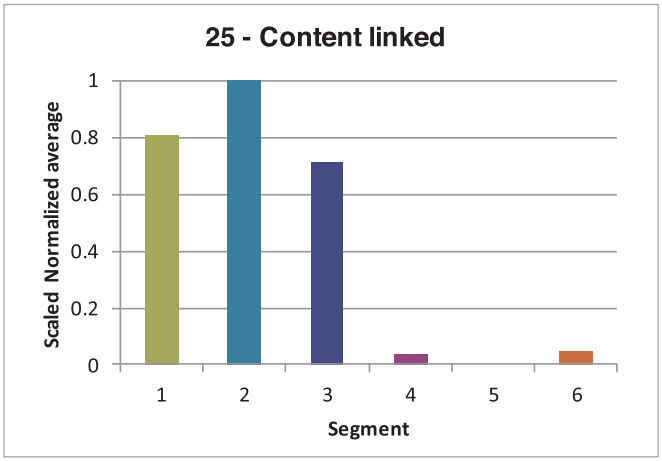

15 - Donation
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

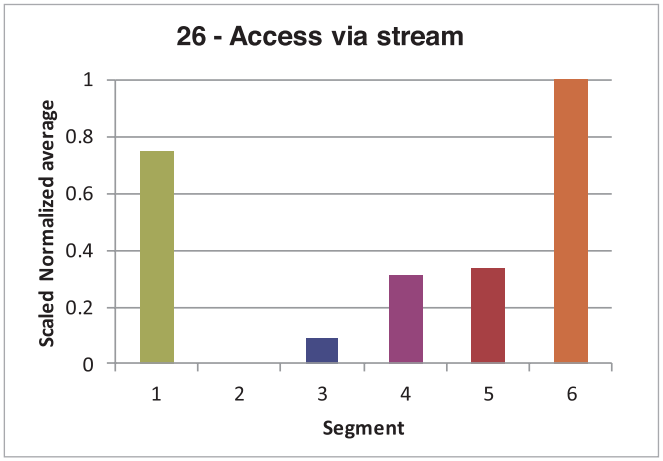

16 - Transaction
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

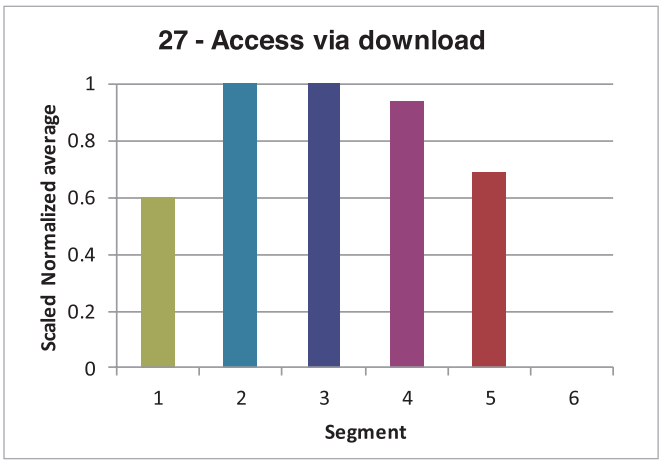

18 - Free access
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

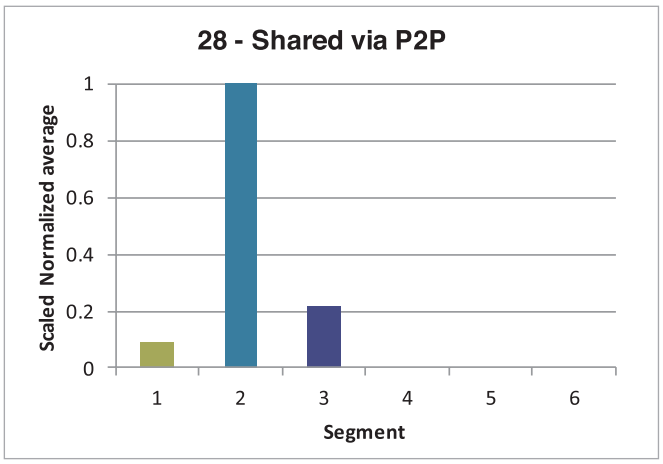

23 - Takedown mechanism
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

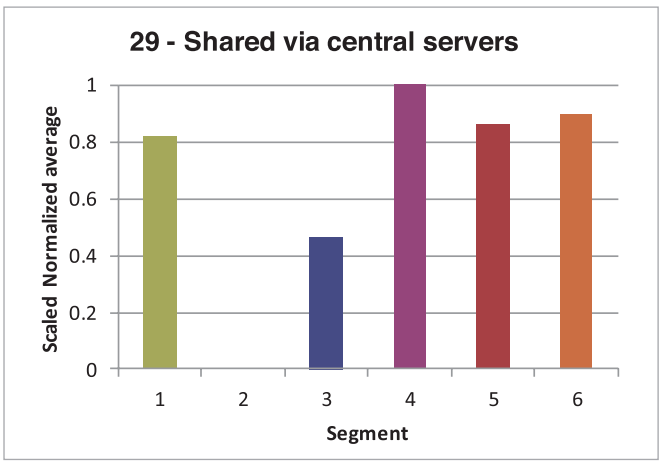

24 - Content hosted
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

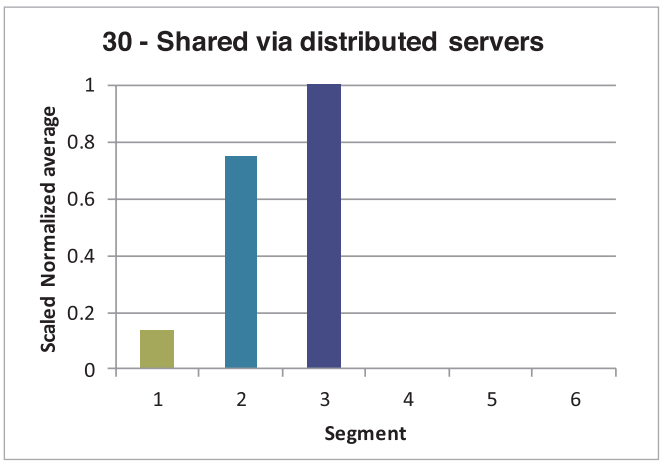

25 - Content linked
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

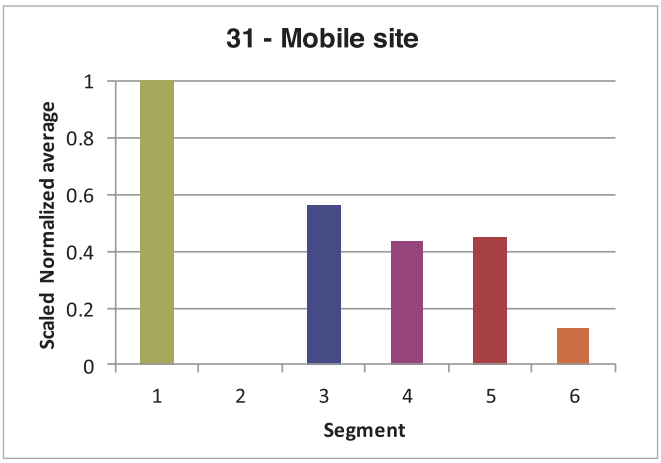

26 - Access via stream
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

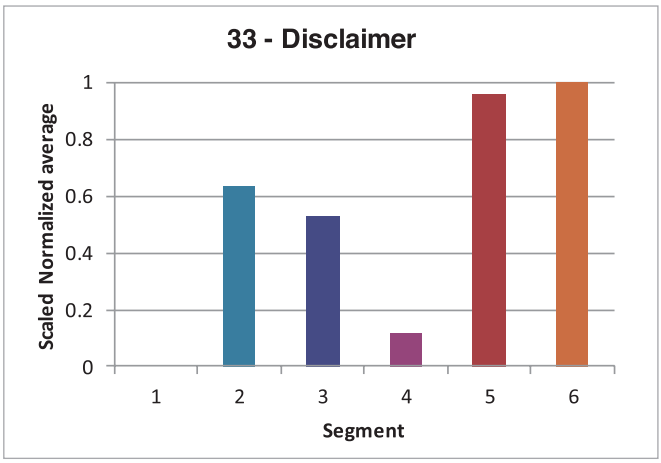

27 - Access via download
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

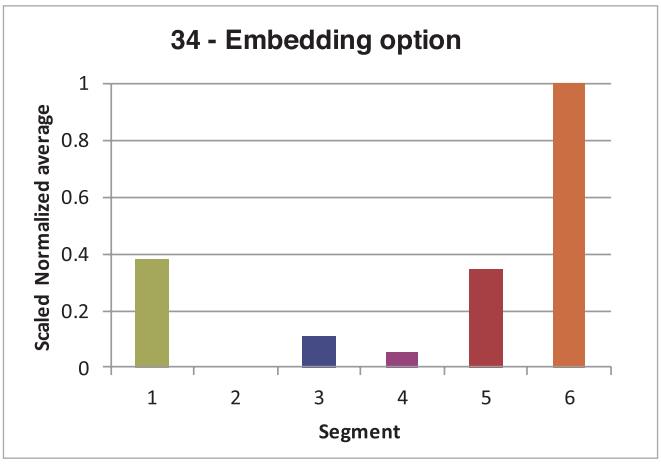

28 - Shared via P2P
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

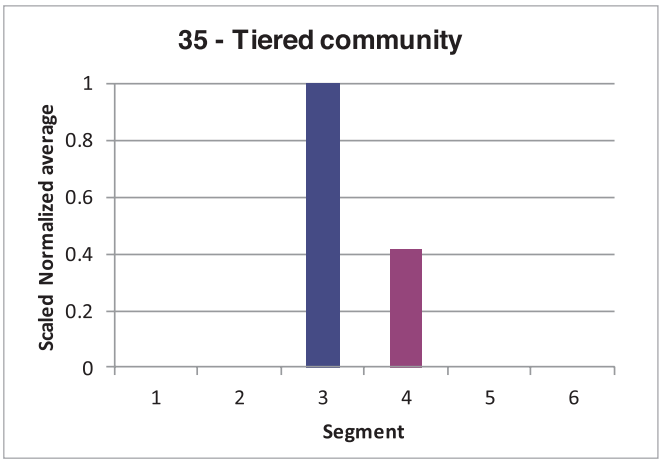

29 - Shared via central servers
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

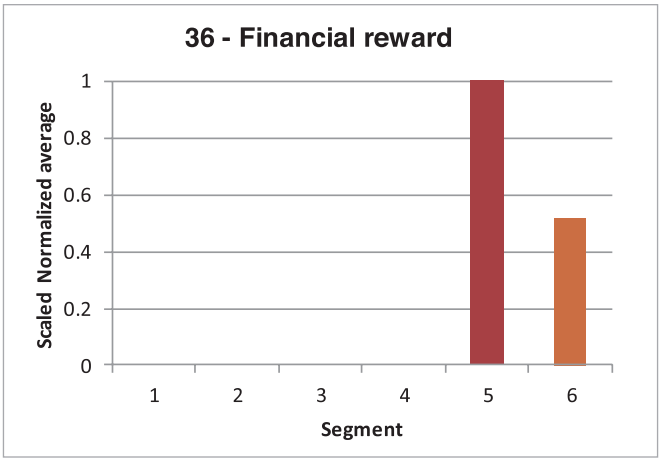

30 - Shared via distributed servers
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

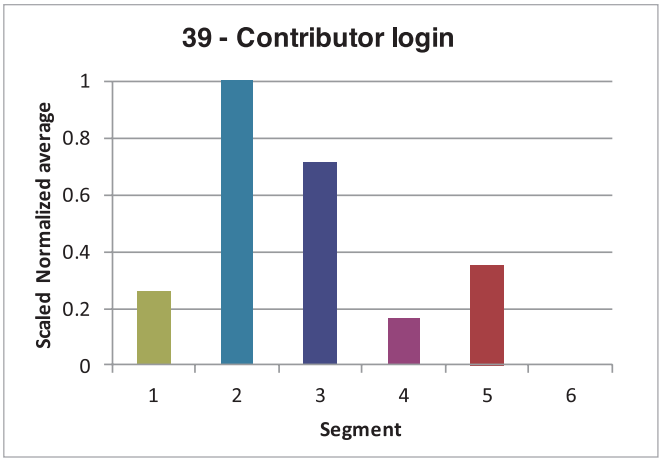

31 - Mobile site
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

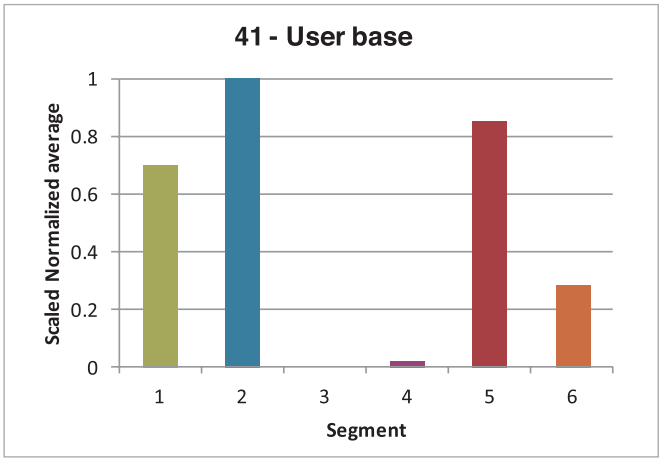

33 - Disclaimer
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

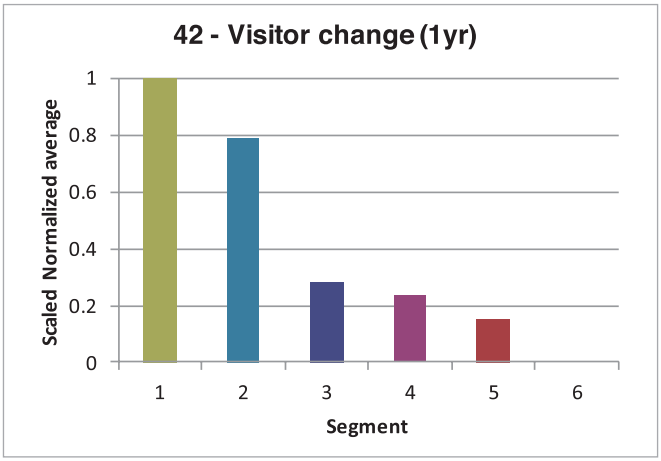

34 - Embedding option
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

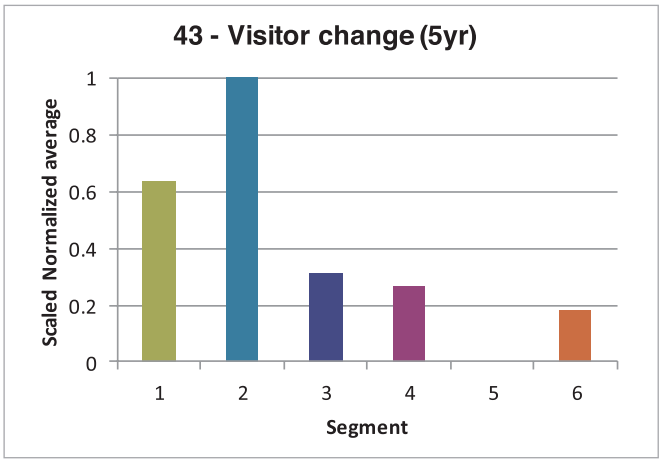

35 - Tiered community
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

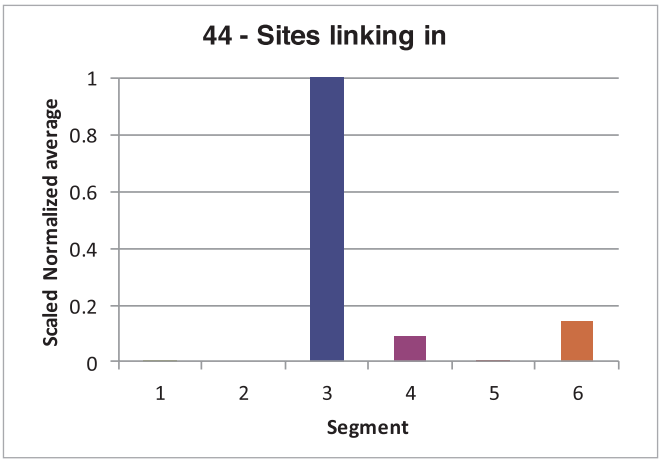

36 - Financial reward
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

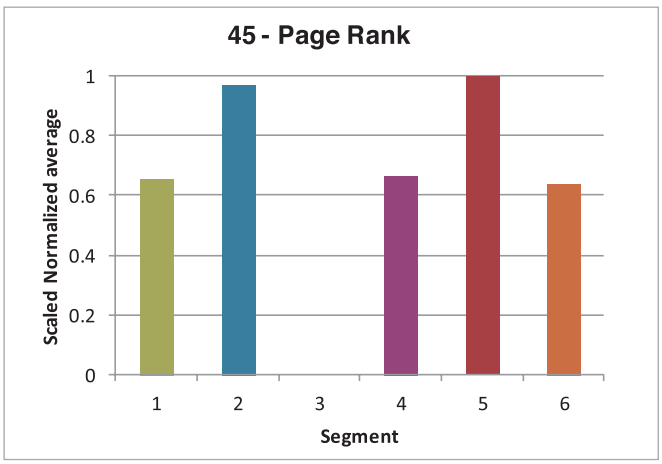

39 - Contributor login
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

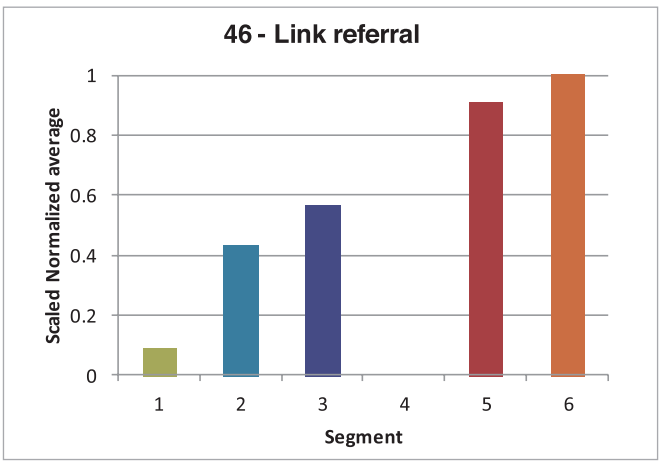

41 - User base
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

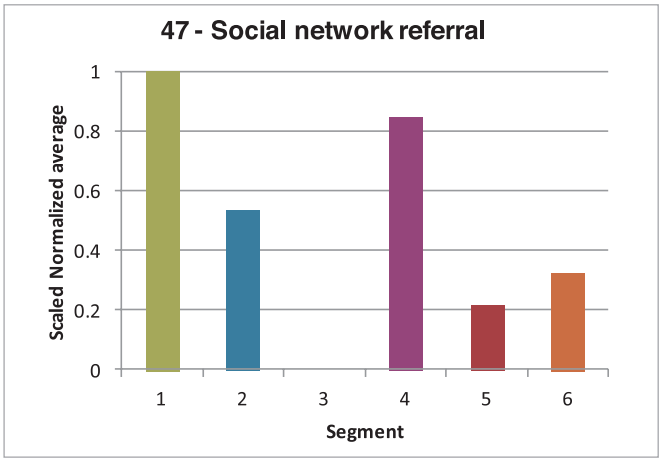

42 - Visitor change (1yr)
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

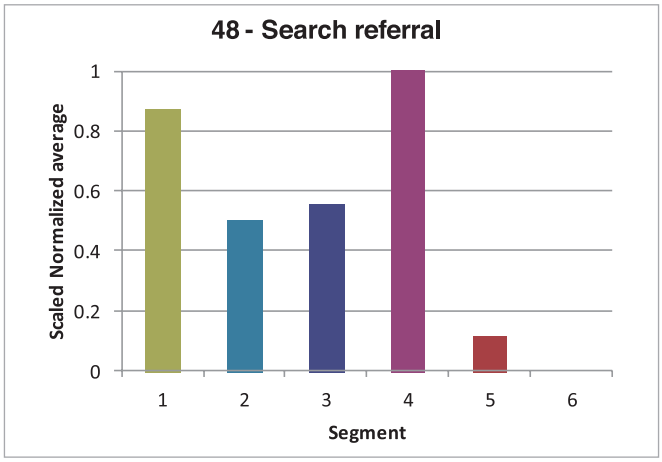

43 - Visitor change (5yr)
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

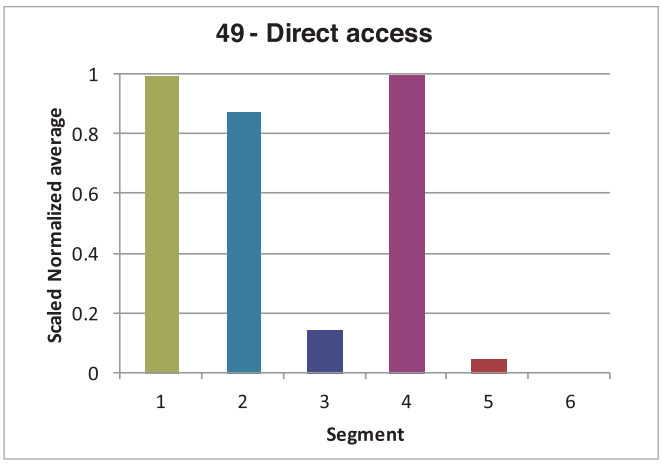

44 - Sites linking in
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

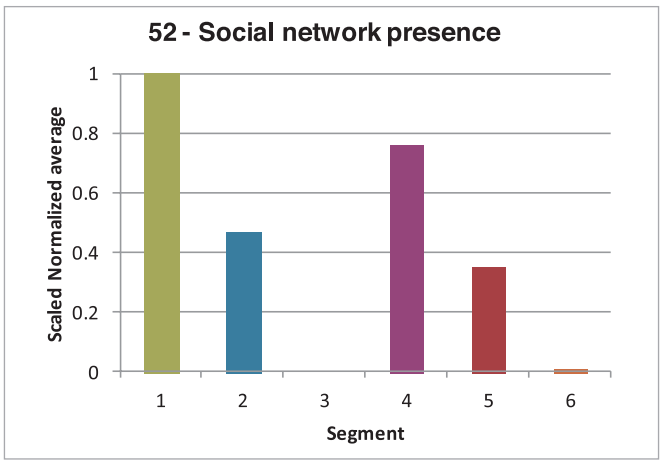

45 - Page Rank
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

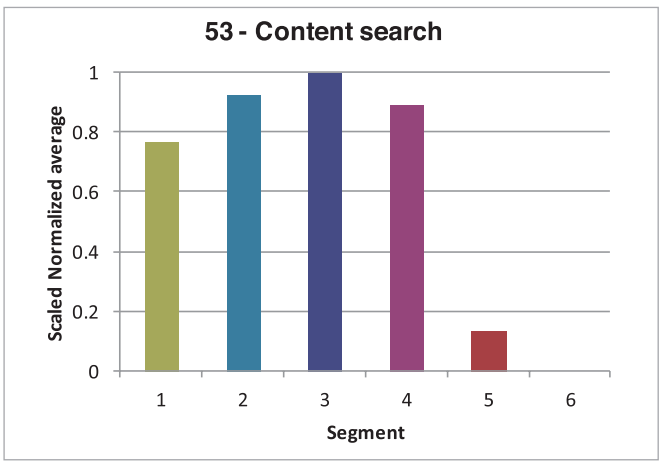

46 - Link referral
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

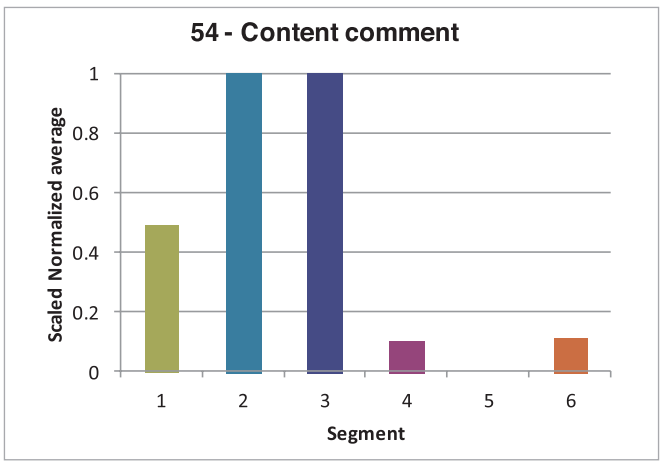

47 - Social network referral
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

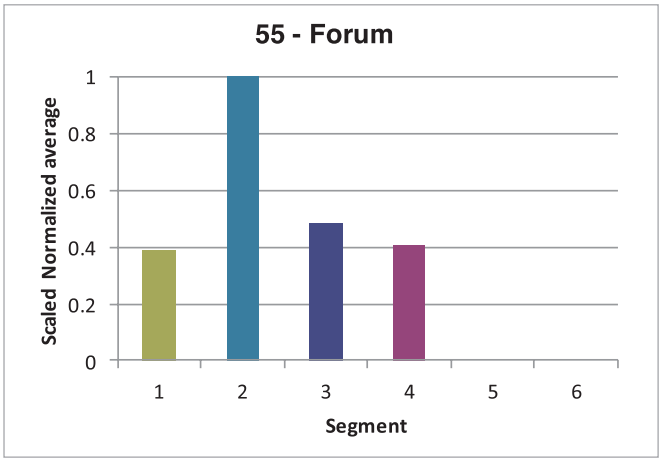

48 - Search referral
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

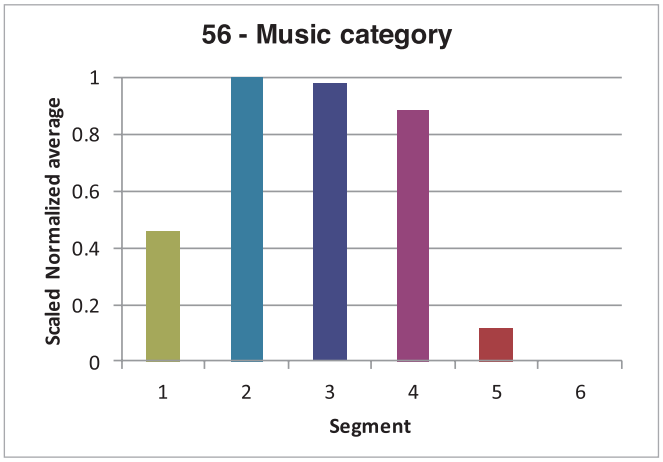

49 - Direct access
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

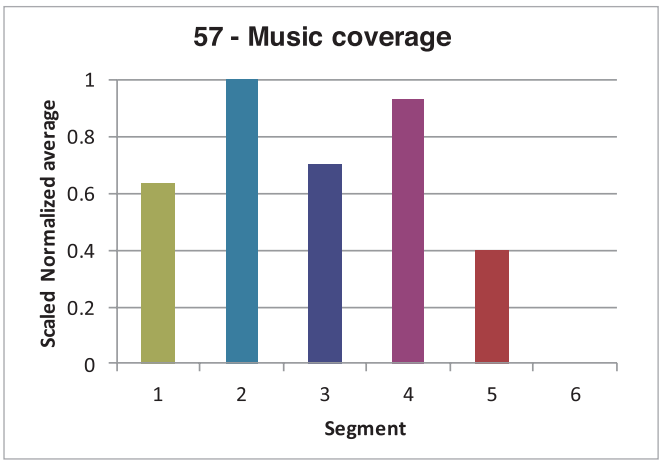

52 - Social network presence
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

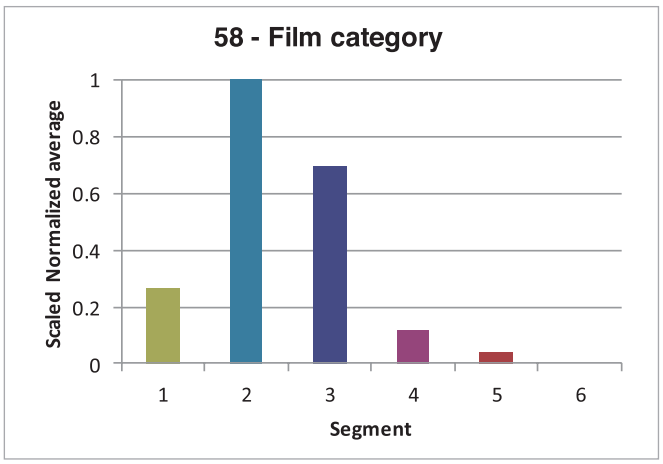

53 - Content search
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

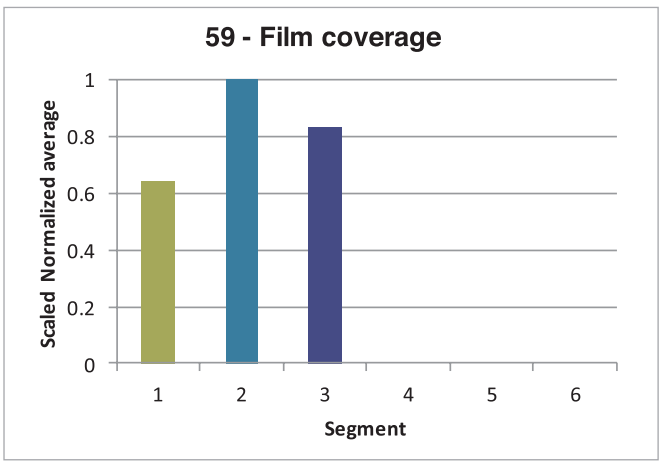

54 - Content comment
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

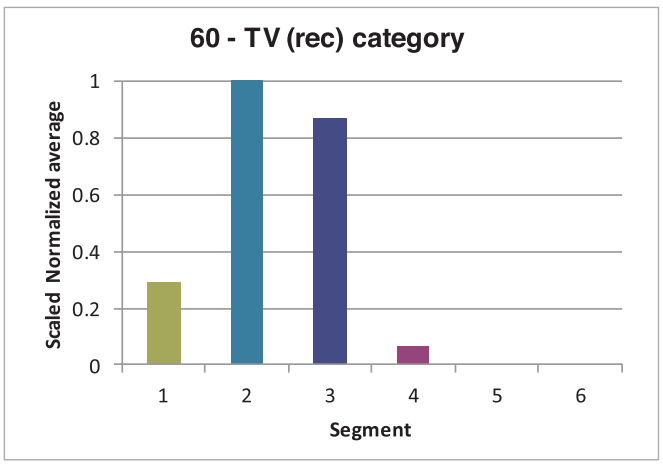

55 - Forum
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

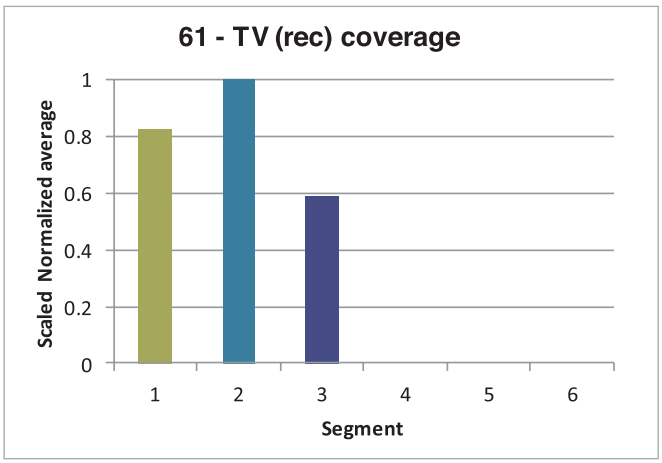

56 - Music category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

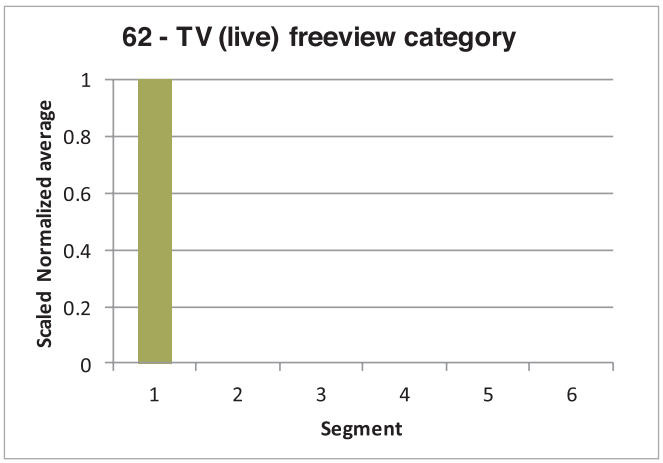

57 - Music coverage
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

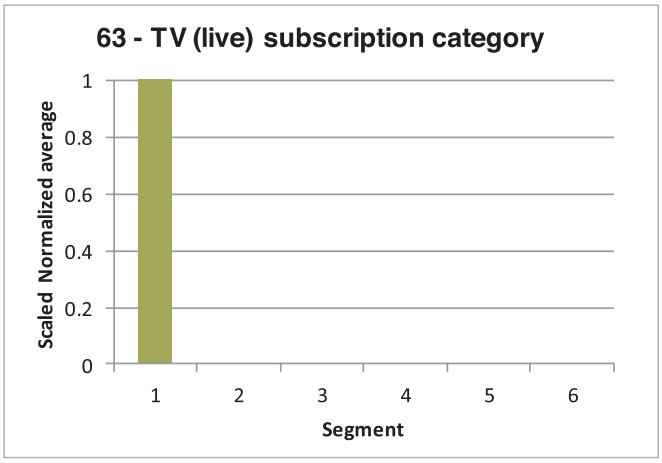

58 - Film category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

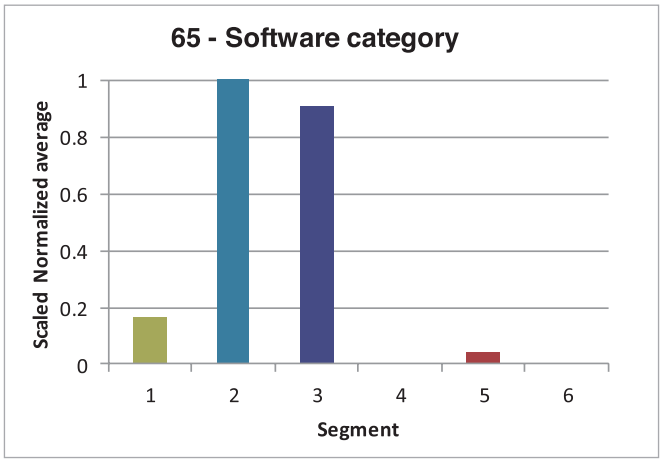

59 - Film coverage
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

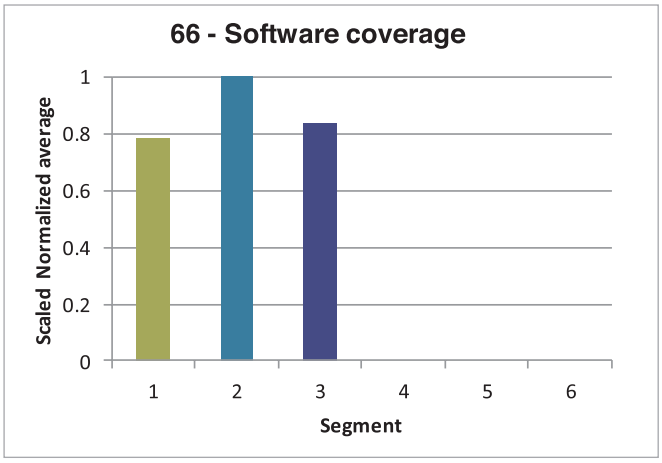

60 - TV (rec) category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

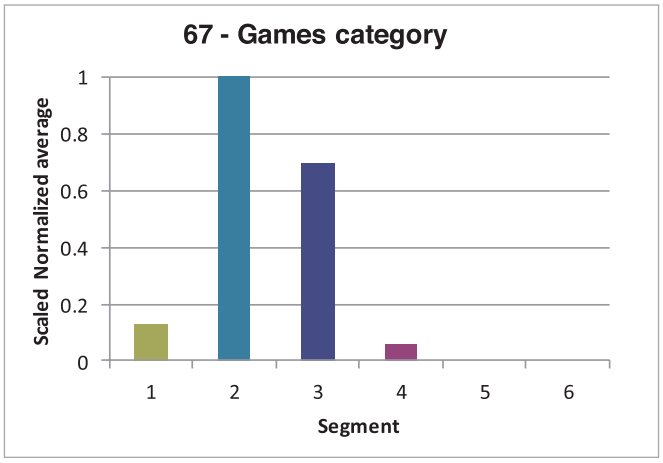

61 - TV (rec) coverage
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

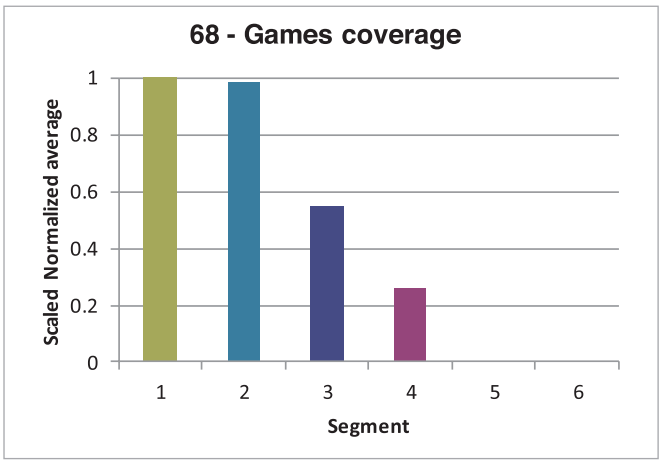

62 - TV (live) freeview category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

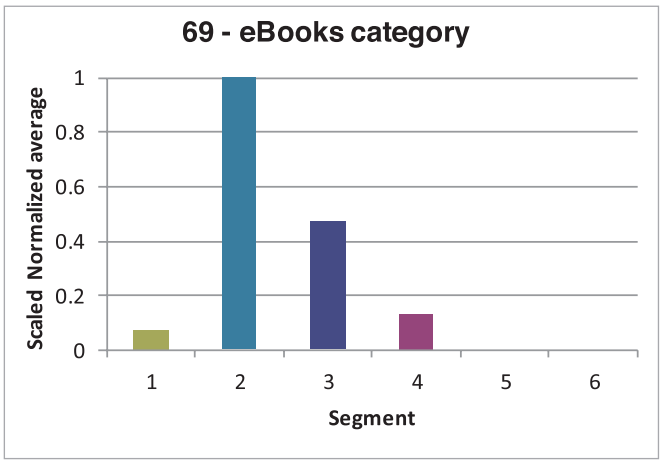

63 - TV (live) subscription category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

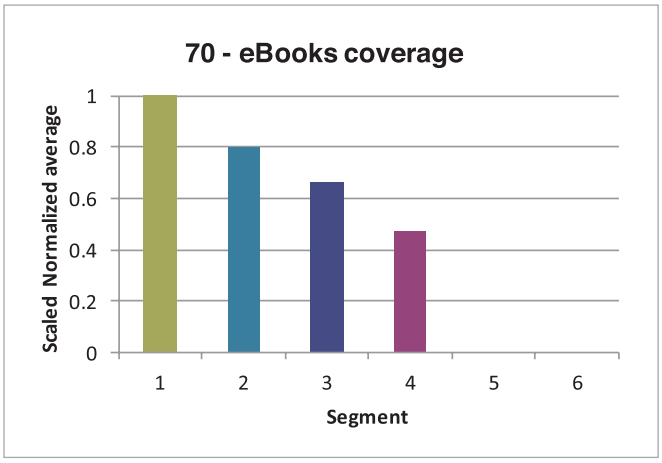

65 - Software category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

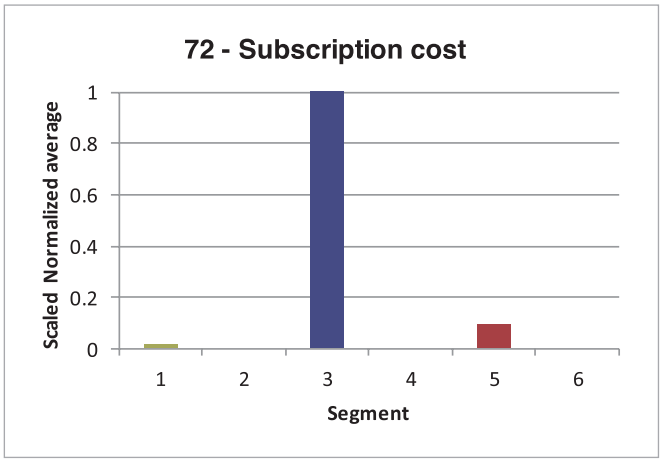

66 - Software coverage
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

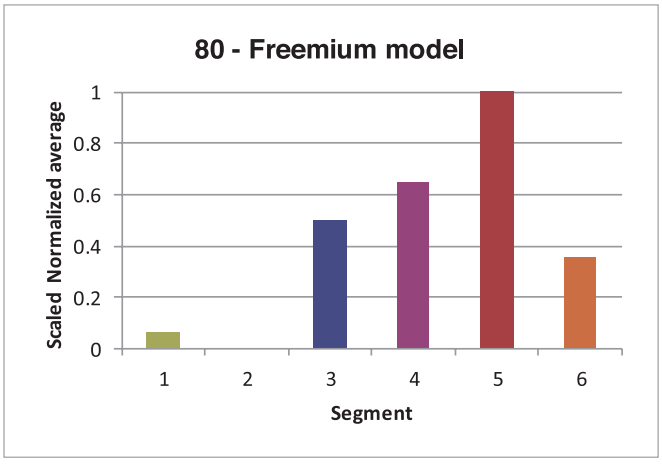

67 - Games category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

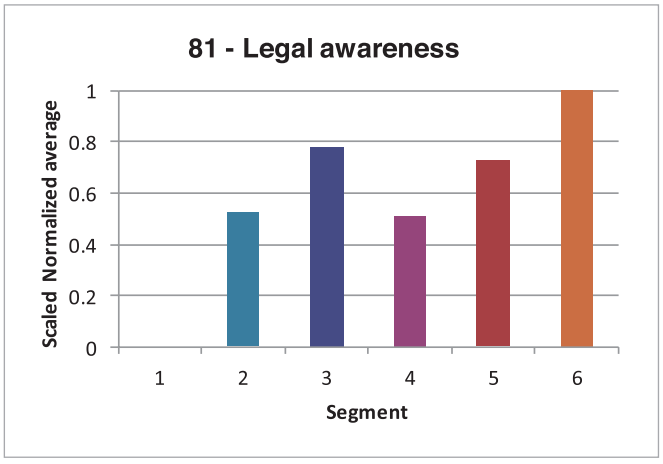

68 - Games coverage
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

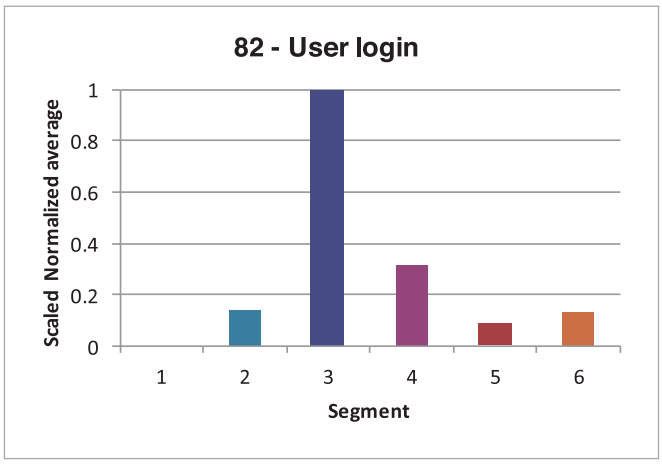

69 - eBooks category
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

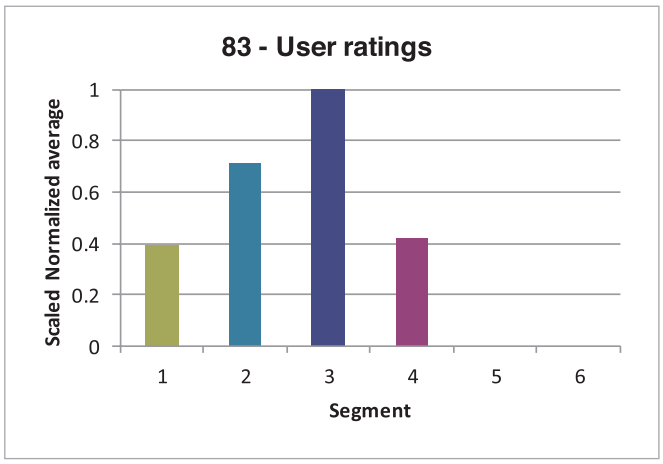

70 - eBooks coverage
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

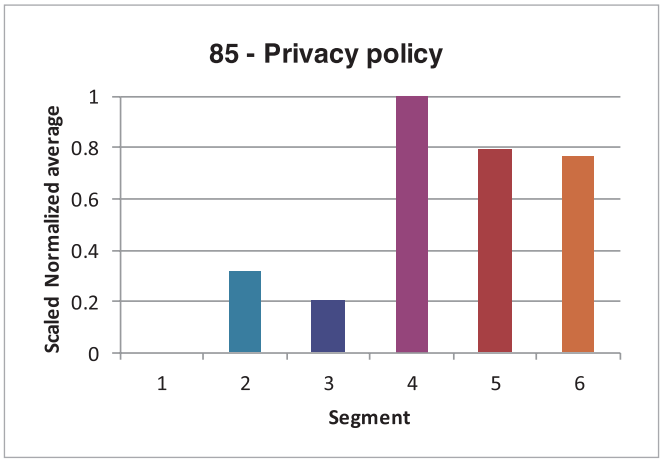

72 - Subscription cost
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

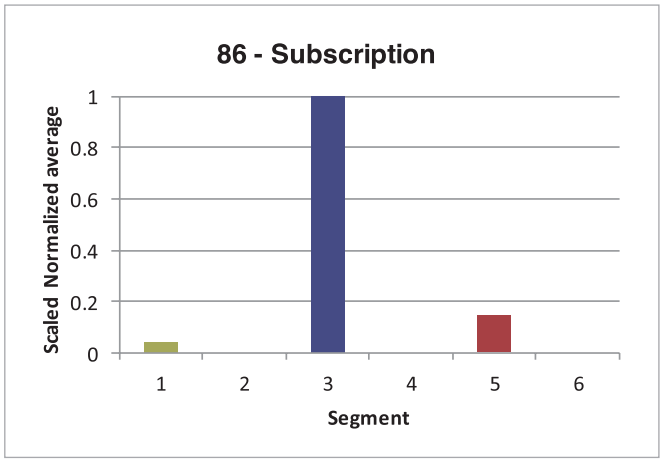

80 - Freemium model
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment


81 - Legal awareness
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment


82 - User login
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment


83 - User ratings
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment


85 - Privacy policy
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment


86 - Subscription
Scaled Normalized average
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
Segment

# C. Glossary

| Term | Definition |
|---|---|
| Actor | A role played by a user or any other system that interacts with the subject, in this case the domain or website being investigated. |
| API | Application Programming Interface - a specification used to communicate with software or online tools and in this research, used to retrieve specific data from third parties. |
| Attribute | A quality or feature regarded as a characteristic or inherent part a domain. |
| Cluster | See Segment |
| CPC | Cost Per Click – the cost to purchase 1000 advertisement click-through (i.e. 1000 instances of a user selecting an advert by clicking on it in their browser) |
| CPM | Cost Per Mile – the cost to purchase 1000 advertisement impressions (i.e. 1000 instances of an advert being loaded in a user's browser) |
| DCMS | UK Government department for Media, Culture and Sport. |
| Decision Tree | A modelling method that works by splitting data into increasingly homogeneous groups using rules based on the various measures. |
| Dissimilarity (distance) matrix | A matrix rather like a road map distance table showing the distance between observations. It is obtained as 1-similarity matrix. |
| Domain | A domain name is an identification string that defines a realm of administrative autonomy, authority, or control on the Internet. In this report the term domain is used to refer to the single string selected by the owner of that domain. It should be noted that additional subdomains (additional strings prior to the domain) are not included in this definition. Example domains:<br>URL: http://example.co.uk → Domain: example<br>URL: http://mail.example.com → Domain: example<br>URL: http://www.mywebsite.net/apage.hmtl → Domain: mywebsite |
| Hierarchical clustering | A segmentation technique whereby clusters are formed by joining together data points one by one sequentially. |
| HTML | Hyper-Text Markup Language |
| HTTP | Hyper-Text Transfer Protocol |
| K-Means | A statistical clustering algorithm using numeric metrics that clusters observations around mean values. |
| Metric | A numeric value which can be measured for a specific entity, in this case a domain. |
| Motivation | A reason for completing a defined action. |
| P2P | Peer-to-peer, a term used to describe networks in which content in distributed from one user to another without being stored on central servers. |
| PCA | Principal Component Analysis |
| Persona | An example of an Actor which represents an extreme of character. |
| Proxy | An indirect method for measuring an attribute |
| RAG | Red, Amber or Green – a traffic lights inspired classification system applied to an entity to illustrate its status where Red normally denotes a negative status and Green denotes a positive status. |
| Random Forest | A modelling method based on randomisation methods which amongst other things can be used to segment data. A Random Forest is a collection of decision trees. |
| Segment | Also called a cluster. A group of entities derived from investigation of their properties or characteristics. In this report a segment typically describes a group of domains which have been derived through algorithmic analysis. |
| Similarity Matrix | A matrix showing how similar data points are to each other. Random forest produces a similarity matrix by measuring the proportion of times, two observations end up grouped together, when the individual random models are run. |
| SOM | Self-Organising Map, otherwise known as Kohonen map. Similar to k-means, but observations are grouped on a two dimensional grid. |
| URL | Uniform Resource Locator |
| WHOIS | A query and response protocol that is widely used for querying databases that store the registered users or assignees of an Internet resource, such as a domain name or an IP address block. |
| WSS | Within Sum of Squares. A method for measuring the tightness of a clustering solution. |

# D. Actor motivations

| ID | Actor | Actor Motivations |
|---|---|---|
| 1 | Website Owner | Site Revenue |
| | | Site Costs |
| | | Business Model |
| | | Content Control |
| | | Freedom of Content |
| | | User Base Size / Site Awareness |
| | | Risk of Prosecution |
| | | New Technology |
| 2 | Contributor | User Base Size / Site Awareness |
| | | Legal Awareness |
| | | Ease of Use |
| | | Community |
| | | User Reward |
| | | Cost of Access |
| 3 | Consumer | User Base Size / Site Awareness |
| | | Ease of Use |
| | | Content Range |
| | | Content Quality |
| | | Community |
| | | User Privacy |
| 4 | Rights Holder | Artist Exposure |
| | | Legal Revenue Impact |
| | | Observability of Consumption |
| 5 | Service Provider | Vendor Revenue |

# E. Actor attributes

| Actor | Persona | Persona Motivations | Actor Motivations = Attributes |
|---|---|---|---|
| Site Owner | Venture Capitalist | - Site Revenue<br>- Site Costs<br>- Content Control<br>- User Base Size / Site Awareness<br>- Risk of Prosecution<br>- Business Model | - Site Revenue<br>- Site Costs<br>- Business Model<br>- Content Control<br>- Freedom of content<br>- User Base Size / Site Awareness (owner)<br>- Risk of Prosecution<br>- New Technology |
| | Altruist | - Site Costs<br>- Freedom of Content<br>- User Base Size / Site Awareness<br>- Risk of prosecution | |
| | Innovator | - Site Costs<br>- User Base Size / Site Awareness<br>- Risk of Prosecution<br>- New Technology | |
| Contributor | Accidental | - Ease of Use<br>- Risk of Prosecution | - User Base Size / Site Awareness<br>- Legal Awareness<br>- Ease of Use<br>- Community<br>- User Reward<br>- Risk of Prosecution<br>- Cost of Access |
| | Egotist/Geek | - Community<br>- User Reward<br>- Risk of Prosecution | |
| | Altruist | - Cost of Access<br>- Ease of Use<br>- Content Range<br>- Risk of Prosecution | |
| | Profiteer | - User Reward (Note: not present in Pilot sites) | |
| Consumer | Unintended | - Ease of Use<br>- Content range<br>- Cost of access<br>- Legal perception<br>- Risk of prosecution | - User Base Size / Site Awareness<br>- Easy of Use<br>- Content Range<br>- Content Quality<br>- Cost of Access<br>- Legal Awareness<br>- Risk of Prosecution<br>- Community<br>- User Privacy” |
| | Casual | - Cost of Access<br>- Content Range<br>- Risk of Prosecution<br>- Ease of Use<br>- Legal Awareness | |
| | Pathological | - Content Quality<br>- Community/Networking<br>- Cost of Access<br>- User Privacy<br>- Risk of Prosecution<br>- Easy to use<br>- Content range | |
| Rights Holder | Amateur | - Artist Exposure<br>- Legal Revenue Impact | - Artist Exposure<br>- Legal Revenue Impact<br>- Observability of Consumption |
| | Corporate | - Content Quality<br>- Legal Revenue Impact<br>- Observability of Consumption | |
| | Independent | - Artist Exposure<br>- Legal Revenue Impact | |
| Service Provider | Dark Ad Agency | - Vendor Revenue | - Vendor Revenue |
| | Mainstream Ad Agency | - N/A | |
| | Payment Provider | - Vendor Revenue | |

# F. Exclusion of applications

During this research a number of stakeholders and industry experts referenced Application software (app), which they consider to be significantly infringing copyright as another entity which should be considered by this research. We conducted a feasibility investigation of two apps to establish whether they could be directly included in the model.

| Application Type | Application description |
|---|---|
| Music mp3 catalogue | Native application for Windows platform |
| | Claims to provide free access to 100 million tracks in mp3 format |
| | No login or subscription required |
| | No evidence of advertising within the application |
| Specific music album streaming | Native Android application |
| | Claims to provide streamed version of specific chart album |
| | No login or subscription required |
| | Advertising present within the application user interface |

*Table F-1: Results of testing the feasibility of including 'apps' in the report*

We found that the methodology presented in this report would be applicable to 'app' segmentation. However, the data available for applications differs significantly from website data and as such the algorithmic segmentation approach being applied in this research could not be used across both groups. This approach requires a consistent and complete data set to be defined for all entities being segmented and this would not be the case for websites and applications.

We decided that, while the segmentation of the copyright infringing application market is potentially feasible and is likely to be of value, it would not form part of the research presented in this report. This is an area that warrants further study.

# G. Collected metrics

| | | |
|---|---|---|
| 1 | Monthly advertising revenue | (<Display CPM Rate> / 1000) x (<Monthly page views> / <Number of pages on user journey>) x <Number of Display Ads on user journey><br>+<br><Text CPC Rate> x (<Monthly page views> / <Number of pages on user journey) x <Number of Text Ads on user journey> x <Text Ad Click Rate> |
| 2 | Monthly subscription revenue | <Unique visitors per month> x <Monthly minimum subscription rate> |
| 3 | Monthly Donation revenue | if(<Donation exists>,<Unique visitors per month>,0) |
| 4/10a | Sum of transaction on revenue | Sum of metrics 4-10 |
| 4 | Music transaction revenue | <Monthly page views> x <transactional value of an album (cheapest)> |
| 5 | Film transaction revenue | <Monthly page views> x <transactional value of a film (cheapest)> |
| 6 | TV (rec) transaction revenue | <Monthly page views> x <transactional value of TV (rec) (cheapest)> |
| 7 | TV (live) transaction revenue | <Monthly page views> x <transactional value of TV (live) stream (cheapest)> |
| 8 | Software transaction revenue | <Monthly page views> x <transactional value of software (cheapest)> |
| 9 | Games transaction revenue | <Monthly page views> x <transactional value of game (cheapest)> |
| 10 | eBooks transaction revenue | <Monthly page views> x <transactional value of eBook (cheapest)> |
| 11 | Equity value | <Brand Rank> |
| 13 | Advertising | If(<advertising present>,1,0) |
| 14 | Subscription | If(<subscription present>,1,0) |
| 15 | Donation | If(<donation present>,1,0) |
| 16 | Transaction | If(<transaction present>,1,0) |
| 18 | Free access | if(<content access is free>),1,0) |
| 19 | A record location | <Country location of A record IP> |
| 20 | Top Level Domain (TDL) location | <TLD Country> |
| 21 | NS location | <Country locations of NS IPs> |
| 22 | Hosting provider | <ASN number of hosting provider> |
| 23 | Takedown mechanism | if(<Takedown mechanism exists>,1,0) |
| 24 | Content hosted | if(<content hosted>,1,0) |
| 25 | Content linked | if(<content linked>,1,0) |
| 26 | Access via stream | if(<content stream available>,1,0) |
| 27 | Access via download | if(<content download available>,1,0) |
| 28 | Shared via P2P | if(<content shared via P2P mechanism>,1,0) |
| 29 | Shared via central servers | if(<content shared via central server mechanism>,1,0) |
| 30 | Shared via distributed servers | if(<content shared via distributed server mechanism>,1,0) |
| 31 | Mobile site | if(<mobile site linked to from homepage>,1,0) |
| 33 | Disclaimer | if(<there a disclaimer\warning to users not to upload illegal content>, 1,0) |
| 34 | Embedding option | if(<site allows content to be embedded elsewhere>,1,0) |
| 35 | Tiered community | if(<tiered access based on previous contribution>,1,0) |
| 36 | Financial reward | if(<adding content provides financial reward>,1,0) |
| 37 | Financial reward value | <upper limit on reward per item of content per month> |
| 39 | Contributor login | if(<user login required to contribute>,1,0) |
| 40 | Contribution cost | <cost to contribute one item (cheapest) of content> |
| 41 | User base | <Unique visitors per month> |
| 42 | Visitor change (1yr) | (<Page views Feb 2012> / <Page views Feb 2011>) - 1 |
| 43 | Visitor change (5yr) | (<Page views Feb 2012> / <Page views Oct 2009>) - 1 |
| 44 | Sites linking in | <Alexa reputation score (sites linking in)> |
| 45 | Page Rank | <Google Page rank> |
| 46 | Link referral | <Percentage of site visits from link referrals> |
| 47 | Social network referral | <Percentage of site visits from social network referrals> |
| 48 | Search referral | <Percentage of site visits from search referrals> |
| 49 | Direct access | <Percentage of site visits from direct access (typed in URL)> |

| | | |
|---|---|---|
| 52 | Social network presence | if(<link to Facebook.com on homepage>,80.4,0) +<br>if(<link to Twitter.com on homepage>,20.7,0) +<br>if(<link to Linkedin.com on homepage>,16.6,0) +<br>if(<link to live.com on homepage>,11.8,0) +<br>if(<link to myspace.com on homepage>,8.5,0) |
| 53 | Content search | if(<content search present>,1,0) |
| 54 | Content comment | if(<content comment>,1,0) |
| 55 | Forum | if(<forum present>,1,0) |
| 56 | Music category | if(<music category exists>,1,0) |
| 57 | Music coverage | <Number of top 10 Official Charts Company UK Albums available> |
| 58 | Film category | if(<Film category exists>,1,0) |
| 59 | Film coverage | <Number of top 10 US box office available> |
| 60 | TV (rec) category | if(<TV (rec) category exists>,1,0) |
| 61 | TV (rec) coverage | <Number of top 10 Hulu programmes this week available> |
| 62 | TV (live) free to air category | if(<TV (live) free to air category exists>,1,0) |
| 63 | TV (live) subscription category | if(<TV (live) subscriber category exists>,1,0) |
| 65 | Software category | if(<Software category exists>,1,0) |
| 66 | Software coverage | <number of top 10 Amazon.co.uk software titles available> |
| 67 | Games category | if(<Games category exists>,1,0) |
| 68 | Games coverage | <number of top 10 Amazon.co.uk game titles available> |
| 69 | eBooks category | if(<eBooks category exists>,1,0) |
| 70 | eBooks coverage | <number of top 10 Amazon.co.uk kindle eBook paid titles available> |
| 72 | Subscription cost | <Monthly minimum subscription rate> |
| 80 | Freemium model | if(<subscription or transaction upgrade option exists>,1,0) |
| 81 | Legal awareness | if(<site contains a legal statement>,1,0) |
| 82 | User login | if(<user login required>,1,0) |
| 83 | User ratings | if(<user rating exist>,1,0) |
| 85 | Privacy policy | if(<privacy policy linked to from homepage>,1,0) |
| 86 | Subscription | if(<subscription required>,1,0) |
| 96 | Ad provider type | if(<Ad Choices logo present on or around first ad on user journey>,1,0) |
| 98 | Ad geography | <Country location of A record IP for ad URL> |
| 99 | Ad agency revenue | Monthly advertising revenue |
| 100 | Card processor logo | if(<VISA / MasterCard / American express logo displayed on payment page>,1,0) |
| 101 | Payment service provider logo | if(<PayPal logo displayed on payment page>,1,0) |

# H. Proxy and complex metrics

Below are descriptions and justifications of:

- How we calculated proxy metrics which indirectly estimate particular metric
- Why we excluded particular metrics
- How we calculated complex metrics

To better understand which category each metric was to be placed we used a RAG status to define a level of confidence in a metric relating to both the underlying data and completeness of the data or calculation where:

R = No metric defined or unable to measure metric

A = Metric has questionable correlation with attribute or is based on non-definitive data

G = Metric correlates with attribute and is definitive

## H.1 Proxy metrics

Below is a list of the metrics for which proxies have been identified. Proxies have been used as it was not possible to directly measure the underlying attribute.

| ID | Metric title | Metric calculation | Proxy justification |
|---|---|---|---|
| 3 | Monthly Donation revenue | if(<Donation exists>,<Unique visitors per month>,0) | It is assumed that all sites take the same average donation size and have the same user conversion rate for donations. This means that the unique visitors will provide a good comparative proxy for donation revenue. |
| 4-10 | Transaction revenue | <site visits per month> x <transactional value of an item of content (cheapest)> | It is assumed that all sites have the same conversion rate of visits to purchases. This means that the cheapest cost of an item (which can be determined from inspection of the site) can be multiplied by the site visits to provide a proxy for transactional revenue. |
| 11 | Brand Rank | <Brand Rank> | It is assumed that the Brand Rank, which is based on site quality, content and user perception provides a good indication of the equity value of the site. This proxy has been used as it is not possible to assess the true equity value of a site due to a lack of data on initial investment costs and potential sales value for the sampled websites. |
| 52 | Social network presence | if(<link to Facebook.com on homepage>,80.4,0) +<br>if(<link to Twitter.com on homepage>,20.7,0) +<br>if(<link to Linkedin.com on homepage>,16.6,0) +<br>if(<link to live.com on homepage>,11.8,0) +<br>if(<link to myspace.com on homepage>,8.5,0) | The presence of a link to the primary website for a social networking site (e.g. “facebook.com”) was selected as it provides a good assessment of the site having a presence on that social network. Selection of the top five social networks in the UK provides a good overview of the range of available social networks. Finally, weighting each network by the total unique visitors (in thousands) ensures that the scale of the social network and its impact on consumer behaviour are taken into account. |
| 57, 59, 61, 66, 68, 70 | Content coverage | <Number of top 10 content available> | It is assumed that content within the relevant ‘top 10’ (e.g. the UK Album top 10, or US Box Office top 10) will provide a good indication of the breadth of content available on a given site as well as provide an indication of the timeliness of content delivery on that site. |
| 73 – 79 | Content cost | <Cost to access one item (cheapest) of content> | Some websites will have a very high number of price points for their content. As data is not consistently available on the sales volumes at each price point the minimum price point was selected as the best indication of the cost of content. |

| | | | |
|---|---|---|---|
| 87 – 94 (not 90, 91) | Content legal impact | <Number of downloads of top content of 2011> | It is assumed that the legal revenue impact of a site is proportional to the number of downloads or streams of content. This means that measuring the total number of downloads or streams of a specific item of content, in this case the top item in the relevant content chart, will provide a comparative proxy for the legal revenue impact. |
| 90 – 91 | TV (live) legal impact | if(<TV (live) category exists>,<Unique monthly visitors>,0) | Live TV infringement has three key effects on legal revenue: to remove subscribers from pay TV services; to disperse advertising revenue; and to put licensing arrangement within the industry at risk. The 3rd affect has a negligible NET impact on the industry in the long term and has therefore been discounted. The first two affects are proportional to the unique monthly visits and this has therefore been defined as the proxy for this metric. |
| 96 | Ad provider type | if(<Ad Choices logo present on or around first ad on user journey>,1,0) | To assess the advertisement provider type, and specifically whether the provider is mainstream or not, the Ad Choices scheme was selected as a suitable measure. Principle I.2.a of the scheme framework states that “Third Parties should provide enhanced notice of the collection of data for [Online Behavioural Advertising] purposes via the Icon in or around the advertisement”. Visual inspection for the Ad Choices logo on the first advertisement on the user journey was therefore selected. However we were advised that that not all signatory companies are yet compliant and compliance is more than just the logo, but the logo should prove a good proxy for the metric We also considered that some sites use multiple Ad providers and as such different ad spaces, or in the case of exchanges a single ad space, may have different compliance statuses depending on the ad provider. Nevertheless the EDA will be maintaining a public list of Ad Choices signatories and those companies which have self-certified – this should be available from June and may be useful in future work |
| 100 | Card processor logo | if(<VISA / MasterCard / American express logo displayed on payment page>,1,0) | The user perception of a site will be altered by the display of a main stream card processor logo. Although it is recognised that the display of a logo does not provide conclusive evidence of the use of that card processor service, the logo will have an effect on user perception and this is what is being measured. |
| 101 | Payment service provider logo | if(<PayPal logo displayed on payment page>,1,0) | PayPal has been selected as representative of the non-card processor Payment Service Provider (PSP) market. It is reported to be the 3rd largest transaction processors after VISA and MasterCard. Therefore, display of the PayPal logo on a sites payment page is likely to significantly impact the users’ perception of that page. |

*Table H-1: The proxy metrics and their justification for use in the model*

## H.2 Excluded Metrics

Below we list the metrics or attributes which we have excluded from the model. The specific justification for exclusions varies, but is generally associated with a lack of available data.

| ID | Metric title | Metric calculation | Exclusion justification |
|---|---|---|---|
| 12 | Hosting cost | <Peak site visits per month> x if(<content hosted or streamed>,1,0) | From investigation of bullet proof hosting services, it is apparent that bandwidth is the primary cost driver. It is assumed that bandwidth is proportional to site visits and as such a proxy based on the peak site visits has been defined. However, peak site visits data was not available within the required timeframe for this research. Future research may wish to estimate the hosting costs. |
| 17 | Content validation | if(<content validation required>,1,0) | It is not appropriate to contribute valid content to the sample websites as part of this research and as such the content validation process could not be tested. |
| 32 | Bleeding edge technology | Unknown | Due to dynamic nature of technology base in this market, it was not feasible to assess this metric as part of this research. Each website would require individual expert investigation and may be worthy of consideration for future work. |
| 38 | Reciprocal reward | if(<adding content provides non-financial reward>,1,0) | It was not possible to define a 'non-financial' reward to enable this to be accurately measured across a large number of websites. |
| 50 | Search terms | <Percentage of selected keywords which appear in top 10 search terms for site> | The top 10 search terms are only available for a relatively small number of sites (from Google's data sets) which would not provide adequate coverage of the sampled websites. Furthermore, no unbiased method for identifying the keyword list could be identified and this would be highly subjective. |
| 51 | Metadata keywords | <Percentage of selected keywords which appear in metadata keywords list> | No unbiased method for identifying the keyword list could be identified and this would be highly subjective. |
| 64 | TV (live) coverage | Unknown | Due to the transient nature of live content (e.g. a sports event may be broadcast for 1 hour only), consistent measurement of its availability across a large number of websites is extremely challenging. In this research it was not possible to capture this data but it may be of interest to future research. |
| 71 | Unknown (Attribute is "Content Quality") | N/A | It was not possible to identify a measurable metric for the 'Content Quality' attribute. This is primarily due to the fact that there are a very high number of content quality levels and no consistent way to assess the content held on a website. |
| 99 | Ad agency revenue | Monthly advertising revenue | It is assumed that Ad agency revenue is proportional to the Ad revenue of a site and that this is consistent between the sampled websites. This means that the Ad revenue of the site (metric number 1) can be used to measure this metric as only a relative value is required. |
| 102 | Payment provider revenue | = Site Subscription + Donation + Transaction Revenue | It is assumed that the Payment Service Provider (PSP) revenue is proportional to the sum of the site subscription, donation and transactional revenues. However, it has not been possible to confirm absolute values for the site subscription, donation and transaction revenue (see the proxy justification for metrics 2-10). It is therefore not possible to sum these and retrieve a figure for PSP revenue. |

*Table H-2: The justifications for the exclusion of some metrics and attributes from the model*

## H.3 Complex metric calculations

In addition to the proxy and excluded metrics, we identified three groups of metrics for specific discussion due to their complexity.

### H.3.1 Monthly Advertising Revenue (Metric 1)

The following calculation was used to estimate the website Ad Revenue:

(CPM Rate / 1000) x (Page views / Number of pages on user journey) x Number of Display Ads on user journey

+

CPC Rate x (Page views / Number of pages on user journey) x Number of Text Ads on user journey x Text Ad Click Rate

Online advertising is a highly complex and dynamic industry in which, generalisations around values such as CPM need to be treated very carefully. The formula defined above was defined by Detica having sought expert advice from Gartner Inc., Nielsen and Enders Analysis. It uses a number of generalisations, including a single blended CPM rate, and this research recognises the limitations of this approach. However, although the absolute values generated in this way cannot be taken in isolation, there is significant value in using this calculation as a comparative measure across the sampled sites. The data points and sources we used in this calculation are:

| Date point | Source |
|---|---|
| Page views | Google website data |
| Number of pages on user journey | Detica direct inspection |
| Number of display and text ads on user journey | Detica direct inspection |
| Text Ad Click Rate | 0.5% |
| CPC rate | $0.55 = £0.34 |
| CPM rate | £3.03 (multiple sources – see below for derivation) |

*Table H-3: The Ad Revenue calculation data points and their sources*

The blended CPM rate of £3.03 was derived through consideration of the four key types of advertising which are present on websites:

- Premium – Served directly by the website and immediately viewable on page load with full screen browser at 1440x900 pixels
- Standard – Served directly by the website but not immediately viewable on page load with full screen browser at 1440x900 pixels
- Ad Network – Served by a third party Ad Network
- Unsold (dormant inventory) – Ad space with a placeholder or no ad content

As shown in Table H-4 below, through measuring the proportion of each type of advertising present on the user journey of the four pilot sites and using industry standard CPM rates[5], we derived a blended rate of £3.03.

| Ad Type | CPM | Percentage of Ads |
|---|---|---|
| Premium | £8.00 | 23.3% |
| Standard | £4.00 | 13.3% |
| Ad Network | £1.00 | 63.3% |
| Unsold (dormant inventory) | £0.00 | 0.0% |
| Blended | £3.03 | 100.0% |

*Table H-4: The derivation of the blended CPM rate*

To assess the accuracy of this approach and the blended CPM rate, we compared the calculated ad revenue of one of the pilot sites with an equivalent licensed service for which true revenue figures were available via published company accounts. We expected that the pilot site should have similar per page view ad revenues as the licences equivalent as there is no reason to believe the CPM rates would be significantly different.

The comparison showed that the pilot site had 15.8 times the number of page views of the licensed service and 15.1 times the ad revenue. This is near parity and therefore consistent with our expectation. Any remaining discrepancy may be explained by the reduced ad market available to the pilot site due to the perception of its involvement in copyright infringement.

### H.3.2 Website Technology (Metrics 24 – 30)

As part of the 'new technology' attribute seven metrics were defined to assess the content location (24 and 25), delivery mechanism (26 and 27) and network structure (28, 29 and 30). These metrics together measure three seemingly related elements, however, upon consideration are independent and only together give you a full view of the technology and implementation being used to share content. The details of these metrics are:

- Content location – is the content (defined in this case as the copyrighted material and not any intermediary link or file) hosted on the website and/or linked to from the website?
- Delivery mechanism – once the user reaches the content, regardless of whether they are on the original website or not, can they download and/or stream the content?
- Network structure – when the user accesses the content, do they do so via a P2P network, from a central server and/or from a distributed set of servers?

*5 Evans, David S., (2009) The Online Advertising Industry: Economics, Evolution, and Privacy. Journal of Economic Perspectives, Forthcoming. Available: http://ssrn.com/abstract=1376607 [18 May 2012]*

Examples of these metrics for hypothetical websites are provided below.

| Example website | Content location | | Delivery mechanism | | Network structure | | |
|---|---|---|---|---|---|---|---|
| | Hosted | Linked | Stream | Download | P2P | Central | Distributed |
| Torrent index | | ✓ | | ✓ | ✓ | | |
| Usenet reporting | | ✓ | | ✓ | | | ✓ |
| Sports streaming | ✓ | ✓ | ✓ | | | ✓ | |
| Digital locker | ✓ | | ✓ | ✓ | | ✓ | |

*Table H-5: Examples of website technology metric results for our test websites*

### H.3.3 Site Referrals (Metrics 46 – 49)

A key element of consumer site awareness is the method by which users are referred to a site. There are four technical mechanisms which we have identified as relevant to this research for site referrals. These are: direct access; search referral; social network referral; and link referral from another category of site.

We engaged Kantar Media Compete[6] to provide the percentage of site traffic from each of these four mechanisms. They gathered this data from a range of sources and a panel of ~2 million US consumers[7]. This data has been harmonised, projected and normalised to provide the figures that were used in this research.

The Kantar Media Compete referral mechanism was established through monitoring the IP traffic on a user's connection and not through monitoring actual user interaction (e.g. clicking a link in a browser). Kantar Media Compete defines a user session as 30 minutes and this determines the classification of direct access. Technical definitions of the 4 metrics are as follows:

- Direct access – user navigates directly to the sample website after 30 minutes of no activity.
- Search referral – user navigated directly to the sample website within 30 minutes of having accessed one of the top eight search engines[8].
- Social Network referral – user navigated directly to the sample website within 30 minutes of having accessed on of the top seven social networks[9].
- Link referral – user navigated directly to the sample website within 30 minutes of accessing any other website (i.e. non Search or Social Network).

Due to technical limitations, this measurement is not able to distinguish between a user visiting two sites consecutively but independently (e.g. by using their bookmarks) and a user clicking a link on a site to access another site. This will mean that there will be a proportion of noise in the referral data. It is assumed that this noise will be evenly spread across all sites and as such will not impact the comparative value of this metric.

*6 Kantar Media Compete, (2012) description, [Online], Available: http://kantarmedia.compete.com/ [18 May 2012]*

*7 After assessment of available UK consumer data, this project proceeded with US data as it offered far higher coverage of the sample domains. US consumer data provides a good estimate of UK data as US and UK consumers have broadly similar internet penetration and language profile. If UK consumer data coverage improves, future research may wish to utilise it.*

*8 In April 2012 Kantar Media Compete define the top eight search engines as: Ask, AOL, DuckDuckGo, Bing, Yippy/Clusy, Yahoo, Dogpile, Google*

*9 In April 2012 Kantar Media Compete define the top seven social networks as: Facebook, Twitter, LinkedIn, MySpace, Pinterest, Google +, Ning*

# I. Data collection methods

## I.1 Overview

To enable a relatively large sample of websites to be included in the segmentation, a range of methods were employed to reduce the manual effort required. These methods were:

- Automated HTML capture and keyword checking – capturing the page source for a defined list of website Uniform Resource Locators (URL) and searching for a set of keywords within the visible text, HTML or any URLs contained in the source code.
- Automated site search and results capture – querying the websites search function through construction of a search URL string specific to that website and capturing the resulting pages for manual inspection.
- Third party API Queries – querying APIs such as 'WHOIS' for technical information regarding a website.

We completed each of these techniques using the Python scripting language. Generic scripts were developed, which we used across all websites to collect publically available data. A list of URLs were required as input to these scripts.

## I.2 Manual data collection

In certain instances the automation techniques were not applicable. For example, it was not possible to complete website search and results capture for sites without search functionality.

In cases where it was not possible to capture a data point for a specific website, a '0' value would be recorded. This was done to ensure that the algorithmic segmentation would tend to group these websites together. For example, if 10 websites were tested for content coverage we might find the following:

- Websites have content search functionality and would receive a value from 0 to 10 depending on how many of the top 10 of that content were returned by the search function.
- Websites that did not have content search functionality and would be assigned a value of N/A.

It should be noted that a website with search functionality may be assigned a value of 0, as it does not return any of the top 10 items of content. This is treated differently from a website without search functionality, which receive a value of N/A. When algorithmic segmentation is completed these websites will be segmented appropriately and websites without search functionality will tend to group together (depending on the values of the other metrics).

Website specific scripting or techniques such as web browser automation were not developed for use in this research. Scripts of this type would allow more data to be captured automatically, but would not have been an efficient approach in this project as this research only seeks to capture data for each website at one point in time. The data that would require this approach was therefore captured manually and is discussed in the following section. If future research in this area seeks to repeat the data capture or run data capture over a period of time, further investigation of site specific code or scripting would be recommended.

Examples of manual data points and the reason that automation could not be applied are listed below:

| ID | Data point | Reason for manual inspection |
|---|---|---|
| 72 | Monthly minimum subscription rate | Automation was applied to remove websites which contain the string "subscription" in their source code. However, without website specific code or scripting the minimum subscription value cannot be automatically recorded and manual inspection is therefore required for the remaining websites. |
| 54 | Content comment | Although the string "comment" could be used as a keyword in automation, it was deemed too common in everyday use to be used in that way and full manual inspection was therefore required. |
| 18 | Content access is free | *No keywords could be defined to assess or filter for this metric and full manual inspection is therefore required.* |

*Table I-1: Examples of data points that require manual inspection and the reasoning*

In addition to collecting these manual data points, we assessed the content coverage metrics by manually inspecting the search results pages that were scrapped as part of the automation process.

## I.3 Rights holder list obfuscation

In order to obtain additional third party data we provided Google and Kantar Media Compete with an obfuscated rights holder list for enrichment. We took approximately 50,000 randomly selected websites from the Alexa Top 1 million sites and added them to a consolidated rights holder list. After consulting with Detica statistics experts we used the following approach based on statistical sampling:

- Sampling should include at least 5% of the population to provide a "forceful" conclusion[10].
- This means that having a sample of less that 5% of the population will not provide a statistically accurate understanding of the population.
- If we reverse this logic, we find that if our sampling rate is 2% we will not have a statistically sound sample.
- Applied to our problem of obfuscating the rights holder list, we can treat the list as a sample of 2% which means we should use a ratio of 1:50 of our list to our random websites to achieve obfuscation.
- If we have 1,000 websites on our rights holder list we will need to use 50,000 random sites from the Alexa Top 1 million lists to obfuscate the data.

## I.4 User journey and search URLs capture

We manually captured a list of the URLs on three pre-defined user journeys to act as input for the automation scripts. This process ensured that we captured the relevant website information as the defined user journeys covered all pages which consumers were likely to interact with. The three user journeys captured were:

1. From website homepage to content via search function.
2. From website homepage to content via browse function.
3. From website homepage to legal and terms and conditions information.

*10 National Audit Office, (2012) Sampling Guide, [Online], Available: http://www.nao.org.uk/idoc.ashx?docId=60e06674-ecfa-4aa2-9fc5-ea61a3d64728&version=-1) [18 May 2012]*

In addition to the user journey URLs, we recorded the search URL for a website, together with the structure of that URL and how the search query was contained within it. We used the search URL in the automation phase to retrieve the results pages for each of the top 10 items of content. In a number of instances we observed that certain websites did not present an identifiable search URL. In such cases we manually collected this data.

## I.5 Automated data collection

We passed the user journey URLs for each website through the automated scripts which captured the relevant source code. The retrieved HTML was then analysed by completing one of the following actions:

- Full source code searched for defined keywords
- Visible text (i.e. non HTML code) searched for predetermined keywords
- URLs contained in the source code searched for defined keywords

This exercise either provided the data point required for use in a metric calculation or provided a filter meaning that some websites could be excluded from further manual inspection.

The search URLs were also used to construct search queries for each item of content in the top 10 lists defined in the model. The scripts then captured the source code of the results page of a search query for manual inspection. As highlighted above, a significant number of data points could not be automated and manual inspection was necessary to obtain these data points.

## I.6 Additional third party data sources

We used third party data sources for data points which were not readily available online or were more easily acquired directly from the data owner. We obtained data points in this way from several sources:

- Google – Historic page views, Ad Planner data and brand ranking.
- Kantar Media Compete – Website referral information.
- Alexa – Reputation Score.
- Robtex/DNS/ WHOIS lookup – IP address and Website data.
- Team Cyrmu Community Services[11] – ASN and Country codes.
- IANA[12] – Top level Website data.

The table below highlights the proportion of each data source making up the metric-based segmentation model.

## I.7 Metric calculations

Once we collated all the data points through automation, manual inspection and third party data capture, we calculated the model metrics. As discussed in Chapter 4 model metrics are either: single data values; composite calculations; or logical statement. Depending on their types, they require one or more captured data points.

The metrics resulting from these calculations form the dataset on which algorithmic segmentation was completed.

## I.8 Data quality

We consider data quality very important and carried out several reviews to ensure the data was of sufficient quality to be used in the model.

Where data points were missing from the Google data we applied the following methodology:

### I.8.1 UK Ad Planner data

The page view and unique visitors data is used in a wide range of metrics and is critical to the modelling. It is therefore important that there are no gaps in this data and the average value for the top 100 websites was used for websites with missing data. Whilst this is not ideal, it reduces the impact on segmentation for these sites.

### I.8.2 Historic Global Page views data

The global page view data provided is used in calculating two visitor change metrics. Any missing data points were supplemented by additional data collected from Domain Tools[13]. Information about a websites history was used to replace missing data in the following way:

1. If the website existed, but there was no Google data present – the average visitor change value for the top 100 websites would be applied.
2. If the website did not exist then we replaced the missing data with a '0'.

### 1.8.3 Collection of validation data

In order to provide a level of confidence in the clustering solution, we undertook at second data collection exercise, with the aim of utilizing this data to validate any clusters identified in the training data.

This additional list of websites was obtained in the same manner previously outlined in this chapter. During collection of this validation data, we excluded 16 website addresses in most cases this was because the website was no longer in use. The validation data set consisted of 104 websites.

To ensure the data set was of sufficient quality for segmentation we placed both data sets (the training and validation sets) through a series of quality control methods.

We verified the collected data by using multiple individuals to ensure that there were no conflicting results for similar metrics and that the data was collected consistently. In the first instance we manually verified a randomly selected 25% of all of the automated and manually collected data points that were used to calculate the metrics.

The data collected through running automation scripts was verified to ensure that there were no systematic errors present. However, we identified a significant number of false positives during this exercise. Therefore, we decided that manual verification would be conducted for all sites, and any false positives would be replaced. On inspection, the false positives were generally present on websites containing more complex HTML e.g. those employing significant amounts of JavaScript. It was not however necessary to replace the automated results from the Top 10 content searches as these required manual inspection as part of the metric calculation.

| Proportion of data captured | Google Data | Kantar Media Compete | Alexa Data | Robtex Data | Team Cyrmu Community Services | IANA | Automated collected Data | Manually collected Data |
|---|---|---|---|---|---|---|---|---|
| Training data [153 websites] | 7% | 4% | 1% | 3% | 1% | 3% | 17% | 64% |
| Validation data [104 websites] | 7% | 4% | 1% | 3% | 1% | 3% | 6% | 75% |

*Table I-2: The proportion of each data source that contributes to the metrics*

*11 Team Cyrmu Community Services, (2012) description, [Online], Available: http://www.team-cymru.org/Services/ip-to-asn.html [18 May 2012]*

*12 Internet Assigned Number Authority, (2012) description, [Online], Available: http://www.iana.org/domains/root/db/ [18 May 2012]*

*13 Domain Tools, (2012) description, [Online], Available: http://domaintools.com [18 May 2012]*

# J. Algorithm selection

The field of segmentation methods, also called clustering, is rich in algorithms and we were faced with an enormous choice. We made a determination of the optimum approach to segmentation based on accepted research and a detailed examination of how well the data fit with respect to the key alternate approaches.

## J.1 Review of segmentation approaches

One class of method starts with a data table of observations and measures derives a metric, usually Euclidean, to describe the distance between observations. Attempts are then made to cluster the observations that are closest to each other.

K-means and Self Organising Maps (SOM, also known as Kohonen maps) follow this approach. K-means is commonly used, in part because it is one of the most efficient computationally and often the only feasible approach in this era of big data sets. SOMs are also frequently used because they employ an algorithm similar to k-means but constrain the solution to lie on a 2 dimensional grid, which make the solution amenable to graphical interpretation. For this study, we dealt with hundreds rather than millions of web sites so this constraint was not paramount and we could consider other methods.

Another class of algorithms perform hierarchical clustering. These need not start from the raw data, but from a dissimilarity or distance matrix that describes how far apart the observations are. The clusters are formed sequentially by joining together the observations that are closest to each other.

When it comes to choosing which clusters to join together, there are a variety of ways of defining the distance between clusters, each giving rise to a different method of hierarchical clustering. A key step in all these methods is how to define 'distance', however, the nature of the variables measured on the websites themselves made this anything but straightforward.

The measures were a heterogeneous mix of continuous, binary and categorical. Particularly problematic were the measures which, though continuous, were not applicable to certain sites because they simply did not do the activity in question. This was not missing data in the sense that there is a value and we do not know what it is, it was missing in the sense of 'Not Applicable' and was clearly a key characteristic for segmenting the sites which we did not want to lose in the analysis.

It is important to note that setting up an extra binary variable to flag applicable/not applicable does not get around the problem satisfactorily because a value was still required as a substitute for the missing value of the original variable and whatever arbitrary number was chosen would affect the result.

## J.2 Testing alternative segmentation methods

We evaluated k-means and SOM using dummy variable flags to describe the categorical variables in the standard manner.

For the hierarchical method we decided to pre-process the data using Breiman's Random Forests[14] to produce a similarity matrix, or proximity matrix, between the sites. Breiman points out that a dissimilarity matrix (calculated as 1-similarity matrix) yields a Euclidean distance matrix that is suitable for use in cluster analysis and gives examples in his Wald lecture[15]. We then used agglomerative clustering on this distance matrix to produce a hierarchical clustering solution for the segmentation. Random Forests, as the name suggests, works through intensive use of the principle of randomisation. The model is fit many times - in our case we used two thousand - on many samples of data and many choices of measures. The patterns that come through consistently over all the randomisations are the ones that determine the final outcome. The spurious ones are averaged out. The underlying model used by random forests is a decision tree. Decision trees have the great advantage of being able to cope with the heterogeneous mixture of measures that we were faced with in this study. They also require minimum data preparation. This was particularly attractive because, as mentioned previously, many of the alternative methods of dealing with messy data require the analyst to make arbitrary decisions as to how to convert the categorical data into Euclidean distances, whereas Random Forests is data driven. Furthermore, as the individual trees making up the "Forest" are splitting algorithms, the results are invariant to scaling of the measures and give the same result under any order preserving transformation of them.

## J.3 Random Forests and agglomerative clustering

We found that, having obtained a distance matrix from Random Forests, agglomerative clustering yielded a rich description of how the segments or clusters were formed.

Agglomerative clustering is a hierarchical method, so called because the solutions are nested in a hierarchy. If A is in the same cluster as B in the 6 cluster solution, then A will be in the same cluster as B in the 5,4,3,2 cluster solutions.

K-means and SOM do not have this property. When using k-means, we had to state how many clusters we wanted up front and run each solution for a given number independently. Similarly in SOM we had to choose a grid for the solution up front. There was no guarantee of a logical ordering of the clusters. Hierarchical clustering allowed us to plot a dendrogram. This showed the order in which items were joined and allowed us to define sub-segments within the major segments without having to rerun the analysis.

The various flavours of hierarchical clustering arise from how the distance (dissimilarity) between the clusters is defined as they are forming. We used the 'complete' method; in this the dissimilarity between 2 clusters is defined as the largest dissimilarity between any 2 members, one from each cluster. The clusters that are joined next are then the ones that are least dissimilar. The complete method is strongly biased toward producing compact clusters with roughly equal diameters, and it can be severely distorted by moderate outliers. In our case we were starting from a dissimilarity matrix so the outlier problem which is a feature when starting with raw metrics was not a concern. The complete method does ensure that all items in a cluster are within some maximum distance (dissimilarity) of one another, which was a desirable feature for our particular purpose. We wanted to avoid chaining whereby site A is similar to site B because they share feature X and site B is similar to site C because they share feature Y but sites A and C have nothing in common for example.

When comparing the results of the three approaches, the combination of random forests and hierarchical clustering were the most satisfactory. The k-means and SOM approaches gave no indication that they had found a natural set of clusters leaving the question of how many segments unanswered; the various groupings were not amenable to simple interpretation and were inconsistent from solution to solution. The results from the Random Forests and agglomerative clustering did, however, suggest six interpretable clusters (segments). This difference in performance was expected to be due to the ability of Random Forests to factor in the non-metric measures into the distance matrix and in particular capture the fact certain sites did not participate in certain activities into the result in a logical way.

## J.4 Analysis of Within Sum of Squares

We decided on Random Forests as the optimum approach to segmentation and ran the initial segmentation using the training sample of 153 web sites. This produced the similarity (proximity) matrix. We then ran a hierarchical clustering on the distance matrix derived from this and had to decide on the number of clusters for the solution.

Selecting a definitive number of clusters as a solution is a problem that is still being extensively researched and many of the techniques make assumptions which do not hold here. A simple approach used to compare different clustering solutions is to calculate the total

*14 Breiman, L. (2001) Random forests. Machine Learning 2001, 45:5-32*

*15 Breiman, L. (2002) Looking inside the black box, [Online], Available: http://stat-www.berkeley.edu/users/breiman/wald2002-2.pdf [18 May 2012]*

Within Sum of Squares (WSS) for each cluster and plot this against the number of clusters. This is defined as:

∑(over clusters k) { ∑ (over points within cluster i) (Xik – m k)2 },

where $X_{ik}$ is an individual point and $m_k$ is the mean of the points within the cluster.

If the clustering is tight then the total WSS for the clusters should be much smaller than the sum of squares for the whole population and with very tightly defined clusters a WSS plot will be L shaped, displaying a distinct elbow point.

As we started from a data table of heterogeneous measurement types, this measure is not perfect as only the reduction in the within sum of squares for continuous measures can be incorporated, the mean value not being computable for the categorical data points.

We plotted the WSS for the numeric variables only, aware that the clustering due to categorical variables was being discounted. These measures were all scaled to prevent scale effects dominating the result, but the formula is not scaled by the number of data points, so that a larger data set will tend to have a total larger within sum of squares value.

We examined the effect using different numbers of clusters with a WSS plot. We did not find an L shaped with a distinct elbow point; this told us that we did not have a few tightly defined clusters. We did, however, find a step decrease in the WSS when going from 5 clusters to 6. As a result, we chose six clusters. This served the purpose of creating segments of sites within each of which the site profiles were broadly similar.

## J.5 Dendrogram visualisation

A dendrogram is a useful graphical device to show the results of hierarchical clustering. In this report, the trees are upside down. The vertical axis labelled height is a measure of the dissimilarity at which the clusters were joined. It is on a scale of 0 to 1, 0 being identical and 1 being as different as possible. At the tip of each branch of the upside down tree is a number. This is the observation number for the website in question. The dendrogram shows you at what level of dissimilarity it was joined onto either another website or another cluster (segment) being formed.

The joins made towards the bottom of the diagram are those between the most similar websites. Choosing any particular cut off on the height axis will give you a solution with a corresponding number of clusters.

If you plot a horizontal line at the cut off level chosen, then the number of clusters formed below that line + the number of yet to be joined observations will be the number of clusters.

In the detail below sites 52 and 61 are joined at a dissimilarity of about 6.0 closely followed by 71 and 75 and so on.

If we chose a cut off of 0.9 then we would have 9 clusters (1,6) (11,25) (28,80,79 39+points below the detail) (8,29,93,21,94,74) (48,67,82) (10,90,12,91) (17,89,92) (52,61,71,75,62,73,96,32,81).

A cut off of about 0.97 gives the 3 clusters show by the red rectangles in Figure 1.

We produced the dendrogram shown in Figure J-2 to summarise the hierarchical clustering solution to match the six segment solution indicated by the Within Sum of Squares analysis. Again, the presence of a small number of very tight clusters is not evidenced, as the larger clusters were formed with higher cut offs for height. Instead the picture is one of segments of web sites that are broadly similar in their profiles, but with some differences as we expected. The sites within each cluster for the six cluster solution are delineated by the red rectangles.

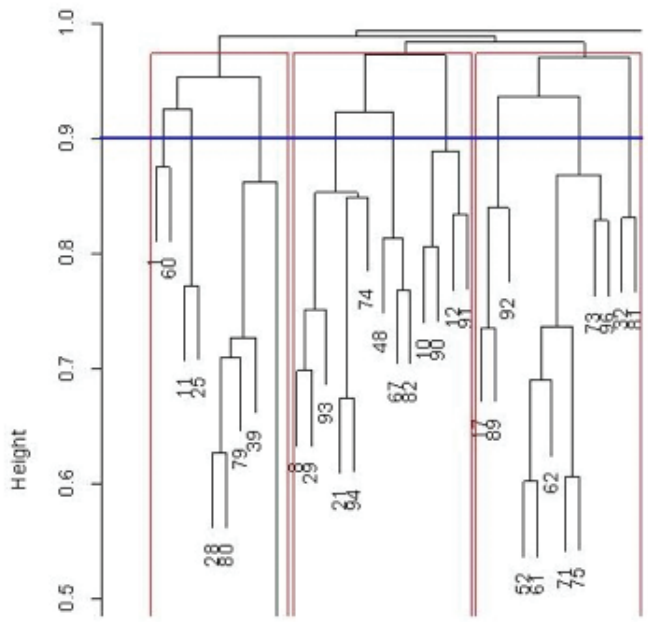



*Figure J-1: Example dendrogram showing three segments*

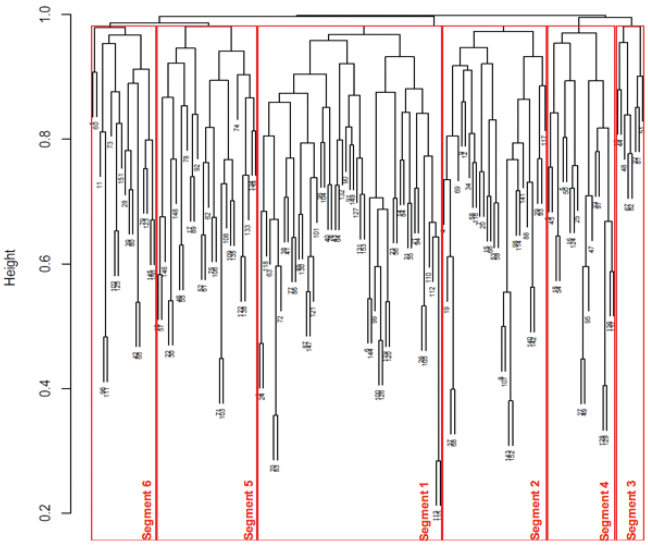



*Figure J-2: Dendrogram of 153 sampled sites, showing the six segments*

# K. Principal component analysis

Principal components analysis (PCA) is a dimension reduction technique with many applications. In the area of data analysis and exploration it is frequently used to summarise highly dimensional data in fewer dimensions to enable the data to be examined graphically, which is what we have done in this report.

PCA is based on the method of decomposing a square matrix into its eigenvalues and eigenvectors (German eigen=singular). In the case of data such as ours where a wide variety of scales has been used it is generally advisable to perform PCA on the correlation matrix of the data. Mathematically the result of the process can be described as follows:

## X=RT*A*R,

where X is the original p x p square matrix, R is a p x p rotation matrix of eigenvectors and A is a p x p diagonal matrix with the eigenvalues running along the diagonal.

Correlation matrixes are positive semi-definite and it can be shown that in such a case the eigenvalues will all be >= 0 and the eigenvectors will be orthogonal. That is RT*R=I. Also |X|=|A| and trace(A)=p.

The eigenvectors can be used to rotate the original data into this new orthogonal space. These are the principal components and are in effect a weighted sum of the original measures making up each observation. The general practice is to plot the principal components corresponding to the two largest eigenvalues against each other as these will display the directions of maximum spread of the data.

Visually, if you can imagine your data as a rugby ball shaped cloud, then PCA will simply rotate the ball so that the longest axis and then the second longest axis line up with the directions required for plotting. It will do this for as many dimensions as the data has choosing in turn the direction of maximum variation (spread) orthogonal to all the directions already chosen.

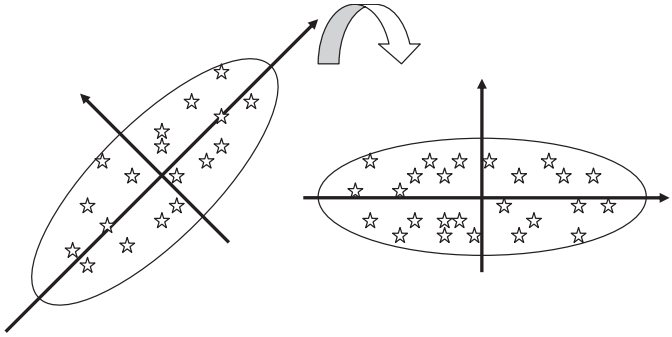


*Figure K-1: Schematic describing Principal Component Analysis*

Interpretation of the principal component axes is possible in a limited way by computing the correlation coefficient of the principal component with each of the original measures in turn. These are called the loadings and the larger ones have been annotated on the plots. There are a wide variety of techniques for creating plots with even more interpretable axes, one of the most popular being varimax which discards the components corresponding to small eigenvalues and rotates the data again in order to make as many of the loadings either close to +/– 1 or 0. As the object of doing PCA in this report was to give a broad overview of the differences between the segments and we have provided narrative elsewhere of the differences found, we did not consider it necessary to do this.

Figure K-2 below containing the 153 sites in the training data in the space of the 1st two principal components. The locations of the individual sites have been colour coded according to their segment membership.

It must be emphasised that this PCA plot is a simplification that gives a rough idea of what is happening with the data. The interesting feature in this plot is that Segments 2, 5 and 6 appear more tightly defined than the other segments. Additionally, Segment 5 and 6 appear to share a number of features in common.

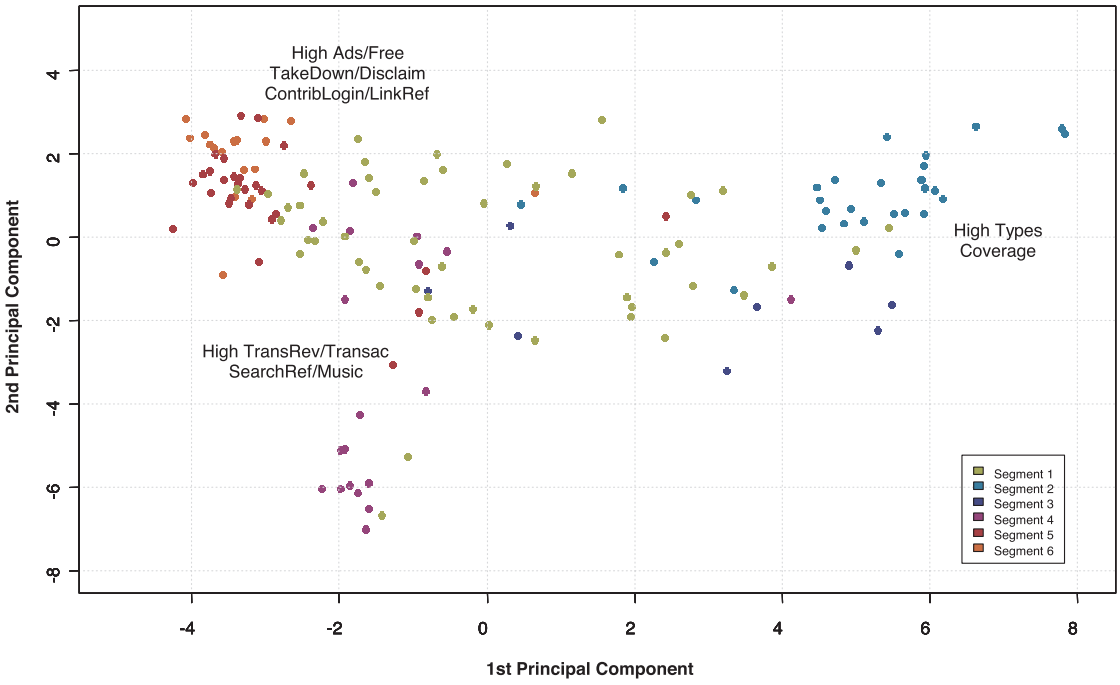



*Figure K-2: The six segments highlighted within the first 2 principal components in order to validate the segmentation*

This document was updated on the 4th July with the following corrections:

- Page 21 & 22: X-axis of Figures 3-5 and 3-6 updated to correct presentation error
- Page 23: Section number on Figure 4-1 corrected
- Page 33: Percentage values for first sector updated to correct presentation error