Norwegian University of Science and Technology

Master's thesis, spring 2014

# User adaptation in an anonymous application setting

*Author:*
Jonas Myrlund

*Supervisors:*
Prof. Heri Ramampiaro
Prof. Helge Langseth

June 2014

*"C is for cookie, and cookie is for me; C is for cookie, and cookie is for me."*

Cookie Monster, Sesame Street

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# *Abstract*

Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Computer and Information Science

Master of Science in Computer Science

**User adaptation in an anonymous application setting**

by Jonas Myrlund

The goal of the project in this thesis is to explore the viability of an approach to user adaptation where the application context is significantly more constraining than in most cases seen in previous academic work.

In this project I describe a system for rolling out product features incrementally in an optimal way, based on feature adoption statistics within user segments. In other words, the described system allows for simple personalization of the product while experimenting with different feature variations.

Identifying how different classes of users use an application differently can be useful on several levels, and it is often the case that some are more desirable than others. This can be due to its associated users generating more revenue, using the product more, inviting their friends, or similar. This project describes a system capable of not only identifying user classes based on user behavior, but more importantly: a framework for identifying the most effective ways of adapting the product to these user class segments, in effect driving users in a desirable direction.

We find that there are indeed clear differences in feature adoption across the identified user segments. However, due to uncertainty caused by domain constraints, it is uncertain to what extent the results generalize.

# *Acknowledgements*

First and foremost, I would like to thank my supervisors, Heri Ramampiaro and Helge Langseth. While providing me with heaps of space to do things my own way, they have been an immense help in finding the academic angle for the project, and in continuously steering me back on course every time I've lost track of the goal.

Telenor Digital AS have allowed me to write this thesis with them, and for this I am very grateful. Especially big thanks to Svein Yngvar Willassen, my supervisor at Telenor Digital, who both talked me into doing a master's project with appear.in, and assisted me greatly along the way.

Obviously, a big thanks to the entire appear.in team. It is extremely inspiring to do research when you thoroughly believe in the application you are working to improve. Ahead of me throughout the entire project period, they have improved the application immensely and attracted huge amounts of users who generated lots of data – without which this thesis would have been impossible to write!

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The goal of the project in this thesis is to explore the viability of user adaptation in applications significantly more constraining than in most cases seen in previous academic work.

The application case is an anonymous, web-based video conferencing service called `appear.in`. The most important part of that description is the term "anonymous", as we not only need to deal with a complete lack of demographic information about the users – we do not even have the ability to consistently identify them.

Previous work in the area of user adaptation has often adhered to the concept of improving the recommending of some kind of resource – often being reducible to the task of finding relevant items to sell to a given user. The goal of this project, however, is not to *find good things*, it is to *optimize the user experience of an application*. More specifically, the goal is to be able to predict which version of the application will lead to better performance for each user, down to individual feature level.

Let us start off with some background on why I believe this is an especially exciting time for user adaptation in the context of "simple" web applications, before section 1.2 introduces the specific application case of appear.in. Finally, section 1.3 will provide an introduction to the realm of user adaptation.

## 1.1 Background and motivation

The web is an interesting venue in which to explore new ways of user adaptation and personalization. Even today, when we are able to build almost every kind of application

right in the web browser, a fundamental concept remains: there is still a server serving the application code every single time a user opens the site[1].

This is the modern web: a place where applications are nearing the power of traditional desktop applications, but with the potential of being continuously tailored to suit each user individually.

Moreover, as there is no installation step – no technologically required "setup" – many simple web applications take this further by not requiring any kind of identification, removing the unnecessary friction it would otherwise entail. As the web moves more and more in the direction of sophisticated web applications, more and more applications of this sort pop up.

I believe this new breed of web applications has come to stay, and want to investigate to what extent we can apply user adaptation techniques to further enhance them. Can we do any interesting user adaptations without actual users and without consistent identity? And if so, what can we hope to achieve?

### 1.1.1 The modern web

When we say that web applications are nearing the power of traditional desktop applications, we are of course talking about the introduction of HTML5.

HTML was primarily designed as a language for describing scientific documents, and little more. The last decade, however, the concept of web applications has thoroughly established itself, while lacking clear standardization efforts from the W3C. The HTML5 specification is an attempt to remedy this [1], by providing standards and guidelines for the browser vendors on how to implement a wide range of common APIs.

In practice, these APIs lets websites do things like:

- play audio and video[2]
- generate graphics[3]
- access your webcam and microphone[4]
- handle and manipulate arbitrary files[5]

[1] Caching aside, of course.
[2] http://dev.w3.org/html5/spec-author-view/video.html
[3] http://www.w3.org/TR/2dcontext/
[4] http://www.w3.org/TR/mediacapture-streams/
[5] http://www.w3.org/TR/FileAPI/

- send and recieve data over full-duplex socket connections[6]

...as well as a wide range of other things.

In general, HTML5 enables web applications to do most of the things one would need plugins or native applications for just a few years ago.

One of these new API specifications is called WebRTC[7], and it is the last piece of the API puzzle enabling applications like appear.in.

## 1.2 The application case

For a long time, developing video conferencing services was an extremely challenging discipline, and as a result the market has consisted of a correspondingly low number of actors. However, this trend is currently in the process of being shaken and turned on its head with the introduction of HTML5; more specifically, with the introduction of the WebRTC specification.

### 1.2.1 The enabling piece of technology: WebRTC

As the name implies, WebRTC handles real-time communication, but for our case, the important aspect of the technology is that it is able to do so peer-to-peer. Although it is designed to be a protocol for exchanging arbitrary data between peers, it is particularly geared toward multimedia. For instance, the traditionally cumbersome task of setting up a two-way audiovisual connection is now a matter of dropping around 40 lines of boilerplate Javascript into a web page[8].

Although the WebRTC specification at the time of writing still officially is a *working draft* in the W3C[9], most of the large browser vendors have already implemented it. Consequentially, a plethora of applications leveraging this technology are already available, with ever more being launched every month.

[6] http://www.w3.org/TR/websockets/
[7] Web Real-Time Communication.
[8] For an excellent introduction, see: http://www.html5rocks.com/en/tutorials/webrtc/basics/.
[9] The latest specification can be found here: http://dev.w3.org/2011/webrtc/editor/webrtc.html.

### 1.2.2 Introducing appear.in

One of these applications is called appear.in, and it will serve as the main use case in this thesis. Like many other WebRTC applications, appear.in concerns itself with video conferencing.

The idea is simple enough: a conversation happens between users who are in the same room at the same time. The central idea, though, is that the *conversation* is identified solely by the URL in use, and not in any way by the peers connecting. There is no notion of *calling* someone – you simply meet up in a room and talk. As an example, if any two people are visiting `https://appear.in/ntnu` at the same time, they will see and hear each other and can start chatting away.

For a showcase and a thorough breakdown of the application, please see section 2.1.1.

This view of a conversation as not really being an enitity in its own right, but rather an *effect* of people being in the same room at the same time, breaks with the traditional model of audiovisual communication. Traditionally, talking to someone not present has been a process of one person *calling* the other one, with group conversations usually being nothing more than an extension of this concept. appear.in doesn't concern itself with distinguishing between callers and callees, has no simple concept of a "conversation", and generally does not enforce any particular way of using the service – apart from requiring the conversation venue, the "room", to be identifiable by a string of characters.

### 1.2.3 Usage patterns

Until appear.in, this particular way of thinking about audiovisual conversations hasn't been a commonly seen pattern. However, the simplicity of the room concept opens the service up for a wide variety of uses: in addition to traditional video calls, we've already seen it used for everything from virtual offices and team meeting rooms to baby monitoring and remote tutoring, just to name a few.

These wildly varied use cases are where the motivation for this research project stems from:

1. If the users' behaviors are quantified, will any clear and distinct usage patterns emerge?
2. If so, can the different uses be better served by dynamically adapting the product to fit each of them?

### 1.2.4 Anonymity and privacy

Before moving on, let's define some words and concepts that will be central to this section, and to the rest of the thesis. The following definitions have been taken from the Merriam-Webster online dictionary[10].

**Anonymous**
: Lacking individuality, distinction, or recognizability.

**Identity**
: The distinguishing character or personality of an individual.

**Pseudonymous**
: Using a pseudonym: a name that someone uses instead of his or her real name.

More plainly, this thesis will use these words in the following ways to describe the user base.

**Anonymous**
: Not being recognizable as a person from the collected user data.

**Identity**
: The ability to be distinguished from other users over time.

**Pseudonymity**
: Being able to identify users without actual personal information.

appear.in is an anonymous communication service. No personal information is ever collected about the users, and not even IP-addresses or geolocational data is logged on an individual level. By tracking individual *browsers* using cookies, however, we can track users – or more precisely, browsers – over time.

By logging various events that we deem interesting along with a cookie value identifying the browser, we can reconstruct user sessions, and connect them to the application users pseudonomously. By "pseudonomously" it is meant that we identify the users solely by a random string set in their browser.

This all opens a wide series of questions bordering to sociological aspects of web usage:

1. To what extent can an anonymous web service be adapted?

[10] http://www.merriam-webster.com/dictionary/

2. Is user behavior enough to provide a satisfactory personalized user experience?
3. Will the identifying cookies live long enough to enable a worthwhile adaptation effort?

These issues will be revisited throughout the thesis.

#### 1.2.4.1 Absence of identity

Several others have written about how privacy constraints impact personalized systems [2, 3]. In many ways, the very nature of anonymity is an extension of privacy. However, the absence of *identity* presents an entirely different challenge.

There are ways of coping with the absence of user identification. Kobsa, for instance, describes an approach that makes heavy use of pseudonyms designed around this problem [4]. However, appear.in is not only a pseudonymous service – it is a fully anonymous service; there are no user accounts, and there is no login.

This all makes tracking users very problematic.

#### 1.2.4.2 Tracking users

appear.in runs in the browser. The first time a user enters the site, a cookie is set with a random value uniquely identifying the browser over time. This value is sent with every event to the instrumentation service used for user analytics.

Unfortunately, using only cookies for identification has its clear downsides. Should the user clear the browser cache, switch browsers, use the browser "incognito"[11], or simply use multiple machines, then different user ids will be generated for each case.

Without tracking more user information – IPs and locations, for instance – there is no easy way of tying these user ids together[12], and to reason about them as a single user. However, collecting this information about users goes against appear.in's privacy policy, and it has been deemed more interesting to see what can be done without crossing this line.

The impact of tracking users based solely on cookies is discussed further in section 2.1.2.2.

[11] Google lingo for private browsing, where previously set cookies are not available (among other things).

[12] Some suggestions on how to solve this is found in the suggestions for further work, in section 5.3.

## 1.3 User adaptation

Before answering the question of *how* to do user adaptation, let's look into the more fundamental issue of *why*. Why would we want to adapt a product to its users?

Dijkstra once described a fundamental way in which humans and computers differ with a short anecdote involving piano playing and his mother [5]:

> To end up my talk I would like to tell you a small story, that taught me the absolute mystery of human communication. I once went to the piano with the intention to play a Mozart sonata, but at the keyboard I suddenly changed my mind and started playing Schubert instead. After the first few bars my surprised mother interrupted me with "I thought you were going to play Mozart!". She was reading and had only seen me going to the piano through the corner of her eye. It then transpired that, whenever I went to the piano, she always knew what I was going to play! How? Well, she knew me for seventeen years, that is the only explanation you are going to get.

Humans constantly monitor the world around them, and model the objects and people in it. The kind of implicit modeling and predictive behavior is something computers generally have a hard time dealing with.

However, the digital world around us is becoming more and more complex, and this complexity will at some point overwhelm users. Adapting to the user's expectations may help in minimizing friction and confusion as users meet this increase in complexity [6].

For a survey of the field of user adaptation, see section 2.2.

### 1.3.1 A new context

The application case of appear.in presents a few novel qualities for an adaptive system to deal with. Firstly, we want to do feature-level user adaptation of a web application, and secondly, we want to do so with extremely challenging privacy constraints.

What being an *application* running on the web entails for the relevance of previous research, is deferred to section 2.2.2. The impact of tight privacy constraints, on the other hand, requires some discussion before we take on the concrete problem specification of this thesis.

As we shall see throughout chapter 2, most of the surveyed adaptation systems have a notion of *users*, who can be tracked over time. Furthermore, most traditional user

modeling systems even assume either the availability of demographic information or some notion of *consumption of resources*, which in turn both can aid in constructing user models.

We, however, do not enjoy such luxuries. A large part of this thesis will look at how far we can get with only behavioral data, when even identity is at best an unstable notion.

Section 2.2.5 will take a closer look at previous research on the notion of "privacy versus personalization".

## 1.4 Problem specification

An essential part of the research will consist of determining whether anonymous users, like the ones described for the appear.in application, will be clearly separable into clusters based purely on their behavior – the only data available about them.

Given a set of identified user classes and a set of product variations, we want to find out how each variation affects each user class. When launching a new feature, for instance, will it be adopted differently by different classes of users?

To find out, we perform the following steps:

1. Separate users into user classes using segmentation techniques.
2. Perform A/B tests of applicable feature variants, independently of user classes.
3. Look for significant variation between user classes given their designated variation, as measured by some performance measure.

If significant variation exists for a feature's variations, we can consistently select the highest-scoring feature variation for the members of the various user classes. The reasoning behind this leap can be found in section 3.4.

The A/B testing regime proposed in this work has the advantage of not requiring any sort of reasoning about the actual users in each clusters. By all means, this may well be interesting from a business intelligence perspective, but will not be necessary to be able to effectively roll out an optimized feature set for each user class.

This project describes a system capable of not only identifying user classes based on user behavior, but more importantly: it describes a framework for identifying the most effective ways of adapting the product to these user classes, in effect driving users in a desirable direction.

Although the project implementation will specifically target the video conferencing service appear.in, a major research question will be to what extent the results generalize.

### 1.4.1 Research questions

With the context outlined above, we formulate the following three research questions:

1. Is it possible to clearly divide users of simple, anonymous web applications into user classes based solely on their behavior?
2. To what extent can such an application be adapted to better suit their way of using it?
3. When identity is unreliable, is the effort worth it?

It is clear that our ability to answer each subsequent research question depends on the answer to the one before it. Consequently, the thesis is organized in a similarly stepwise manner.

## 1.5 Organization of the thesis

This paper is organized as follows.

Chapter 2 surveys similar research, similar applications, and provides an in-depth analysis of the available data. Chapter 3 describes the system design and the reasoning behind central design choices, choices of algorithms et cetera. In chapter 4 we'll analyze the discovered user classes to help answer the first research question, and see how they differ in their adoption of two central application features. Chapter 5 summarizes the most important takeaways, and suggests further work.

# Chapter 2

# Survey

This chapter will survey first the application case in section 2.1, then move on to each major research field that will be touched upon through the thesis.

## 2.1 The application case

Although the case in question is a specific service, the techniques in use and the limitations needing to be dealt with should apply to many kinds of applications.

Firstly, let's pick apart the workings of appear.in to thoroughly understand what kind of an application we are dealing with. This should provide a good basis for understanding the generated data and the applicability of the results for the general case.

### 2.1.1 The inner workings of appear.in

As illustrated in figure 2.1, the appear.in architecture is quite simple. It is built on a simple peer-to-peer (P2P) architecture, with signaling done through a centralized server endpoint.

The instrumentation service has been included, simply to illustrate the fact that the data we have available is generated and sent directly from the clients.

By "signaling" in the context of appear.in, we mean everything not directly related to the media streams between the peers. This includes:

- Managing which peers are in which room.
- Setting up new peer connections when a client joins a room.

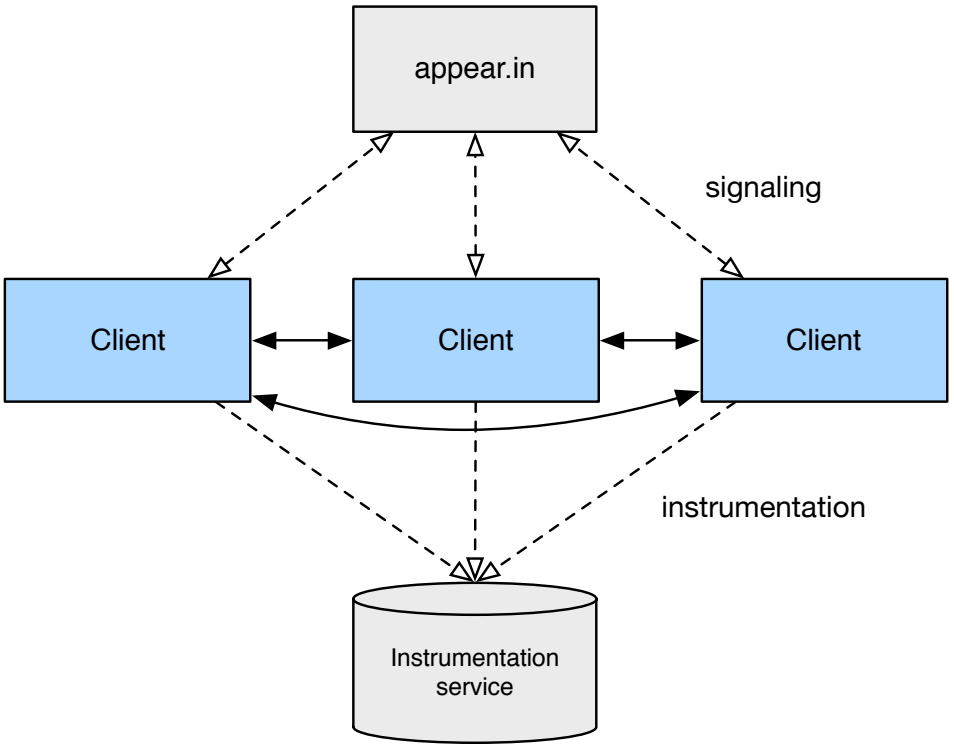



FIGURE 2.1: The appear.in architecture, illustrated for a conversation between 3 peers. The black arrows indicate media data flow, and the dashed arrows indicate metadata flow: signaling data and instrumentation data, respectively.

- Tearing down peer connections when a client exits a room.
- Distributing various metadata that needs to be in sync across peers.

The user interface is also quite simple. It consists of a landing page (a screenshot of the current version can be seen in figure 2.2), and a "room page" (see figure 2.3). For the sake of simplicity, let's go through them separately.

#### 2.1.1.1 The landing page

The landing page's objective, as for most landing pages, is two-fold: to *evoke interest*, and to *activate the user*. Although we cannot directly measure them, we can indirectly measure the degree of interest and the activation rate in two ways:

1. The ratio of users going from the landing page to a room (interest).
2. The ratio of users going from the landing page to a room to a conversation (activation).

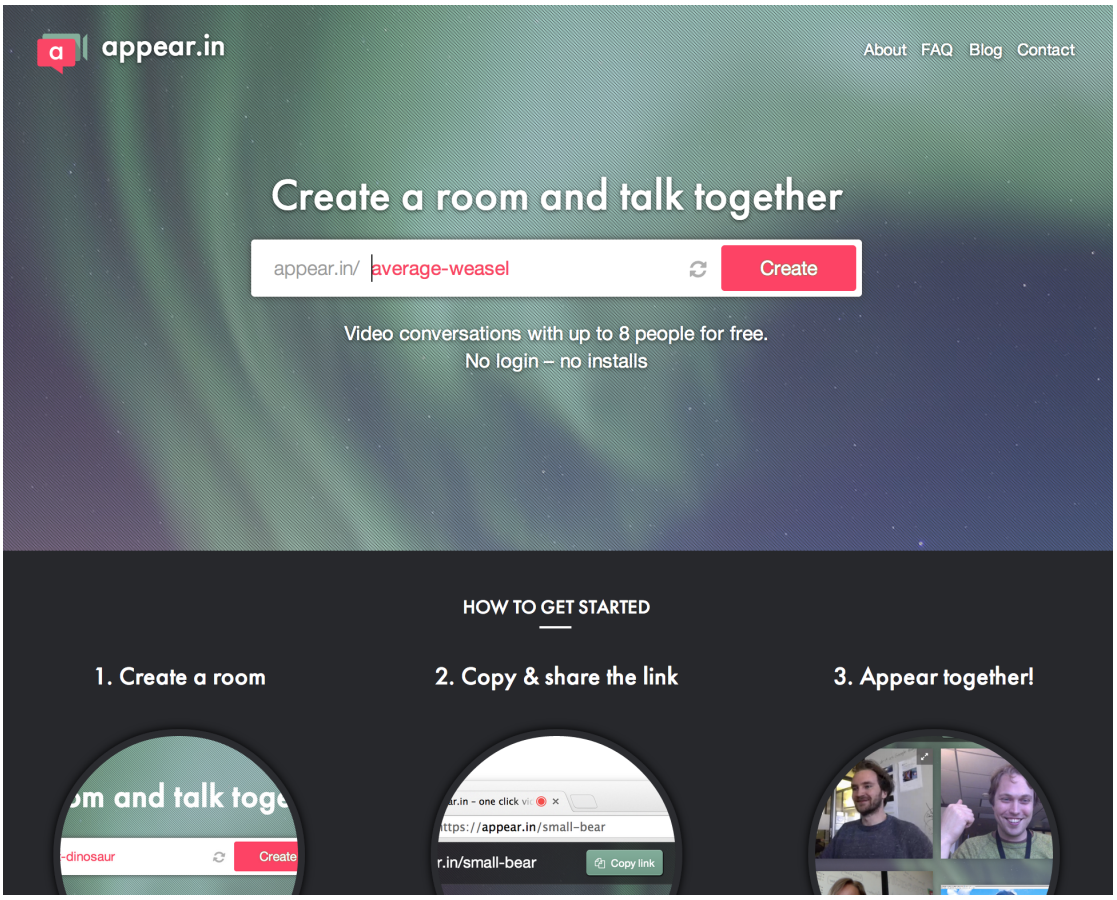



FIGURE 2.2: The appear.in landing page (as of June 2014).

The concept of evoking interest and of activating the user are universal terms that should generalize well to many other web applications.

#### 2.1.1.2 The room page

The room page, the actual product user interface, is composed of several parts. Each participant resides in his or her own video control, and various room controls are placed along the top and bottom parts of the page.

As the quality of the video conferencing part of the application is largely governed by the browser and other low-level technicalities, we will mostly focus our efforts on the functionality augmenting the content: effectively, the rest of the UI.

*Please note that all features discussed in this section are continuously subject to heavy revision, and should not in any way be viewed as a permanent or final set of features.*

The leftmost part of the top bar consists of a URL copying control, as shown in figure 2.4a. Many users utilize this area when copying the page URL to invite their peers

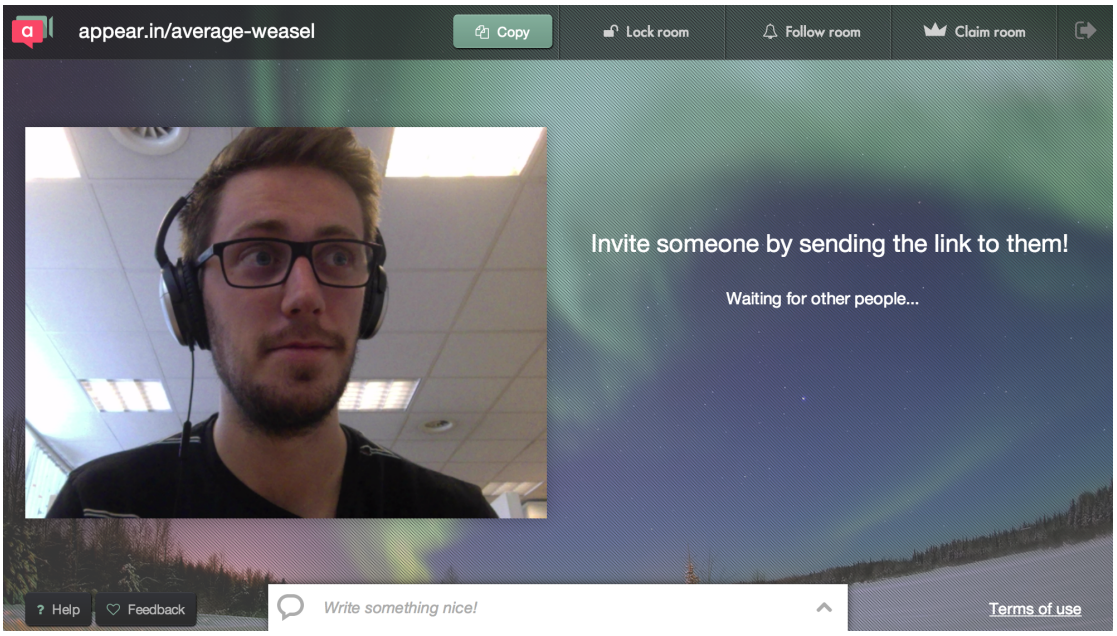



FIGURE 2.3: The author in an appear.in room (as of June 2014).

to the room. However, seeing as the same effect is easily achieved by copying the address field of the browser – which we cannot track – use of this control does not give a complete picture of users' sharing behavior.

To the top right is a row of buttons, as shown in figure 2.4b. Respectively, they allow the user to "lock", "follow", "claim", and leave the room. Of these, only the first three are of particular interest, as the "leave room" button essentially does nothing but close the window, severing the connection.

All these buttons alter the state of the room. Let's briefly walk through them.

**Lock**

When locking a room, one prevents other people who stumble upon the room's URL from entering.[1]

**Follow**

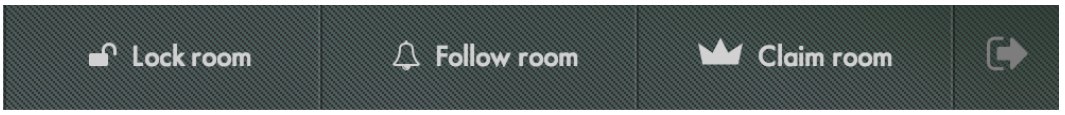
Users following a room are notified when other people enter it. A room's followers may also follow the room chat without being present in the room, by using a browser extension.

**Claim**

Users can claim a previously unclaimed room, and essentially take ownership of it. This enables them to customize the room in a number of ways.

[1] They are, however, able to knock their way in, much in the same fashion as one would enter a locked room in the physical world.


(A) The room URL copying control.




(B) The top button row. Each button serves its own purpose and fires its own event.


(C) The chat control.

FIGURE 2.4: The most important UI parts.

The last piece of the feature puzzle is the chat control, depicted in figure 2.4c. When users post a message, it becomes visible to other members. Chat messages are written to a centralized store, and persist as long as there are people in the room.

### 2.1.2 Generality of the application case

To improve our ability to generalize the functionality described in the previous section, we will focus on the application features augmenting the main video conferencing functionality. These functions essentially either *manipulate the room resource* or aid in *exporting information* about it. Thus, the analyzed feature set should apply equally well to other applications where these are central user tasks.

However, as touched upon in section 1.2, the *main* challenge of appear.in as an adaptation use-case is 1) the absence of demographics, and 2) the unreliability of user identity. The next sections look at how well these aspects generalize.

#### 2.1.2.1 The absence of demographics

A key facet of all user adaptation and personalization is adapting to a user's interests, and so it is imperative to learn about the user. Montgomery and Srinivasan introduce a distinction between active and passive learning to aid in categorizing approaches to the issue [7]. Whereas active learning results from direct questions to the user, passive learning is the opposite: learning about the user without asking.

Active learning has several disadvantages in the general case:

1. It requires too much effort on the customer's part.
2. The user may indeed not know the answer to the questions, either lacking the proper knowledge or experience to evaluate the alternatives.
3. The user may be unwilling to reveal correct answers.
4. It is inefficient, as it typically ignores information consumers reveal about their preferences in their past interactions and purchases.

The first point applies especially to the case of appear.in. An important part of the product is the simplicity of the application, ie. the small amount of friction. Thus, introducing questionnaires or similar approaches to collecting active feedback from users has not been viewed as a positive tradeoff up to this point.

So, we will be limited to passive learning. What does this leave us with in practice?

Montgomery identifies three major sources of information from which it is possible to learn passively: transaction data, clickstream data, and email.

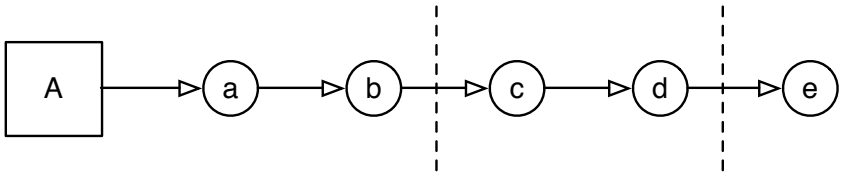
appear.in, being a so-called single-page web application (SPA), has no clickstream in the traditional sense, a traditional clickstream being the series of pages navigated to. Transactional data, on the other hand, is usually related to e-commerce, and to some sense of "items bought". This is similarly irrelevant in this particular interpretation. However, we *can* choose to view the series of interactions with the application as a clickstream of sorts – an event stream – and use it as a data source in much the same way.

#### 2.1.2.2 The unreliability of user identity

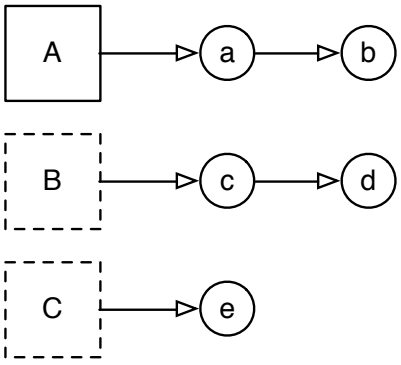
Being an anonymous service, appear.in has no explicit login. As explained in section 1.2.4.2, appear.in uses cookies to track users over time, which introduces some challenges when it comes to building user models.

As we shall see, most users periodically clear their browser cookies (see 2.2.5 for numbers), which causes an abrupt end of the perceived event stream from the browser user, never to be reassumed. The effect is illustrated in figure 2.5 for a hypothetical case, in which user $A$ clears the browser cache twice, effectively cutting off the event stream each time. There is no obvious way in which to consolidate the three event streams of the three perceived users $A$, $B$, and $C$.

In practice, this leads to sparse usage data and quite a bit of noise in the event stream used as the basis for the user model generation. Furthermore, the shortening of event




(A) An event flow from $a \rightarrow e$, as seen from a user $A$. The dashed vertical lines indicate a point at which the user clears the browser cache.




(B) The same event flow as perceived by the system.

FIGURE 2.5: The impact that clearing the browser cookies has on the chronological event stream for a single user $A$.

streams also introduces a bias to the clustering algorithms, as the user model vectors will be shorter than is actually the case. We will return to how this issue affects our approach in section 3.4.

These are all issues that apply especially to appear.in, but that may also apply to many other anonymous systems.

## 2.2 The field of user adaptation

This section will present an overview of the field of user adaptation. From the rather short history of the field, we move on to the conceptual frameworks and methodologies that modern adaptive systems base themselves on, before surveying the state of the art applications.

### 2.2.1 A brief history of adaptive systems

Vrieze discerns the history of adaptive systems into three eras [6]: early research, the pre-Internet era, and the Internet era. His historical dissertation is loosely based on Kobsa [8]. This summary will do the same, but the main focus of attention will be on

what Vrieze calls *the Internet era*, which is when user modeling and the Internet began to join forces.

#### 2.2.1.1 Early research

The first work within user modeling research was conducted in the nineteen-eighties. Strongly influenced by the field of artificial intelligence, the groundwork for modern user modeling was laid.

The common denominator for the user modeling systems composed in this era was their tight integration with their respective production systems. The end of the era, however, saw systems such as GUMS [9], the first standalone user modeling system. These early standalone systems are most widely known as User Modeling Shell Systems, a term coined by Kobsa in 1990 [10].

#### 2.2.1.2 Pre-Internet research

The nineteen-nineties was a decade widely dominated by User Modeling Shell Systems, which carried on the tradition of GUMS from 1989. Mostly, stereotype-based approaches were used, which sought to deduce logical connections between user models and the application domain.

One system, however, Doppelgänger [11], stands out in this regard, being the only system to employ a probabilistic model, and not a logic-based approach [8, 12, 13].

*@TODO: Architecture underlining the "pre-Internet" name?*

#### 2.2.1.3 Internet-time research

@TODO: What are recent developments and key enablers?

### 2.2.2 Personalization on the web

When the field of user adaptation meets the web, we have traditionally tended to dub it either "web personalization" or "adaptive hypermedia". The terms are defined as follows:

**Web personalization**

Any action that adapts the information or services provided by a Web site to the

needs of a particular user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site [14].

**Adaptive hypermedia**

A hypertext or hypermedia system, with a user model, able to adapt the hypermedia using the model [15].

The first of these definitions assumes the website to be a hierarchy of content. However, large parts of the modern web have indeed shifted towards the application domain, as discussed in section 1.1.1. The traditional concept of a website being some kind of structured set of "content pages" is completely outdated. In this light, the definition of web personalization is a bit dated.

Adaptive hypermedia, on the other hand, is a wide and general enough term to cover our application case, and much research within its scope applies well to our application case.

We will be using the terms "user adaptation" and "personalization" without relating it to the domain in which it is visible to the user, as this is quite irrelevant where a clear separation of the application and adaptational component is in place. However, as the "hypermedia" term implies, we base our adaptations on interactions with the interface.

### 2.2.3 Designing adaptive graphical user interfaces

@TODO: Rephrase intro below. "As user modeling work often manifests itself as user interfaces, naturally..."

There are those who approach the issue of adaptive interfaces from a slightly different angle, and many of them naturally come from the design world.

There has been extensive research in this area, which differs from the more user model oriented angle, as described above, in two important ways:

1. They tend to focus on ways of applying concrete interface adaptations, and their respective effects on actual users.
2. The conclusions are often of qualitative nature, basing themselves on user testing.

Results from this kind of research tends to measure an adaptation's success not in terms of transactions and ROI, but in terms of user satisfaction and the quality of the user experience [16, 17].

As we shall see, the system described in this thesis is indeed inspired by this approach, particularly in its use of feature experiments to power user adaptations. We will come back to this point in section 3.5.

### 2.2.4 Implementation methodology

As mentioned, there are many approaches to the task of user adaptation. Due to the constraints outlined in the previous chapters, we will be honing in on the approach termed "web usage mining", or sometimes "clickstream analysis".

Traditionally, this has entailed tracking the way users traverse a website hierarchy, looking for path patterns among them, and using this information to adapt the pages in various ways [14, 18, 19]. While not entirely the same task, this is mostly analogous to the type of adaptation problem we are solving, and the systems proposed in this earlier work solves many of the same problems that this work will need to solve.

### 2.2.5 Privacy versus personalization

The marketing firm Jupiter defines personalization as "predictive analysis of consumer data used to adapt targeted media, advertising and merchandising to consumer needs. According to Jupiter, personalization can be viewed as a cycle of recurring processes consisting of *data collection*, *profiling* and *matching*" [20]. This section will discuss some ethical and political issues surrounding the first step of this cycle, namely *data collection*.

Teltzrow and Kobsa [2] state it plainly: "Personalization systems need to acquire a certain amount of data about users' interests, behavior, demographics and actions before they can start adapting to them." As we shall see, many Internet users are highly sceptical of providing personal information to web sites, and a majority are concerned about web sites tracking their movements and behavior online. This doesn't fit adaptive systems' demand for data collection.

#### 2.2.5.1 Personal information

First, consider the following survey results regarding personal infomation [2]:

1. Internet Users who are concerned about the security of personal information: 83% [21], 70% [22], 84% [23]

2. People who have refused to give (personal) information to a web site: 82% [24]
3. Internet users who would never provide personal information to a web site: 27% [23]
4. Internet users who supplied false or fictitious information to a web site when asked to register: 34% [24], 24% [23]

Although the above numbers aren't directly relevant to the case of appear.in, where no personal information is collected or stored, it underlines a general scepticism towards providing information to web sites.

The fact that such a large portion of users are sceptical of providing personal data tells us that there is reason to believe there is room for anonymous niches within most application areas, serving as a drive towards more applications like appear.in.

#### 2.2.5.2 Tracking

The same scepticism as above is again seen when users are surveyed on their attitudes towards being tracked online [2].

1. People who are concerned about being tracked on the Internet: 60% [21], 54% [23], 63% [25]
2. Internet users who generally accept cookies: 62% [26]
3. Internet users who set their computers to reject cookies: 25% [24], 3% [21], 31% in warning modus [21], 10%[23]
4. Internet users who delete cookies periodically: 52% [26]

As discussed in section 2.1.2.2, any time a user clears the browser cookies, we effectively break the user's event stream, and is from that moment onwards – for all we know – a brand new user.

This makes it hard to know if these numbers are similar to those among the users of appear.in, as they have not yet been surveyed on this subject. Furthermore, our only data source touching the user base is based on the tracking cookies in question – hence we are unable to examine to what extent the above survey applies to our case.

#### 2.2.5.3 The road ahead

Although these numbers are a bit dated, there is little reason to believe that the tracking situation is going to get any easier in the years to come [27–30].

As we have seen, there already exists broad scepticism towards the use of cookies to track users, even on first-party sites. Much of the problem seems to stem from sites allowing third-party cookies, to better serve advertisements to its users – who most often are oblivious to the tracking taking place at all.

These third-party cookies, however, allow these third-parties to track users' activity and behavior across multiple sites, often without their explicit consent. This has recently been deemed as bordering to surveillance, and in recent years extensive legislative restrictions have been introduced to decrease the prevalence of particularly third-party cookies.

This all presents a challenge for services like appear.in, who use third-party cookies not to advertise, but to enhance the service for the user. There is reason to believe this situation will not become easier to deal with in the years to come, but hopefully, there will come about better ways of dealing with the issue.

## 2.3 Clustering techniques

There are many approaches to the task of clustering user models.

@TODO: Some text on hierarchical, flat, density, probabilistic, k-median etc, before presenting k-means as our choice due to: good match with both input and desired output, scales well (parallel optimization), by far most popular in commercial applications [31].

### 2.3.1 Clustering evaluation

In general, any clustering method should search for clusters whose members are close to each other and well separated. Berry and Linoff [32] formulate it in terms of *compactness* and *separation*.

**Compactness**

The members of each cluster should be as close to each other as possible.

**Separation**

The clusters themselves should be widely spaced.

For an algorithm such as $k$-means, which takes the $k$ as an input, the central question when it comes to cluster validity is for which value of $k$ the cluster compactness and separation are optimal.

The most efficient way of measuring cluster quality between several executions of a clustering method, where only its parameters – like $k$ – differ, is to utilize a so-called *relative criteria* [33].

One of the most widely used measures of clusters' relative criteria is the Davies-Bouldin index, which is the one used to differentiate between clusters in this project. It is defined as:

$$\mathrm{DB}_k = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left( \frac{s_i + s_j}{d(c_i, c_j)} \right) \tag{2.1}$$

where $k$ is the number of clusters, $c_x$ is the centroid of cluster $x$, $s_x$ is the average distance of all elements in cluster $x$ to centroid $c_x$, and $d(c_i, c_j)$ is the distance between centroids $c_i$ and $c_j$.

In plainer words, the Davies-Bouldin index formulates a measure of the quality of cluster separation and compactness, while considering the number of clusters in a responsible manner. Thus, it is well suited to distinguish between runs of eg. the $k$-means algorithm, where the value of $k$ differs from run to run.

## 2.4 A/B testing

Controlled experiments embody the best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior [34, 35].

The simplest form of controlled experiment is often referred to as the A/B test. In A/B tests users are randomly exposed to one of two variants: control (A), or treatment (B). Based on data collected, an Overall Evaluation Criterion (OEC) is derived for each variant. Figure 2.6 illustrates the A/B testing process.

One word in the previous paragraph, "randomly", warrants some discussion. To be able to establish a causal relationship between the selected variant and the evaluation, the variant selection must indeed be completely random, and not based on what Kohavi et al. term "any old which way" [34].

This point *must* be kept in mind when designing adaptive systems based on controlled experiments, as in our case.

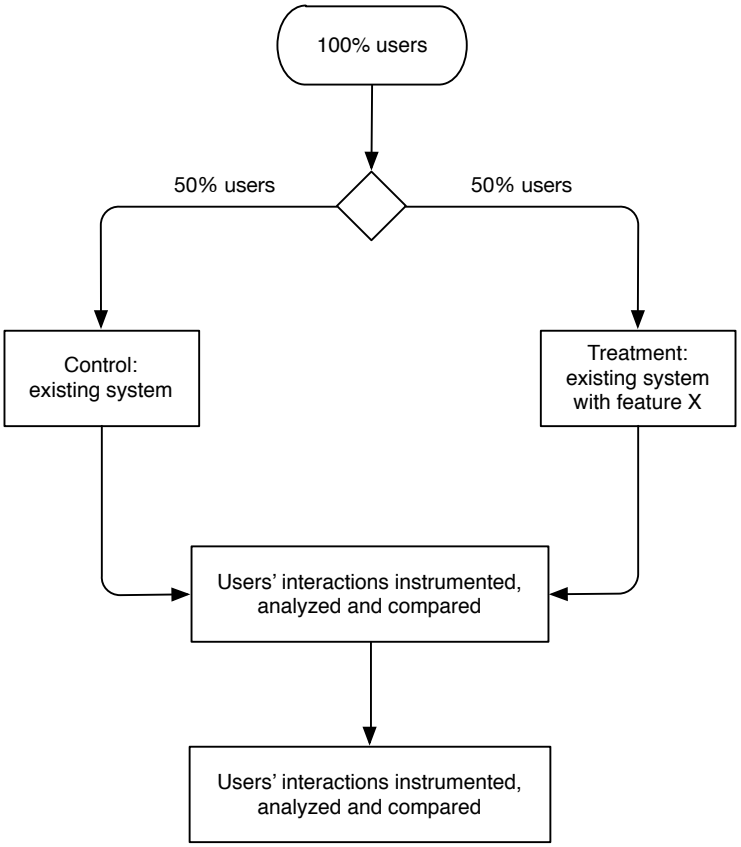



FIGURE 2.6: The flow of an A/B test.

## 2.5 State of the art

*@TODO: Compile a set from the above sections.*

# Chapter 3

# Approach

## 3.1 System overview

The proposed system is constructed around the data flowing through it. This chapter describes each part of the system outlined in figure 3.1.

The system begins by taking the data stepwise through an ingestion pipeline, importing, filtering, and cleaning it. The steps of the ingestion pipeline are discussed in section 3.2, before moving on to the user modeling component in section 3.3.

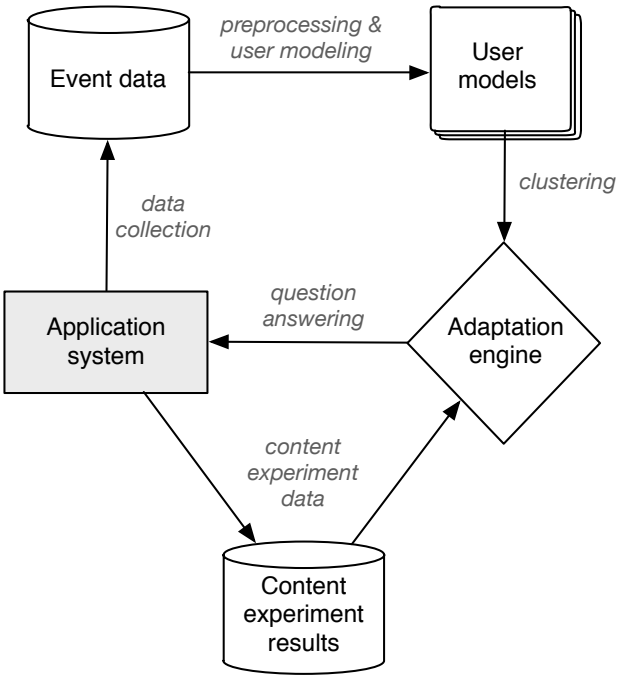



FIGURE 3.1: Overview of the system architecture.

Once user models are in place, the system identifies user segments. These are stored in a database for future use. This process is discussed in more detail in section 3.4.

Apart from clusters, the adaptation component requires content experiment results to predict how users will respond to application variations. This is discussed in section 3.5.

The components come together in the adaptation component, discussed in section 3.6, whose job it is to answer questions about users, effectively completing the feedback cycle back to the application.

## 3.2 Data ingestion and preprocessing

Before the interesting parts of the system can start doing their work, the data needs to be transformed from *a series of chronological raw events* to *a set of user models*.

The amount of data can be arbitrarily sizable, and will grow linearly with user activity. The system architecture has been designed to be able to cope with this; its functional and data-driven nature should be easily adaptable to hugely scalable programming paradigms like MapReduce.

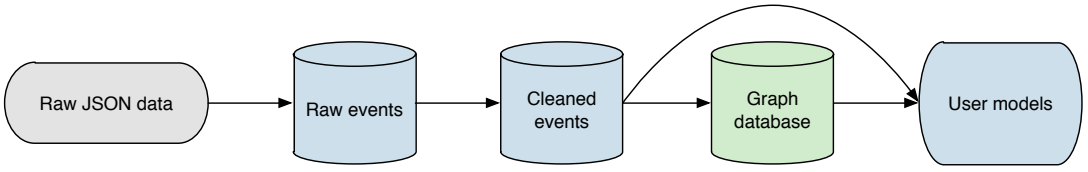



FIGURE 3.2: The ingestion pipeline broken into 4 steps. The color of each node indicates means of storage: *Blue* indicates a RDBMS, *green* indicates a graph database, whereas *gray* is used to indicate flat file storage.

### 3.2.1 Generating the raw data

The system input is a chronological series of raw events sent from the production system. The events are instrumented via an external analysis service called KISSmetrics[1] – a user analysis system designed around tracking individual users' behavior.

The application logs an event by calling the KISSmetrics REST API. The following data is in each instrumentation call:

1. person identifier

[1] https://www.kissmetrics.com/

2. event name
3. user properties (optional)

The consistency of the personal identifier has already been discussed extensively in the introductory chapters, especially section 1.2.4, but a short technical introduction to the actual production system is in order.

To enable effective utilization of the KISSmetrics instrumentation functionality, they supply a client library for the purpose. This client library handles a few central things for us:

1. Person identity storage and loading over subsequent page loads.
2. The low-level instrumentation of events.
3. Simple A/B testing facilities.

When the KISSmetrics client library is loaded, the person identity is automatically either retrieved from the browser cookies, or generated. The identity of a person is a unique randomly generated string, which serves no other purpose than to track the identity of the browser over time.

Whenever something "interesting" happens, an event is sent to the KISSmetrics instrumentation service. An "interesting" event is typically anything that tells us about how the users use the service, both in terms of general activity and in terms of feature adoption. Every event is tagged with the person identity, as well as an event name and a timestamp.

The KISSmetrics service provides several analytical tools to dig into this data, thereamongst funnel reports (example in figure 3.3) and cohort reports.

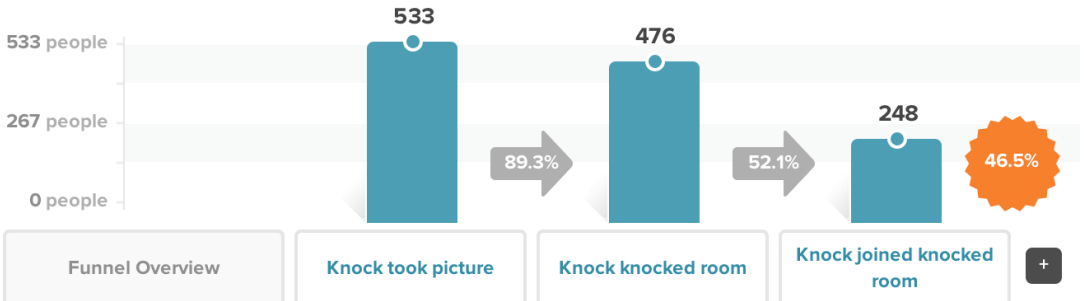



FIGURE 3.3: Example of a simple funnel report.

### 3.2.2 Cleaning the data

KISSmetrics has the ability to export raw event data to data files. This can be used to power completely customized analyses, as will be needed for our particular task.

After the raw data has been aquired, it will need to be cleaned. This is a simple process of churning through each line of each unprocessed data file, parsing and splitting its contents into appropriate data fields, and inserting it into databases.[2]

## 3.3 User modeling

To find clusters of users, we first and foremost need to quantify them. More specifically, we want to represent each user as a numerical feature vector.

### 3.3.1 Feature selection

Given the types of events being collected, we landed on compiling the following features:

1. First degree conversation partners
2. Second degree conversation partners
3. Inviter
4. Invitee
5. Conversations
6. Rooms used
7. Rooms claimed
8. Roomnames generated
9. Chat message sent

To generate user models, the stream of event data were mapped to their associated users for aggregation. That is, the transformation in this step started with data in the form of (3.1).

[2]To facilitate the compilation of network-related user model features, conversation data was loaded into a graph database to enable querying of network structures.

$$\langle \text{event}, \text{timestamp}, \text{person}, \text{metadata} \rangle \tag{3.1}$$

And ended with data in the form of (3.2), where *value* contains the aggregated value.

$$\langle \text{person}, \text{feature}, \text{value} \rangle \tag{3.2}$$

## 3.4 Clustering the users

We assume the following hypothesis holds, as a basis for the clustering approach to the user adaptation problem:

*Hypothesis* 1. The more similarly two people use a service, the more likely they are to respond similarly to it changing.

Thus, we want to segment the users in the following way: users in each segment should be as similar as possible, and as dissimilar those in other segments as possible. This scheme fits well with the two criterias for selecting an optimal clustering scheme, as described by Berry and Linoff [36].

@TODO: Remember the bias introduced in section 2.1.2.

### 3.4.1 Choice of algorithms

Three algorithms were implemented and experimented with: DBSCAN, mean-shift, and k-means. Of these, the k-means clearly proved itself as the most effective one, and as it also managed to produce adequate and meaningful results, it was chosen as the principal algorithm.

#### 3.4.1.1 The k-means clustering algorithm

The k-means clustering algorithm takes as input a preset number of clusters, $k$, a similarity measure function, and a set of data vectors. Initially, it randomly chooses $k$ data vectors as centroids as a starting point for the process, before repeatedly performing a two-pass operation adjusting the centroids until convergence.

Section B.1 shows a simple python-esque implementation of the k-means algorithm.

To select good parameters – the optimal value for $k$, given the input vectors – we will use a relative cluster validation index, like the Davies-Bouldin index discussed in section 2.3.1.

## 3.5 Feature experiments

Ideally at this point, we have identified several significant clusters of users. Next we want to find viable ways of adapting the product to better suit each user.

Given a set of candidate product alterations $A$, we want to give each user the combination of these that maximizes some performance measure $P$. However, we have no predictive bias to start us off on solving this.

The approach taken in this implementation is a simple one, which relies on conducting a single A/B test up front for each proposed product alteration. An important part of this phase is to log each selected variant with the *same person identifier* as is used for other event instrumentation, as discussed in section 3.2.1.

### 3.5.1 Evaluation metrics

When considering which variation of a product feature to prefer, we need to be able to measure their relative success rates.

As for any user-facing product, the main objective is achieving user happiness. However, user happiness is hard to objectively define. For a service like appear.in, though, activity level serves as an adequate indicator of user happiness, as unhappy users have more than enough alternative applications that could cater to their needs.

Simply put, we basically assume that the following hypothesis holds:

*Hypothesis* 2. Happy users use the service more than unhappy users.

Furthermore, we seldom need a measure of *user happiness* to determine the relative success of variations of a product feature. Some parts of the product have clearly defined goals themselves.

The perhaps most obvious example of this is the landing page, whose main objective is to get people to try out the product. We can define the performance of the landing page in terms of the percentage of users that continue on to try out the product.

## 3.6 Adaptation component

In the adaptation component, we want to select the variation which is most likely to provide the best experience for the user.

In other words, given a user $u$ in cluster $c$ and a set of variations $V$, we choose the variation that is more likely to maximize our selected objective function. That is,

$$\text{preferred}(c, V) = \underset{v \in V}{\text{argmax}}\, \text{score}(c, v) \tag{3.3}$$

The scoring function will in our case simply be the ratio of cluster $c$ members who were "converted" during the experiment, ie. who achieved the designated goal, after having been presented with variation $v$.

$$\text{score}_{c,v} = \frac{\text{converted}_{c,v}}{\text{participants}_{c,v}} \tag{3.4}$$

Given a recent set of clusters and experiment results for the users within them, the adaptation component has what it needs to select a preferred variation based on this heuristic.

### 3.6.1 Applying the personalized feature set

Since well before this project, appear.in had a way of toggling features based on external criteria. The existing model allowed for overriding various settings, as well as pre-releasing features internally, using URL parameters.

We call this scheme "URL-based feature switching". Some examples are listed in table 3.1.

| Parameter setting | Effect |
|---|---|
| `?video=off` | Turn off video by default. |
| `?audio=off` | Mute microphone by default. |
| `?followRoom` | Turn on experimental "follow room" functionality. |

TABLE 3.1: By appending feature switches to the room URL, various effects can be achieved.

Enabling user adaptation thus became a simple matter of extending this model by allowing for feature values to be dictated by our adaptation component.

In the current codebase, the setting of each feature value happens like so:

```
var features = {
    // ...
    isVideoDisabledByDefault: $routeParams.video === "off",
    isAudioDisabledByDefault: $routeParams.audio === "off"
};
```

Here, `$routeParams` contains the parameters set in the URL, as described above. In the same way, we can perform an external call with the current user id to the adaptation component described above, and apply the desired feature set in a similar way.

# Chapter 4

# Evaluation

This chapter will evaluate a prototype implementation of the approach presented in chapter 3.

The chapter starts with introducing the main goal and the experimental setup in sections 4.1 and 4.2, before evaluating the actual results in sections 4.3, 4.4, and 4.5. We round up this evaluation chapter with a summary and discussion, in sections 4.7 and 4.8.

## 4.1 Prerequisites

For any of this user adaptation to have any actual use, we will need to find out whether there actually exists significant variation in feature adoption between clusters; it matters little whether users exhibit differing behavior with regard to existing functionality, if this behavior yields no basis for predicting feature adoption in the future. We need an inductive bias.

As described in chapter 3, a way of discovering whether this inductive bias is present in the user base is to investigate whether there is significant variance across the user groups in their users' adoption of new features. Thus, this question of whether the predictive bias exists will be included as a central part of the actual experiment itself, and be thoroughly discussed through the next sections.

## 4.2 Experimental setup

@TODO: Describe figure 4.1 before jumping to hypothesis.

The main hypothesis is:

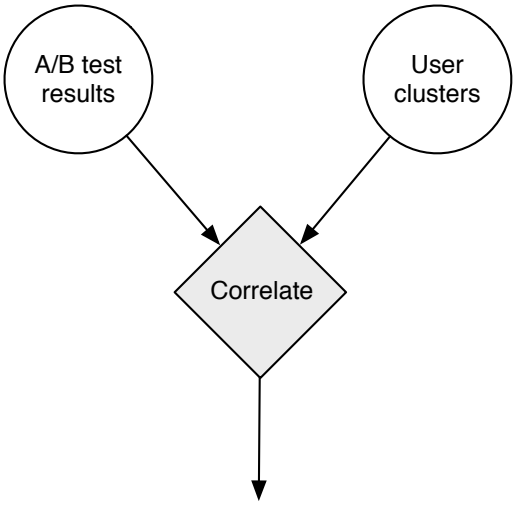



FIGURE 4.1: The experiment setup. The output of the correlation component (in gray) will determine the outcome of the experiment.

> Different variations of a service may perform differently among "different kinds of users". These variations can be used as inductive bias on which general user adaptations can be based.

Two sources of input will serve as input data to test the hypothesis:

1. A/B test results
2. User clusters

The data from the A/B test results describe each person's designated variant, and the resulting performance. See table 4.1 for some example data.

| Person | Variant | Entered room | Entered conversation |
|---|---|---|---|
| p1 | A | no | no |
| p2 | A | yes | yes |
| p3 | B | yes | no |
| p4 | A | no | no |
| p5 | B | yes | yes |

TABLE 4.1: Each person has one assigned variant, and some performance measures.

The user cluster data contain the same persons as the test results in table 4.1 and the clusters they have been assigned to, as illustrated in table 4.2.

These data sources are aggregated with regard to variant and cluster, and the relative success rates compared.

| Person | Cluster |
|---|---|
| p1 | 1 |
| p2 | 2 |
| p3 | 1 |
| p4 | 1 |
| p5 | 2 |

TABLE 4.2: Persons have been assigned to to clusters.

The next sections walk through and evaluate the actual results in the context of the experiment described in this section.

### 4.2.1 Evaluation case: New appear.in landing page

To test the hypothesis, we performed a controlled experiment on the user base of appear.in as we rolled out a new landing page.

The experiment stood between the old landing page, the control treatment, and a new design. We shall from here on refer to them as variation $A$ and $B$, respectively. The two competing variations are depicted in figure 4.2.

## 4.3 A/B testing features

After running through the test period with an even randomized user split between the two variations, the results were in. In total, just over 20,000 people were included in the experiment and were served one of the two variations.

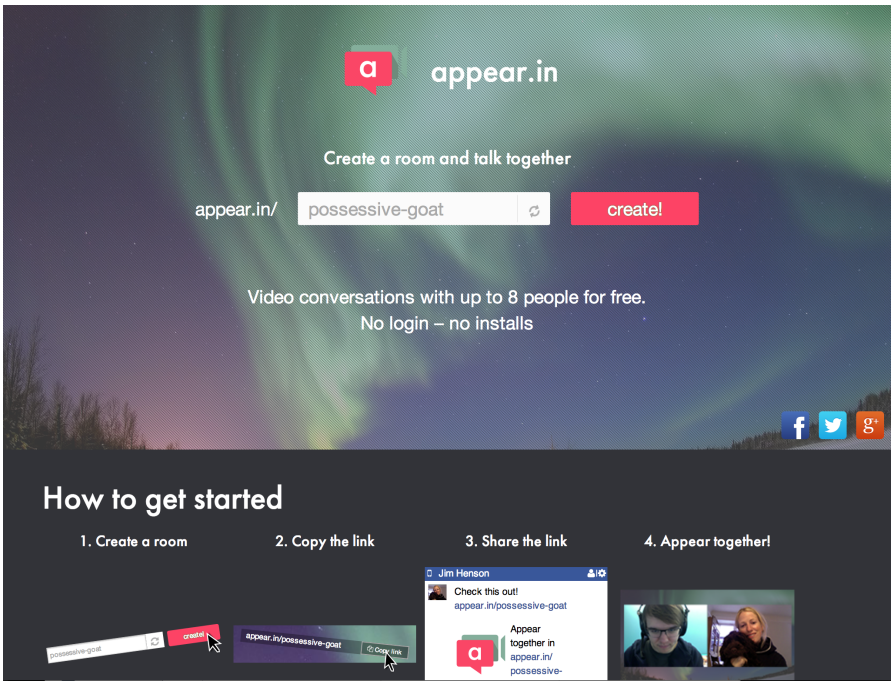
The two were compared head to head for two different performance metrics:

**Visited room**
: The number of users going from the landing page and into a room.

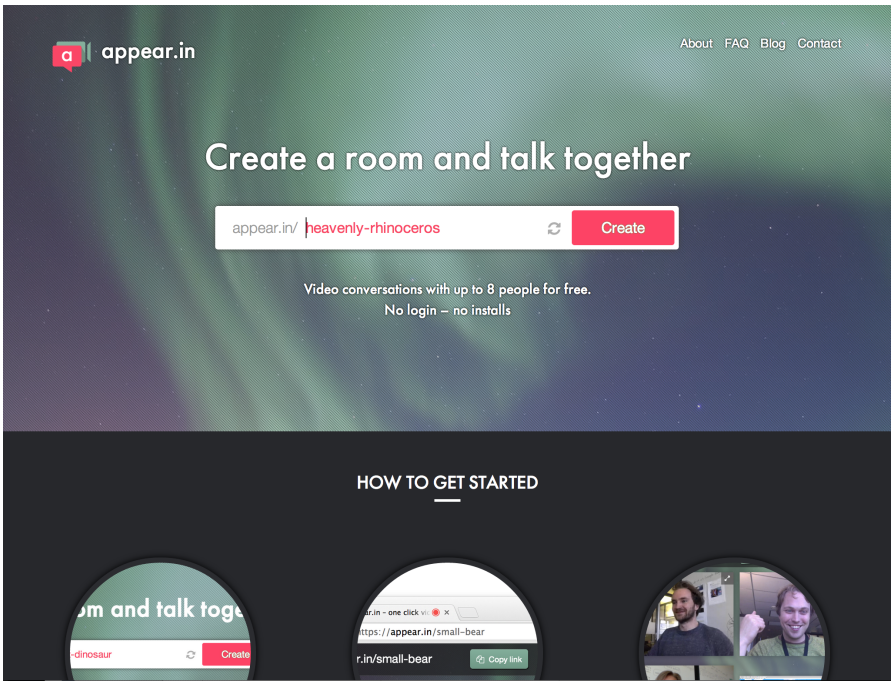
**In a conversation**
: The number of users going from the landing page and into a conversation. This entails that the person in question has understood the concept of how one establishes a conversation, which can be seen as the next step of user activation after entering a room.

The aggregate numbers are shown in tables 4.3 and 4.4.

Figures 4.3a and 4.3b show how the variations performed with respect to the two metrics, both with variation $B$ plotted relative to variation $A$ as the baseline.




(A) Variation *A*.




(B) Variation *B*.

FIGURE 4.2: Variations in the new landing page experiment.

As we can plainly see, variation *B* performs slightly better at moving people into rooms, but evidently fails to communicate the conversation concept, leading to a relative decrease in actual subsequent conversations.

While these numbers and plots set the scene, they are not very interesting from our

| Variation | People | Conversions | Average conversion | Improvement | Certainty |
|---|---|---|---|---|---|
| *A* | 10,100 | 7,946 | 78.67% | - | - |
| *B* | 9,931 | 7,883 | 79.38% | 0.90% | 89.04% |

TABLE 4.3: Comparison of "Visited room" conversion ratios for variations *A* and *B*.

| Variation | People | Conversions | Average conversion | Improvement | Certainty |
|---|---|---|---|---|---|
| *A* | 10,009 | 3,602 | 35.99% | - | - |
| *B* | 10,172 | 3,781 | 37.17% | -3.29% | 95.98% |

TABLE 4.4: Comparison of "In a conversation" conversion ratios for variations *A* and *B*.

adaptation perspective. We are first and foremost interested in seeing whether these numbers vary significantly between segments of the user base, thus providing us with a predictive bias from which we can extrapolate our user adaptations.

Section 4.5 breaks the numbers down cluster by cluster, and investigates to what degree this information can be used to adapt the interface.

## 4.4 User clustering

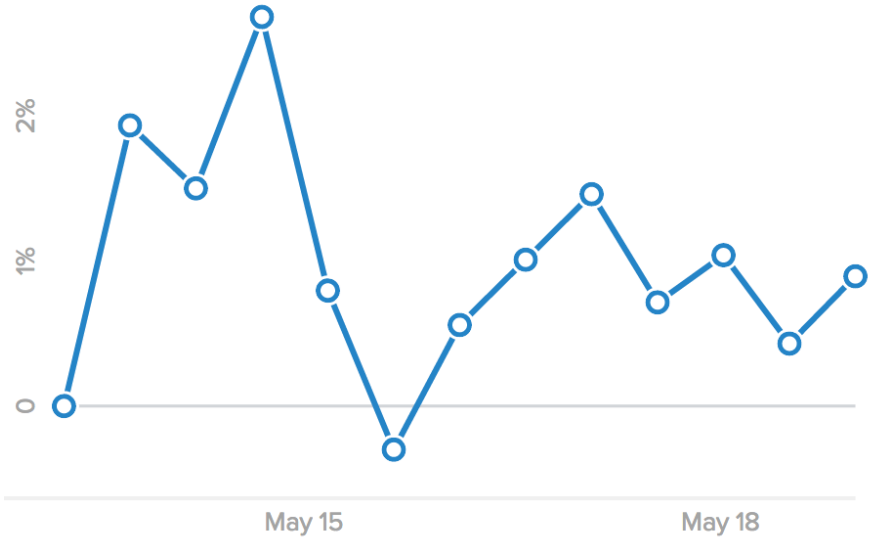
The clustering approach was iteratively improved on over several months. Large amounts of new data was gathered through the developments process, and with it the parameters were tuned and adapted.

Figure 4.4 shows a set of clustering runs. Data was timeboxed per month, from January thru May, and k-means was run for a range of $k$ from 3 up to 10. The other parameters used, as well as the particular resulting cluster set from February, are discussed in further detail in section 4.4.3.

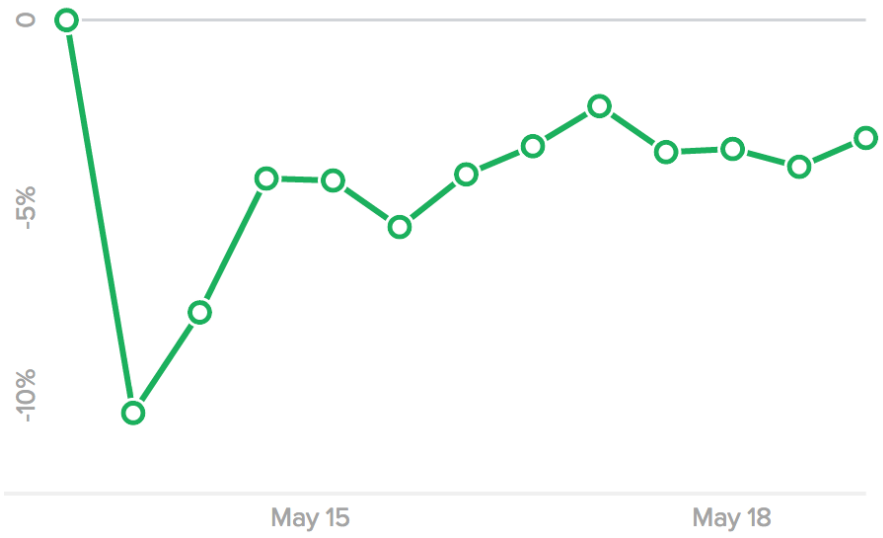
### 4.4.1 Varying demographics

Through the spring of 2014, appear.in received quite a bit of media attention, and different times saw user influx from a large variety of demographic origins.

The coverage varied from technical showcases and industry newsletters to the service being featured in Hungarian newspapers and on BBC World News. Moreover, these spurs of media attention seems to mostly have been independently initiated, and as a result we saw large peaks of visitors from quite specific locations at different times. Overall, though, the traffic was distributed quite evenly over a large number of geographical areas.




(A) Comparison of the variations by the "Visited room" metric.




(B) Comparison of the variations by the "In a conversation" metric.

FIGURE 4.3: Variation $B$ performance, relative to variation $A$ (the baseline).

The *timespan* parameter was used extensively to compare these, and aided in avoiding becoming subject to these various demographical biases.

### 4.4.2 Axis normalization

@TODO

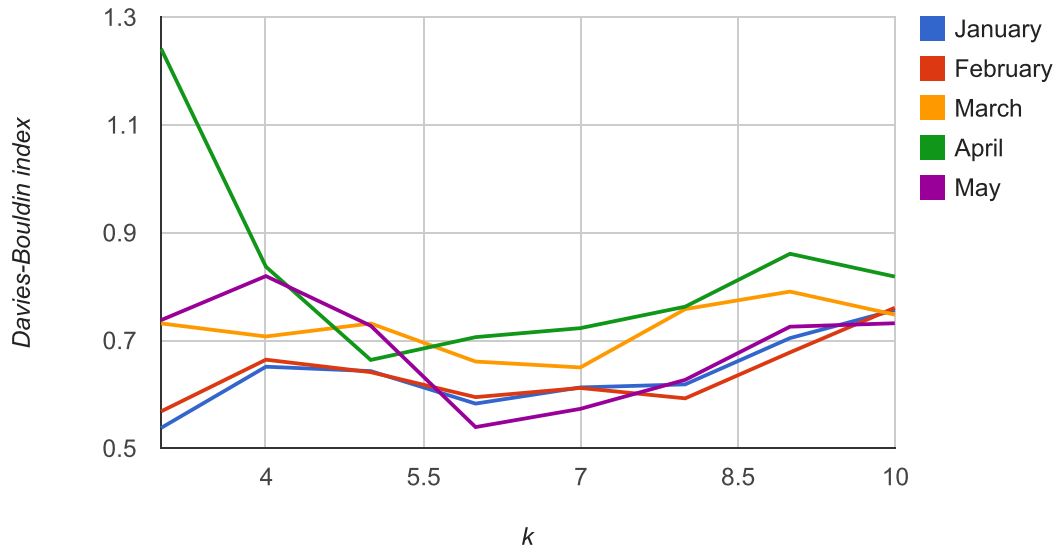



FIGURE 4.4: $k$ vs. Davies-Bouldin index, for clusters with data for January to May.

### 4.4.3 Initial clustering results

The data used in this analysis stems from February 2014.

The following clustering results were found by choosing the best of 5 k-means runs, "best" being defined by their Davies-Bouldin indices (see section 2.3.1 for a brief description of this evaluation metric). This process was performed for $k$ parameter values from 3 to 10, where $k = 8$ yielded the best result. The 2 smallest resulting clusters were omitted in the analysis due to their relatively insignificant sizes.

Although irrelevant to the experiment, a detailed analysis of the clusters depicted in figure 4.5 is available in A.1.

Data from January and March yield more or less the same results, although they are a bit less clear. This could be due to media events and holidays generating more skewed data than usual.

The radar chart in figure 4.5 shows the centroids of 6 large clusters $C$ relative to each other.

Each dimension's centroid values $\mu_i \in \mu$ have been scaled by a factor of $\frac{10}{\max_{c \in C} c_i}$, to fit nicely inside the chart.

The features used in this particular clustering run are (clockwise around the chart):

1. chat messages sent

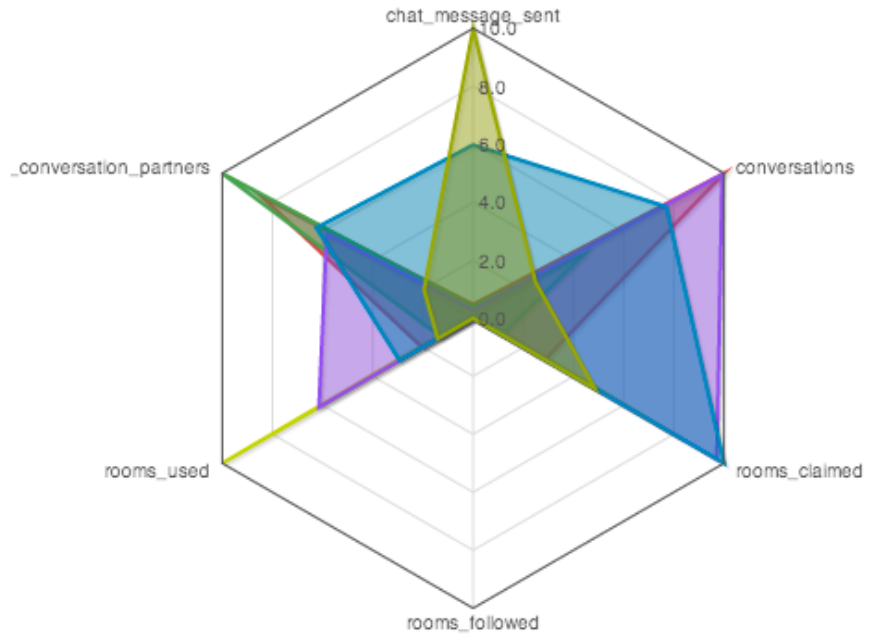



FIGURE 4.5: Radar chart comparing clusters generated from data for February 2014.

2. conversations (2+ persons present in room)
3. rooms claimed
4. rooms followed
5. unique rooms used
6. conversation network size (ie. number of unique other users within 2 degrees of conversation separation)

Most of these features are quantified by counting the number of relevant events logged for each user.

### 4.4.4 The applicability of manual cluster analysis

When analyzing a set of clusters, it is tempting to focus on the extremes – the easily stereotyped user profiles – as we have done above. However, we see that a vast majority of users are indeed too close to the origo to be confidently labelled as some kind of stereotype.

Thus, although tempting from a business intelligence type of perspective, manual cluster stereotyping becomes a highly speculative exercise whose conclusions are extremely hard to confirm without demographic data readily available.

We will return to this issue in chapter 5.

## 4.5 Adapting the application

The task of adapting the application brings together previous A/B test results and the clusters discovered, as described in sections 4.3 and 4.4.

| | off | | | on | | | total | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $C$ | n | c | % | n | c | % | n | c | % | preferred |
| 1 | 42 | 0 | 0.00 | 43 | 0 | 0.00 | 85 | 0 | 0.00 | off |
| 2 | 21 | 9 | 42.86 | 16 | 11 | 68.75 | 37 | 20 | 54.05 | on |
| 3 | 156 | 75 | 48.08 | 152 | 67 | 44.08 | 308 | 142 | 46.10 | off |
| 4 | 151 | 90 | 59.60 | 160 | 113 | 70.62 | 311 | 203 | 65.27 | on |
| 5 | 118 | 106 | 89.83 | 133 | 125 | 93.98 | 251 | 231 | 92.03 | on |
| 6 | 180 | 145 | 80.56 | 165 | 134 | 81.21 | 345 | 279 | 80.87 | on |

TABLE 4.5: Success of the different variations for each cluster $C$.

Table 4.5 shows the success of the different variations for each cluster. We see that there are indeed clear differences between them, and can therefore reasonably state that there exists an inductive bias on which predictions about future behavior can be based.

However, before jumping to conclusions, let us take a look at the number of people included in the experiment and the number of people identified as in the given clusters.

To illustrate the point of the next section, the clusters in table 4.5 were identified based on data from just *before* the A/B experiment was run. Cluster 1, in the first row, contains the same types of users as the ones described in A.1.1, namely users only hitting the front page.

As we can plainly see, the majority of the clustered users did not return to the service within the experiment time period, resulting in an extremely low relative participation rate. The other clusters however, comprised of more active users, see a much higher participation rate.

## 4.6 Identity persistence

It may not be enough to be able to predict what the current user base wants. What we *really* want is to use this information to improve the service for returning users. However, it is not enough to have predictive bias if we're unable to recognize our returning users. This section analyzes the impact of cookie impermanence.

As has been discussed extensively already, a major difficulty in adapting applications such as appear.in to its users is that there is no concrete notion of *a user*. Users are

anonymous, and the only way they are being tracked at all is through a random hash set in a tracking cookie (see section 2.1.2.2 and 2.2.5).

How, though, does this affect our adaptation efforts? One way of gauging this is to look at the number of users we see returning. More precisely: for how many months do we see each tracking cookie return before we never see it again?

Table 4.6 shows the decline over time of the number of returning users, as compared to the user base of January, February and March.

| Users from month | 1 month | 2 months | 3 months | 4 months | 5 months |
|---|---|---|---|---|---|
| January | 23.20% | 5.10% | 2.90% | 1.90% | 0.90% |
| February | 20.20% | 3.80% | 2.10% | 1.10% | - |
| March | 23.20% | 5.40% | 2.90% | - | - |

TABLE 4.6: The degree to which a monthly unique set of users are seen again after $n$ months.

Two plots of these numbers can be seen in figure 4.6, with normal scale and logarithmic scale.

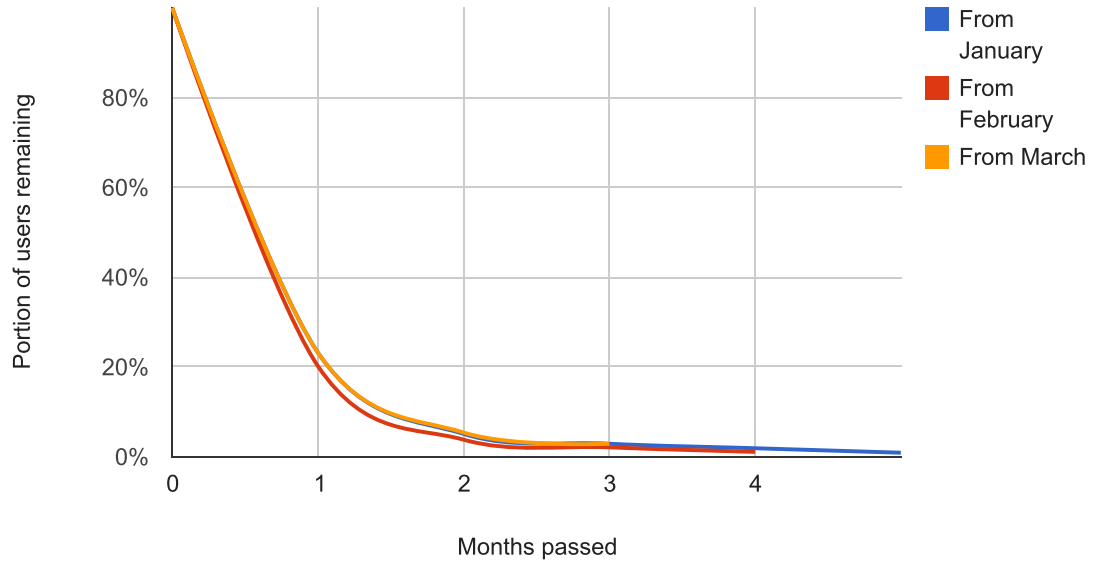
As we see, just over 20% are seen again the month after they first visit the site. Only about 20% of these users again make it through to the subsequent month. After three months, less than 3% remain in all three cases.

The perceived rate of retention is subject to three different factors:

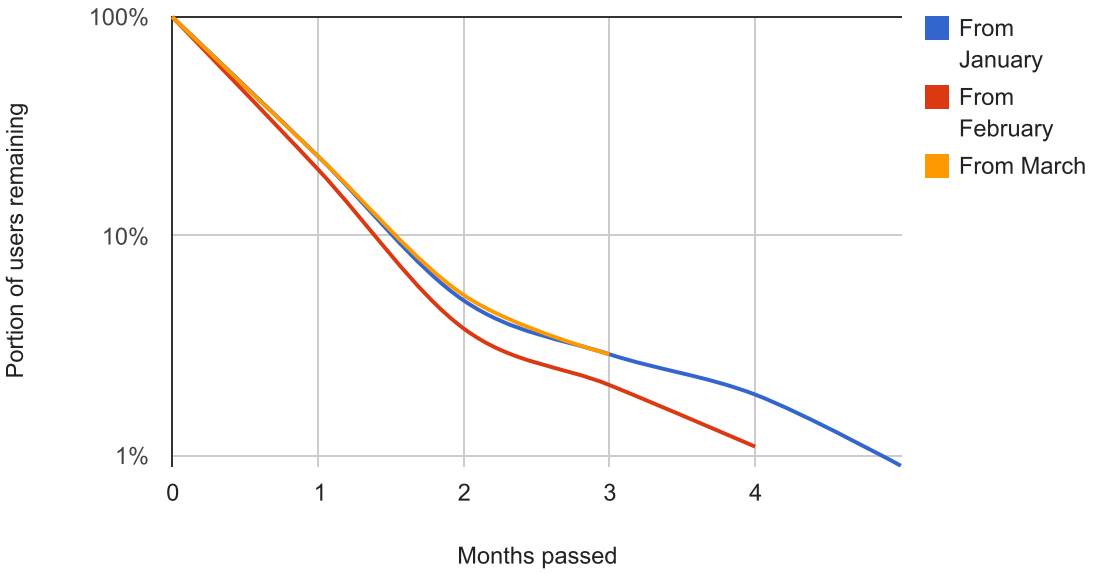
1. The user did not use the service again.
2. The user cleared the browser cookies.
3. The user changed browser or computer.

There is all reason to believe that we are looking at a combination between all three, yet hard to know how much each contributes to the trend. We naturally assume that most users perceived as not returning simply do not return.

However, the survey performed by Teltzrow and Kobsa [2], as discussed in section 2.2.5, states that more than half of all users clear their cookies periodically, and a significant amount of users reject cookies altogether. We will therefore assume that the last two factors also contribute in a significant degree to these numbers.




(A) Perceived user persistence.




(B) Perceived user persistence, logarithmic scale.

FIGURE 4.6: Perceived user persistence, illustrating the identified user dropoff rate.

## 4.7 Summary

## 4.8 Discussion

# Chapter 5

# Conclusion

## 5.1 Summary

As shown in the previous chapter, we find that there are indeed clear differences in feature adoption across the identified user segments.

The domain constraints and the resulting scarcity and unreliability of user modeling data makes for clustering results that are at best hard to intuitively understand. This hinders us in reasoning about the various ways of applying the user adaptations, ie. the ways of distributing the feature variants among the identified user segments.

The controlled experiments of section 3.5 provide us with hard evidence of behavioral correlations in our user base – or indeed, lack thereof. This disentangles us from the error-prone decision making otherwise involved in setting up group adaptations.

## 5.2 Generalizing the system

Section 2.1.2 analyzes the application case, and identifies the two main challenges of providing user adaptation:

1. the absence of demographics
2. the unreliability of user identity

Together, these two factors comprise an unusually harsh environment for an adaptative system to operate in. In ... we predicted that the clusters would be hard to intuitively

understand, as they would effectively represent a noisy and incomplete version of the actual user base.

As we saw in chapter 4, these predictions did indeed match the results.

Due to uncertainty caused by domain constraints, it is hard to tell to what extent the results generalize.

## 5.3 Suggestions for further work

The work presented can be improved in many ways.

Within the domain of appear.in, and applications like it, the event data cleaning and user modeling processes have a lot to gain.

Firstly, it would be interesting to see if similar results are achieved for other applications.

Keep cluster centroids to enable "classification" of new users into groups whose behavior is already known.

# Appendix A

# Evaluation Results

## A.1 Clustering results

This section contains a comprehensive qualitative description of the 6 largest clusters identified among user models generated compiled from February data.

### A.1.1 Cluster 1: Front page hits

The centroid for this cluster is quite simply $\mu_1 = \langle 0, 0, 0, 0, 0, 0 \rangle$.

*Persona* 1. The user has not tried the actual service – most likely hitting the front page and either not using a compatible device/browser, or not finding it interesting enough to try out.

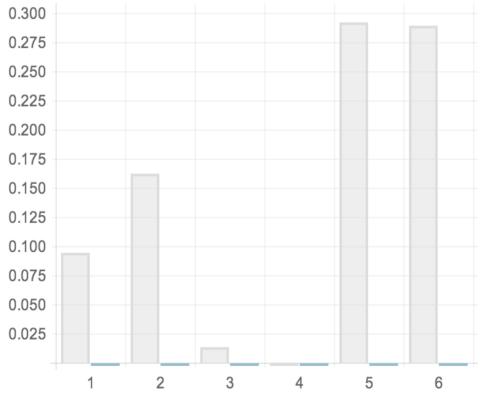
### A.1.2 Cluster 2: Trying out the service alone

As shown in figure A.1b, the users in this cluster score close to 0 on every feature except the number of rooms used – most notably, the number of conversations.

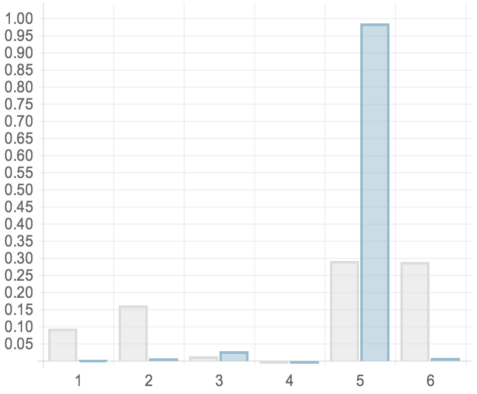
*Persona* 2. The user has tried out the service, but not ever conversed with another user.

### A.1.3 Cluster 3: Simple users with small networks

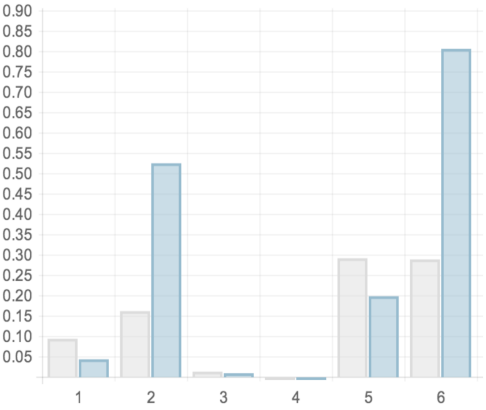
Users in cluster 3 (see figure A.1c) don't use any of the more advanced features of the service, like chatting, claiming or following a room, but on average they have taken part in just over 5 conversations.



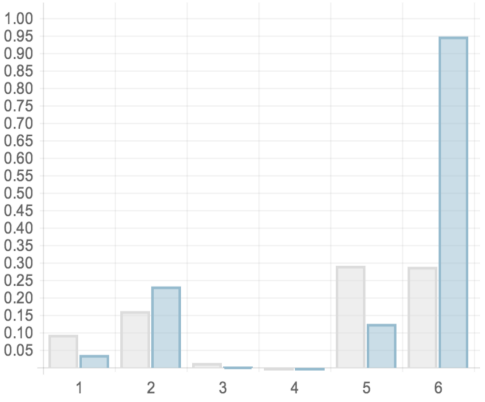

(A) Cluster 1: Front page hitting users



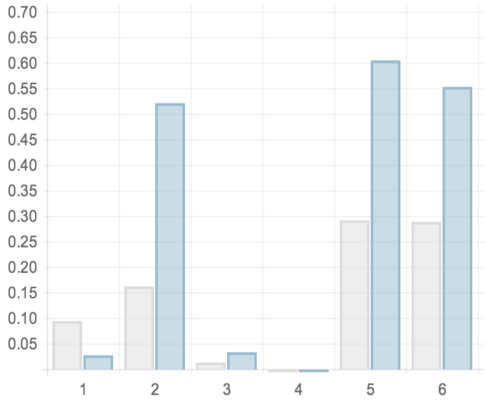

(B) Cluster 2: Users trying out the service alone



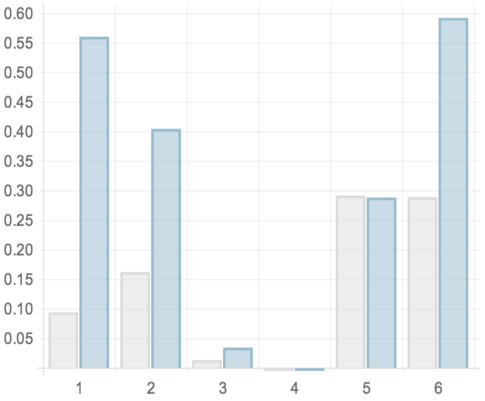

(C) Cluster 3: Simple users with small networks




(D) Cluster 4: Simple users with large networks




(E) Cluster 5: Incognito users




(F) Cluster 6: Chatty, simple users

FIGURE A.1: Cluster centers compared to the unweighted average cluster centers. The features plotted are 1) chat messages sent, 2) conversations, 3) rooms claimed, 4) rooms followed, 5) unique rooms used, and 6) conversation network size.

*Persona* 3. A returning user, only utilizing the bare functionality, conversing only with a very limited group of people.

### A.1.4 Cluster 4: Simple users with large networks

The users in cluster 4 (see figure A.1d) are very much like the ones in cluster 3, but use the service slightly more, and are part of much larger networks.

*Persona* 4. A returning user, only utilizing the bare functionality, part of a large group of people using the service.

### A.1.5 Cluster 5: Incognito users

To track users over time, cookies are required. Whenever clearing the browser cache, or when browsing in "incognito mode"[1], the service will not be able to tie together user sessions. These perceived one-off users should end up in this cluster, close to the pattern shown in figure A.1e.

*Persona* 5. A user browsing in incognito mode, or who clears the browser cache regularly.

### A.1.6 Cluster 6: Chatty simple users

The users of cluster 6 are very much like those in cluster 3, as can be seen in figure A.1f. The significant difference is that they make heavy use of the chat functionality.

*Persona* 6. A user sharing the characteristics of users in cluster 3, except in making use of the text chat functionality.

[1] Browsing without storing any data, including cookies.

# Appendix B

# Source Code

## B.1 The k-means clustering algorithm

```
def kmeans(K, distance_fn, vectors):
    N = len(vectors)
    cluster_members = {}
    memberships = {}

    # Initially set centroids to random data vectors
    centroids = [vectors[randint(N)] for _ in xrange(K)]

    while True:

        # Assign each data vector to the closest cluster centroid
        for vector in vectors:
            k = argmin(K, lambda k: distance_fn(centroid[k], vector))
            if not memberships[vector] == k:
                cluster_members[memberships[vector]].remove(vector)
                cluster_members[k].append(vector)
                memberships[vector] = k

        # Set each centroid to the mean of its members
        previous_centroids = centroids
        for k in xrange(K):
            centroids[k] = mean_vector(cluster_members[k])
```

```
        # Stop computing if we've achieved convergence
        change = 0.0
        for k in xrange(K):
            change += distance_fn(previous_centroids[k], centroids[k])
        if change < CONVERGENCE_THRESHOLD:
            break
```

# Bibliography


[1] W3C. Html5, a vocabulary and associated apis for html and xhtml, introduction, 2014. URL http://www.w3.org/TR/html5/introduction.html.

[2] Maximilian Teltzrow and Alfred Kobsa. Impacts of user privacy preferences on personalized systems. *. . . personalized user experiences in eCommerce*, 5(0308277), 2004. URL http://link.springer.com/chapter/10.1007/1-4020-2148-8_17.

[3] Alfred Kobsa. Privacy-Enhanced Web Personalization. *Communications of the ACM*, 50:628–670, 2007.

[4] Alfred Kobsa and Jörg Schreck. Privacy through pseudonymity in user-adaptive systems. *ACM Transactions on Internet Technology*, 3(2):149–183, May 2003. ISSN 15335399. doi: 10.1145/767193.767196. URL http://portal.acm.org/citation.cfm?doid=767193.767196.

[5] Edsger W Dijkstra. *Selected writings on computing: a personal perspective*. Springer-Verlag New York, Inc., 1982.

[6] PT De Vrieze. *Fundaments of adaptive personalisation*. Paul de Vrieze, 2006. ISBN 9789090211138. URL http://books.google.com/books?hl=en&lr=&id=9wyclzI91McC&oi=fnd&pg=PA1&dq=Fundaments+of+Adaptive+Personalisation&ots=qDOGKmb6hw&sig=b51fHFbRmGXV4nJ1EjkH3-XrvHI.

[7] Alan Montgomery and Kannan Srinivasan. Learning About Customers Without Asking. *Tepper School of Business*, (324), 2002.

[8] Alfred Kobsa. Generic User Modeling Systems. pages 49–63, 2001.

[9] TimothyW. Finin. Gums — a general user modeling shell. In Alfred Kobsa and Wolfgang Wahlster, editors, *User Models in Dialog Systems*, Symbolic Computation, pages 411–430. Springer Berlin Heidelberg, 1989. ISBN 978-3-642-83232-1. doi: 10.1007/978-3-642-83230-7_15. URL http://dx.doi.org/10.1007/978-3-642-83230-7_15.

[10] Alfred Kobsa. Modeling the user's conceptual knowledge in bgp-ms, a user modeling shell system1. *Computational Intelligence*, 6(4):193–208, 1990. ISSN 1467-8640. doi: 10.1111/j.1467-8640.1990.tb00295.x. URL http://dx.doi.org/10.1111/j.1467-8640.1990.tb00295.x.

[11] Jon Orwant. Heterogeneous learning in the Doppelgänger user modeling system. *User Modeling and User-Adapted Interaction*, 4(2):107–130, 1995. ISSN 0924-1868. doi: 10.1007/BF01099429. URL http://link.springer.com/10.1007/BF01099429.

[12] Wolfgang Pohl. Labour - machine learning for user modeling. In *Design of Computing Systems: Social and Ergonomic Considerations (Proceedings of the Seventh International Conference on Human-Computer Interaction), volume B*, pages 27–30. Elsevier Science, 1997.

[13] Wolfgang Pohl. Logic-Based Representation and Reasoning for User Modeling Shell Systems. *User Modeling and User-Adapted Interaction*, 9:217–282, 1999.

[14] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1):1–27, February 2003. ISSN 15335399. doi: 10.1145/643477.643478. URL http://portal.acm.org/citation.cfm?doid=643477.643478.

[15] Peter Brusilovsky. Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3):87–129, 1996. ISSN 0924-1868. doi: 10.1007/BF00143964. URL http://dx.doi.org/10.1007/BF00143964.

[16] Krzysztof Z. Gajos, Mary Czerwinski, Desney S. Tan, and Daniel S. Weld. Exploring the design space for adaptive graphical user interfaces. *Proceedings of the working conference on Advanced visual interfaces - AVI '06*, page 201, 2006. doi: 10.1145/1133265.1133306. URL http://portal.acm.org/citation.cfm?doid=1133265.1133306.

[17] Leah Findlater and Joanna McGrenere. Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, page 1247, 2008. doi: 10.1145/1357054.1357249. URL http://portal.acm.org/citation.cfm?doid=1357054.1357249.

[18] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic Personalization Based on Web Usage Mining. 43(8), 2000.

[19] Alan L. Montgomery and Michael D. Smith. Prospects for Personalization on the Internet. *Journal of Interactive Marketing*, 23(2):130–137, May 2009.

ISSN 10949968. doi: 10.1016/j.intmar.2009.02.001. URL http://linkinghub.elsevier.com/retrieve/pii/S1094996809000322.

[20] C Foster. The personalization chain. *Jupiter Communications, Site Operations*, 3, 2000.

[21] Cyber Dialogue. Cyber dialogue survey data reveals lost revenue for retailers due to widespread consumer privacy concerns. *New York, Cyber Dialogue, November*, 7:2001, 2001.

[22] L Behrens. Privacy and security: The hidden growth strategy. *Gartner G2 Report*, 2001.

[23] S Fox, L Rainie, J Horrigan, A Lenhart, T Spooner, and C Carter. Trust and privacy online: Why americans want to rewrite the rules. pew internet & american life project, washington. *A Pew survey of*, 2, 2000.

[24] Mary J Culnan. The culnan-milne survey on consumers & online privacy notices: Summary of responses. In *Interagency Public Workshop: Get Noticed: Effective Financial Privacy Notices*, pages 47–54, 2001.

[25] Harris Interactive. A survey of consumer privacy attitudes and behaviors. *Rochester, NY*, 2000.

[26] Personalization Consortium et al. Personalization & privacy survey. *Los Angeles*, 2000.

[27] A. Ruiz-Martínez. A survey on solutions and main free tools for privacy enhancing Web communications. *Journal of Network and Computer Applications*, 2012.

[28] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. *2013 IEEE Symposium on Security and Privacy*, pages 541–555, May 2013. doi: 10.1109/SP.2013.43. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6547132.

[29] Ove Sorensen and Christian-albrechts-universitat Kiel. Zombie-Cookies : Case Studies and Mitigation. pages 321–326, 2013.

[30] Van Eijk, Van Der Plas, and Van Der Sloot. Online tracking: Questioning the power of informed consent. 2011.

[31] P. Berkhin. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas, and Marc Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-28348-5. doi: 10.1007/3-540-28349-8_2. URL http://dx.doi.org/10.1007/3-540-28349-8_2.

[32] Michael JA Berry and Gordon Linoff. Data mining techniques for marketing, sales and customer support. john willey & sons. *Inc., 1997, 454 P*, 1996.

[33] Maria Halkidi. On Clustering Validation Techniques. pages 107–145, 2001.

[34] Ron Kohavi, Randal M Henne, and Dan Sommerfield. Practical Guide to Controlled Experiments on the Web : Listen to Your Customers not to the HiPPO. 2007:1–9, 2007.

[35] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, July 2008. ISSN 1384-5810. doi: 10.1007/s10618-008-0114-1. URL http://link.springer.com/10.1007/s10618-008-0114-1.

[36] Michael J Berry and Gordon S Linoff. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, 1997.