Data-Caps and Rationing

Master Thesis
by

Ferdinand Gantert
1453992

At the Department of Economics and Management
Institute of Information Systems and Marketing (IISM)
Information & Market Engineering

Reviewer: Prof. Dr. Christof Weinhardt
Advisor: Jun.-Prof. Dr. Lukas Wiewiorra

14th of September 2014

Contents

List of Figures

List of Tables

Glossary

BNetzA .... Bundesnetzagentur, Federal Network Agency for Electricity, Gas, Telecommunications, Post and Railway, German federal regulatory body for all kind of networks, including telecommunication networks.

DTAG .... Deutsche Telekom AG, former government run monopolist for telecommunication services in Germany, still incumbent in almost all areas of the country.

FCC .... Federal Communication Commission, United States of America

HCI .... Human computer interaction, a research area in computer science examining the optimal interaction between humans and computer systems.

ISP .... Internet service provider.

QoE .... Quality of Experience, QoE is less technical than QoS. While QoS has some measurable characteristics that might be ratio scaled in most cases, QoE can only be measured by comparison and is therefore ordinal scaled. QoE can be important as there are human perceptions of quality that are hard to be measured by numbers.

QoS .... Quality of Service is the more technical term compared to QoE. Therefore it does not describe the user's perception of the service characteristics but some measurable technical characteristics of the service.

VoIP .... Voice over Internet Protocol, a telephone like service realized over Internet infrastructure.

Abstract

Data-cap tariffs have been known for a long time in mobile data communication. A relatively new phenomenon is their use for landline broadband Internet access in developed markets. In the past landline broadband Internet-access in developed markets was dominated by flat-rate tariffication. Flat-rate tariffs have special psychological properties therefore a shift away from them may be especially hard. These effects make flat-rate tariffs very popular among consumers and alternatives hard to establish after the consumer got used to them.

The first part of this thesis illustrates the concept of data-cap tariffs. It first gives an introduction followed by a short examination how flat-rate effects may apply to data-cap tariffs. After that the determinants of data-caps are evaluated. To put the gained insights into perspective and conclude the first part, a short overview of currently used data-cap tariffs is given. Based on that, the most likely properties for all determinants of a data-cap tariff are derived.

The second part builds on the insights of the first and designs an economic model for data-cap tariffs later used to analyze the impact on the consumer. Since data-cap tariffs share many economic properties with rationing, no new model is built but rationing research is reviewed in order to find a suitable model. To do that firstly an introduction to the general economic theory of rationing and how it may apply to data-cap tariffs is given. Following that the rationing models considered are briefly introduced and it is argued why a specific one is chosen. Next, the chosen model is closely presented and modified to the needs of a data-cap tariff. The model is designed as a multi-currency system focusing on the interchange of different currencies. The model is able to determine consumer choice with different restrictions and compute virtual prices for restricted scenarios. Based on this model a trend analysis is conducted. This analysis will be based on a simple service framework and analyzes the most basic trends regarding a consumer who is faced with a switch away from a flat-rate tariff towards a data-cap tariff. The analysis part will conclude with a first attempt to quantify the welfare effects for a consumer resulting from a shift from a flat-rate tariff towards a data-cap tariff.

1. Introduction

1.1. Motivation

The Internet has gone through a development from a network purely for research purposes and communication between researchers to a vital infrastructure for both economy and civil society. Besides these vital functions, Internet infrastructure enables a lot of comfort to a wide variety of tasks. Consequently, access to the Internet is a controversially discussed topic. Politicians, the general public, telecommunication companies, as well as special interest for a "free Internet", and economists all have their own - in many cases opposing - views on the topic. For such reasons access to telecommunication networks is highly regulated in most industrialized countries such as the United States (FCC) and Germany (BNetzA). Such regulations can go as far as an a priori check of tariffication for incumbents. For this reasons any major tariff change is a risk for the telecommunication company, both in regard to legal aspects and their public image.
The flat-rate tariff is the predominant pricing-scheme for access to the Internet today. This kind of pricing has some special psychological properties which make a switch away from it unlikely, even if it would be cheaper for the consumer [LS06]. This is a major problem for ISPs. They argue that flat-rate tariffs encourage wasteful use of Internet-traffic. They further argue that this additional traffic plus the general demand for higher transmission capacities for new services leads to a tremendous demand for network development which is far beyond their means. To resolve that problem they suggest that part of the solution would be to come back to a partly usage based pricing. A tariffication scheme often discussed in this context is the data-cap. But, despite its prominence in the discussion, this term can denote many different tariffication-schemes. The common core to all these tariff schemes is that unrestricted Internet access is limited to a certain amount of transferred data for a given period of time. While such tariffication-schemes have been in use in some parts of the world and for mobile data connections, they are rather new to landline broadband in developed broadband markets, especially in Western Europe. This however will change as the German incumbent for landline broadband-access, DTAG, announced that the standard for their consumer broadband offerings will be data-cap based tariffs in the near future[1].
Recalling that a flat-rate tariff has many special economic properties, a shift away from this type of tariff to another type raises questions. Who will profit from such a change? How does this kind of tariffication alter consumers behavior? And what implications does such a change in consumer behavior have for service providers?

To answer these questions the first step has to be to understand what a data-cap is. To achieve that the first part of this thesis focuses on defining what a data-cap is and of which properties it is made up. Part of this is to have a short assessment how flat-rate effects influence data-cap tariffs and shortly review what kind of data-cap tariffs currently are in use.

[1] They will also still offer flat-rate tariffs but for an increased price.

Based on this knowledge the second part of this thesis will then focus on the economic aspects of such a tariffication-scheme. Therefore the special economic characteristics of data-cap tariffs need to be identified. With that knowledge an existing economic model is modified to fit the needs of data-cap tariffs. This model will then be used to answer the questions about the impact of a data-cap tariff on consumer behavior.

1.2. Research question and methodology

It seems like data-cap tariffs for landline broadband access raise a lot of questions, but how can these questions be handled scientifically in order to get answers?

The first thing is to determine a specific research focus. Otherwise the questions arising from the introduction of data-caps would be to broad to be covered in detail here. As already foreshadowed in the introduction, the main focus in this thesis will be on the effects of the introduction of such a tariff on the consumer. As a result, the strategic variables of the ISP and the service providers of different Internet services[2], such as prices and the size of the data-cap will be taken as given.
As this is a first step towards an economic model, the focus lies on an individual consumer, called a representative agent, rather than the aggregated mass of consumers.
Hence the goal of this thesis is to develop a model to analyze the influence of the introduction of a data-cap tariff on an individual consumer coming from a flat-rate tariff.
Questions that are especially interesting in this context are how the consumption points change and what welfare effects result form such a change. To monitor the change it suffice to observe consumption under a data-cap tariff, so the outcomes of a flat-rate tariff and a data-cap tariff can be compared. This has the advantage that this model generally should be able to compute consumption points under a data-cap tariff.

Since restricted consumption is nothing new to economics, it is not necessary to develop a new model from scratch. It should be sufficient to build on existing research and modify it to fit the needs of data-cap tariffs. An area that seems promising as a starting point for a model is economic rationing research, as a data-cap in the broadest sense is nothing more than a restriction on data consumption. Consequently the search for a model fitting the needs defined in the the first part of this thesis will start with a review of the economic literature on rationing.

[2] The terms Internet service provider, as provider of access to the Internet, and web service provider, as provider of a service using the connection provided by the ISP, are easily mixed up. To prevent that a web service provider is usually referred to simply as web service or service provider.

Part I.

Properties of data-cap Tariffs

2. Introductive classifications

2.1. The terms data-cap, ISP and service provider

The term data-cap implies a limitation of data usage. Sometimes a data-cap will also be called a volume-cap to differentiate it from a cost-cap. While a cost-cap sets an upper limit to the costs of a tariff, a volume-cap sets an upper limit to the amount of data consumable in this tariff.
In this context data-cap means a special kind of two-part tariff which is payed upfront and includes, as first part, a given amount of data (volume), measured in bytes, and, as the second part a policy for what happens after that volume is depleted. The included amount of data is called cap because after this amount of data the connection quality will be decreased to a pre-specified level. Up until the cap is reached there are no restriction to data transmission that could not be present in other tariffs. In other words, prior to the cap a capped tariff behaves like a flat-rate or pay-per-use tariff. However there could be a difference invisible to the consumer at this point, namely QoS, but that will be discussed later.
When talking about capped tariffs, it is usually referred to a cap on data usage, but there could also be a cap on time. So for example today it is still common for cellphone tariffs to include a given amount of talk-time for an upfront payment. This kind of arrangement would also be a capped tariff, specifically an example for a time cap. There are different types of caps, but since this thesis will have a look at Internet economics, whenever a cap is mentioned, it will refer to a caped amount of data in a tariff.

Other terms that will be present throughout the entire thesis are ISP and service provider respectively web service. While the terms Internet service provider and service provider or web service[1] are quite similar they denote quite different things and therefore must not be confused. The ISP provides the connection to the Internet, besides that it is mainly important because it sets the size of the cap. It also sets the price for the Internet access, this however is irrelevant to the model since the ISP is neither consumer nor producer. It is a facilitator connecting the consumer and the web service provider. The service providers are the producers in the model. They offer the consumer different services which increases her utility. Normally the terms "web service" and "service provider" are used interchangeably in this thesis. While the first emphasizes the service itself, the second has a focus on the entity providing this service.

2.2. Mobile and fixed-line communication and tariffication

When talking about data-caps a look at mobile communication tariffs can be useful as data-caps are a rather common tariff property there. Also data-caps seem to be a less

[1] The full term would be web service provider but because this is easily confused with Internet service provider the terms web service or interchangeable service provider are used.

controversial topic there. Reasons for that might be that there never have been flat-rate tariffs for mobile data communication in the mass market or that people might understand that there is only a limited number of frequencies, hence capacity is limited. However, the second argument does not seem to hold up as flat-rate tariffs for all other forms of mobile communication (talk time and SMS) are becoming more and more common. Putting the technical and economic differences between mobile and fixed-line communication aside, it could be worthwhile to summarize some insight obtained from research concerning mobile data-caps which could later be used in the model.

How the flat-rate biases can also be present for caped mobile tariffs was studied by Krämer and Wiewiorra in 2010 [KW10]. Since there will be a detailed section about how flat-rate biases may apply to data-cap tariffs later[2], it is not discussed here.
The economics of cellular networks regarding flat-rate tariffs and caped tariffs were studied by Blackburn et al. in 2013 [BSE$^+$13]. They found caps are a necessity in cellular networks today for two principal reasons. The first being to avoid cross subsidies [BSE$^+$13, page 121] and the second being that consumption of data grows faster than innovations enabling higher data rates [BSE$^+$13, page 121][3]. Despite that they also found that heavy users, while using more capacity as they pay for under a flat-rate tariff, may still add revenue for the provider by attracting consumers who consume less than they pay [BSE$^+$13, page 116]. Although they find caps a necessity out of technical reasons they warn to have a second look at the argument of cross subsidies.

[2] See section 3.2.1.

[3] Both arguments are regularly used by ISPs when lobbying for data-cap tariffs for landline broadband-access.

3. The flat-rate bias and how it may apply to data-caps

3.1. The flat-rate bias

Consumers love flat-rate tariffs. This is even true in many cases where a pay-per-use tariff or some other kind of tariff fitting the needs of the consumer would be cheaper. The reason for this flat-rate bias has been subject of extensive research. A widely accepted explanation for this flat-rate bias was published by Anja Lambrecht and Bernd Skiera in their 2006 article ”Paying too much and being happy about it” [LS06]. In this article they identified four main causes for the flat-rate bias, which will be discussed in the next few paragraphs.

The insurance effect

People purchase insurances to be safe for rather unlikely events. According to Lambrecht and Skiera [LS06, page 213] it is the same with flat-rates. According to their findings people in many cases do not purchase a flat-rate because of their average usage patterns but because of rare peak consumption.
Imagine a middle-aged couple with two kids in college. Under normal circumstances they probably have a rather light Internet consumption and therefore would be better of purchasing a pay-per-use tariff. But during Christmas season and summer break their kids are back home and their Internet usage multiplies. Since they do not know a priori how much they will consume during that time, they want to be insured against high costs. Therefore they pay a premium in order to have an upper limit on costs, hence they purchase the flat-rate tariff.

The taxi meter effect

Riding a taxi in an unknown city can be stressful for several reasons. One of it may be the fact that one does not know the city and therefore the driver can take detours in order to be able to bill more. So the ticking of the taxi meter puts stress on the passenger. This effect is widely known in economical psychology as mental accounting. Mental accounting is the phenomenon that the consumption of a good or service becomes less enjoyable when the consumer is reminded of the costs associated with the consumption. Such effects are of course especially strong with pay-per-use pricing. In contrast to that, Lambrecht and Skiera argue that flat-rate pricing separates usage and payment and therefore eliminates the taxi meter effect and makes Internet usage more enjoyable [LS06, page 213].

The convenience effect

Metered tariffs will result in more transaction costs than flat-rates, for both the ISP and the consumer. The reason for that is that in order to provide a metered tariff the ISP needs to record the exact consumption for billing purposes and the consumer needs to check her bill whether it complies with her actual usage. Since a flat-rate tariff has a fixed uniform price these transaction costs do not occur with flat-rates. Thus Lambrecht and Skiera argue that many consumers value their time high enough that this small convenience is another argument for purchasing a flat-rate [LS06, page 214].

The overestimation effect

Data volume, i.e. measured in megabytes, is an abstract concept and therefore often confusing to consumers [CBB+12, page 3026].This is one of the reasons why consumers tend to overestimate their usage. With a higher estimated usage the flat-rate tariff seems to be more attractive [LS06, page 214].

3.2. Flat-rate bias and data-caps

Strictly speaking a flat-rate tariff is just a data-cap tariff with an infinite cap. Consequently it is possible that some of the effects causing the flat-rate bias are also applicable to data-caps.

3.2.1. How flat-rate biases may apply to data-caps

Some initial research how flat-rate effects may apply to data-cap tariffs was already done by Jan Krämer and Lukas Wiewiorra in 2010 [KW10]. However, they do not study volume-caps but rather cost-caps and the study was conducted for mobile communication services. While both, cost-caps and data-caps, are a hybrid between a pay-per-use tariff and a flat-rate tariff they follow different goals. A cost-cap protects the consumer from unexpected high invoices, a volume-cap shields the service provider from unexpected high usage. Because of this fundamental difference we will make our own argument instead of reproduceing the arguments of Krämer and Wiewiorra.

The insurance effect

The insurance effect is probably also applicable to data-caps because the costs are caped alongside the data volume. How strong the insurance effect will be is subject to the exact properties of the data-cap and the usage perception of the consumer. In all models the price is fixed for a given amount of data, hence the question is what will happen when this amount of data is exceeded.

There are three scenarios for that situation. In the first the connection speed is reduced. In this case it seems very likely that the insurance effect is the same as for a flat-rate tariff, because there will not be any followup costs after the cap is reached.
In the second scenario, after reaching the cap, the connection will be terminated. This scenario might have a huge impact on the behavior of the consumer but regarding the insurance effect there is probably no reaction compared to a flat-rate or the first scenario since the cost is still fixed.
The third scenario is similar the first or the second but additional high speed volume may be purchased. This scenario might have an impact on the insurance effect since additional costs are possible. However additional costs are only associated with additional purchases, so without further research it is impossible to be sure how strong the insurance effect is in this scenario.

The taxi meter effect

As for a flat-rate tariff, the taxi meter effect is probably also not present for a data-cap tariff, due to the fact that it is not a pay-per-use tariff but is payed up front for a given amount of data. On the other hand, at least for the model where the connection is terminated after reaching the cap, this kind of tariff can be regarded as a special form of pay-per-use tariff. On account of that, the tariff can be seen as purchasing a certain amount of data. In that case the price for any amount of data can easily be calculated. Hence it could be hypothesized that such a tariff is an upfront paid pay-per-use tariff. It seems likely the payment upfront eliminates or at least lessens the taxi meter effect, but it can not be said for sure without further research.
There is some indication that consumers perceive the taximeter effect differently for flat-rate and caped tariffs [KW10, page 7].

The convenience effect

Concerning the convenience effect there is a big difference between the different forms of data-caps.
In the first scenario, when after reaching the cap the connection speed is reduced, the convenience effect probably is quite the same as with a flat-rate. This assumption rests on the fact that for the consumer the only difference between this tariff and a flat-rate is the potential change in connection speed during usage. Hence regarding the convenience effect in this model there is no difference between a flat-rate and a data-cap since the consumer makes an upfront payment and gets a predefined service in return. As a result she does not need to check her consumption or similar things to be able to validate the billing. However the scenario drastically changes when there is a different kind of data-cap or even when there is the possibility to buy additional high-speed data-volume. In such scenarios the convenience effect probably suffers whenever there is additional complexity added to the tariff. Hence the other models may have a reduced or no convenience effect. To be sure additional research is needed on that effect as well.

The overestimation effect

How the overestimation effect affects data-caps, or even whether it is present at all, is hard to determine. This is because the overestimation only can be determined in comparison. Because a flat-rate guarantees unlimited use it always compares best in regard to price per byte when very high or high usage is suspected. A caped tariff may compare favorable to a pay-per-use tariff but may compare poorly to a flat-rate tariff when high usage is suspected. So it all depends on how much the usage is overestimated and how different tariffs take that into account.

Summary flat-rate bias and data-caps

Flat-rate tariffs and tariffs with data-caps share some important characteristics. However there is next to no research regarding this topic, so most of the conclusions made here rest on rather strong assumptions. To be sure this research gap should be addressed in the future. If these assumptions are correct a major finding is that different determinants of data-caps have a strong influence on whether such tariffs are perceived flat-rate like and therefore profit from a flat-rate bias.

3.2.2. Choosing between flat-rate and data-cap tariffs

Another important aspect concerning flat-rate effects for flat-rate and data-cap tariffs is how those effects will influence a decision between a flat-rate and a data-cap tariff.
In this scenario it seems likely that the flat-rate tariff will be preferred even in many cases where the data-cap tariff would be cheaper. The reasoning behind that is even though flat-rate effects might be partially present for data-caps in some scenarios, they should always be stronger with flat-rate tariffs. This is especially true for the overestimation effect were the flat-rate will always be the most favorable tariff as it has by definition no boundaries for consumption.
As a consequence a data-cap tariff must offer some other advantages compared to a flat-rate tariff. It could be cheaper, for example.

4. Determinants of data-cap tariffs

The last chapter discussed some possible determinants of data-cap tariffs and their properties. To build a model for data-cap tariffication the determinants of this kind of tariff must be studied in depth. This is necessary to make realistic assumptions for the model or to prove that the model needs no assumptions.

The obvious first determinant of a data-cap tariff is the included data volume. A second important characteristic could be the behavior up from the point the included volume is depleted. And finally, a third and final determinant could be the policies on which data traffic will be taken into account for the data-cap. Or with other words which traffic will not be taken into account when calculating how much of the cap is left.
All three determinants have different aspects and parameters, the details of which will be discussed in the following sections. Probably not all possible properties will be covered, as with increasing experience with such tariffs there will be variation and new ideas. However all possible properties currently used or in the academic discourse will be discussed here.

4.1. Determinant 1: Included data volume

The first determinant of a data-cap tariff is the data-volume included, in other words the cap itself. This determinant, as a result, also offers the possibility of different tiers, in the form of different amounts of included data-volume, to the ISP.

It can be assumed that the utility function for data consumption will have positive decreasing marginal utility. Hence it is important to know the saturation point and take it into account when deciding about the size of the data-cap, otherwise a too low cap will cost the ISP marketshare in a competitive market. In order to determine different tires of data-caps the ISP needs to identify different groups which are characterized by their data consumption. This is nothing new since identifying groups is a standard technique in tariff design [TBAA89].

As Internet traffic is still increasing rapidly and new applications have different demand regarding traffic [Cis12a; Cis12b; GR12] it is likely unsatisfying from a consumer's point of view to have static caps. Consequently the basic differentiation here is whether the cap is static, meaning fixed in the contract for the entire contract period, or dynamic in some way.
An example for a dynamic model could be that when 95% of consumers reach their cap in two consecutive month the cap will be slightly increased in the month following. The same is of course possible for decreasing the cap. Nonetheless such arrangements could be difficult in practice due to legal reasons. And probably more importantly, they most likely set the wrong incentives from an ISP's point of view, as such arrangements would encourage consumers to consume as much data as possible in order to maintain a high cap.

4.2. Determinant 2: Behavior after reaching the cap

As discussed in the last chapter flat-rate and data-cap tariffs share some special characteristics up to a point. One of the most important differences is that a data-cap tariff has a data-limit after which it changes its behavior. In this section the behavior of the connection when the cap is reached is examined closer. To do that the possible strategies are divided into two principle groups. In the first group the connection will be terminated after reaching the cap. In the second group the connection is not completely terminated but some connection properties will be changed to decrease QoE notably. In both cases there is the differentiation whether the original connection properties can be restored or not. The possibility to restore the original connection properties is called a top-up. Such a top-up can be seen as an additional cap, as it will offer the original connection properties for additional volume. So a top-up itself has the properties of the price and data volume.

4.2.1. Connection termination

No top-ups possible

The simplest case is that after reaching the cap the connection will be terminated and there is no possibility to restore the connection until the beginning of the next billing cycle. This kind of tariff seems to be very unattractive as the only advantage over a pay-per-use tariff is a fixed cost limit[1], so some kind of insurance effect. This however comes at the high cost of a strict limit on data consumption. Besides the insurance effect the convenience effect seems to be plausible since there is a fixed price.
Because there is a fixed amount of data for a fixed price the cost per used data unit can easily be calculated. Consequently, there are fixed unit costs and a constant marginal cost, thus a taxi meter effect seems to be plausible as well.

As a result of the relative unattractiveness as compared to a pay-per-use tariff it seems necessary that this kind of data-cap tariff needs some other advantages. For example, a discount on the equivalent amount of data compared to a pay-per-use tariff would be a possibility.
But that model also seems to be unattractive for the ISP, as the only way to increase sales is to up-sell a costumer, thus selling him a tariff with a bigger cap. Consequently such a tariff is to the best knowledge of the author currently not used by any ISP[2].

Top-ups possible

A slight modification to this approach is more common in practice. This model works as follows: the connection will still be terminated after reaching the cap but there is the possibility to buy additional volume. Purchasing additional volume can be done either via

[1] There could also be the advantage of a discount compared to the pay-per-use tariff, but that does not seem relevant in this context.

[2] To be fair this approach somehow is in use. But it can hardly be considered a tariff, as in this case the cap is a upper celling on data usage and multiple overreaches will be considered a breach of contractual agreements and result in complete termination of the service and annulment of the contract. Cf. Comcast's "Fair Use" policy.

a special still accessible website or by interaction with some service employee, for example via phone or in a shop. This model of data-cap tariff could be described as pay-per-use tariff with pre-defined warnings when a certain amount of data usage has been reached. This form of data-caps was quite common in South Africa and is well studied in the article ”You're capped” by Marshini Chetty et al. [CBB+12]. They share the opinion that such data-cap tariffs are a special form of pay-per-use tariffs [CBB+12, page 3021]. Whether or not such a tariff would really not be different from a pay-per-use-tariff is a matter of its own and will be discussed in section 4.2.3.

Despite the question whether or not such a tariff is different from a pay-per-use tariff the question here again is how it compares to a pay-per-use or flat-rate tariff.
When such a tariff is considered as a pay-per-use tariff there is the speciality that there is a rather rude reminder for the amount of consumed data as the connection will be terminated after the cap is depleted. Also to get additional data volume one needs to specially purchase it and pay it upfront. In this respect it could be considered as a special kind of pay-per-use tariff that offers an insurance effect for the cost of potential effort to buy additional data volume in some predefined units.
Compared to a flat-rate tariff this kind of data-cap tariff needs to be cheaper in order to be preferable as it only causes the insurance effect to some extend. As it could be considered a pay-per-use tariff, the taxi meter effect is present in some form. The convenience effect may also be applicable if there is no need to purchase additional volume, but even in this case the convenience effect would be quite weak since there is a need to check the bill because there is the possibility to buy additional data volume.

4.2.2. Decrease in connection quality

Decrease in connection speed

Up until now it was assumed that the connection will be terminated when the cap is reached. Such a model is closely related to a pay-per-use tariff. In contrast to that, there are also data-cap tariffs which are more closely related to flat-rate tariffs. Such tariffs without connection termination will be the subject of this section.

The basic idea is that the tariff includes a flat-rate that will change its connection characteristics after the cap is reached. Hence all effects causing the flat-rate bias should be present. However, it seems quite possible that with the decrease in connection quality some of these effects decrease as well.
The simplest form of such a tariff is that after reaching the cap the connection speed is drastically reduced. This is what DTAG announced in April 2013 as their future standard tariff model for fixed-line Internet access in Germany[3]. This example also shows that when the connection speed is reduced too drastically the tariff seems to loose the perception as a flat-rate and therefore the bias connected with such tariffs. DTAG, due to strong public pressure[4], felt the need to increase the connection speeds of the restricted service after

[3] `http://www.telekom.com/medien/produkte-fuer-privatkunden/184370`, Text in German, last accessed 21st May 2014.

[4] For more information on that topic see 5.1 .

reaching the cap less than two month later, long before the tariffs even were in use[5]. This example illustrates that such a data-cap tariff can be perceived flat-rate like but only as long as the restricted service still is perceived as useable basic access to the Internet.

Restrictions to the number of services available

Another alternative for restricting the service after reaching the cap is restricting the number of services usable or makeing some restrictions within a service. An example for that would be that in South Africa some ISPs with data-cap tariffs only had a cap on international websites. South African websites were still available even after the cap was reached [CBB+12, page 3024]. Such a scheme clearly conflicts with the principle of net neutrality[6] and also touches on the third determinant for data-cap tariffs. Because of that the details of such a tariff will not be discussed here as they can be found in the next section.

General aspects of service limitations

As with the data-cap models with connection termination, there again is the question what influence the possibility to restore the original connection properties will have. One can argue that the need for top-ups will not be that big since there is still a restricted connection. On the other hand a restricted access will either block some service or make some services unusable because of bandwidth limitations. After all the purpose of the limitation is to prevent extensive usage. Also if a consumer needs one of the services that will not be accessible in the restricted traffic class, there is no difference to her whether the connection has be terminated or restricted. Even if the service is still accessible but needs longer to load, the QoE will suffer. Consequently there is a clear incentive to leave the restricted class. This is necessary for a differentiation from a flat-rate tariff and makes the possibility of top-ups desirable for consumers. This thoughts are backed up by findings from HCI research examining the effects on consumers with poor Internet connectivity [WSC+10, page 2596][CADS10, page 3]. Another finding from this research is that high bandwidth is taken for granted and therefore web services are designed under this assumption [WSC+10, page 2593] [CBB+12, page 3029]. But this assumption does not hold when looking at developing countries or, in our case, households that have already reached their caps. Hence such caps could be an incentive for service providers to develop web services that save as much data as possible in order to be accessible even with a throttled connection.

4.2.3. The difference between data-caps with top-ups and pay-per-use tariffs

In section 4.2.1 it was hypothesized that a data-cap tariff which terminates the connection after reaching the cap but offers the possibility to buy additional cap is essentially a pay-per-use tariff. Although this opinion is shared by Chetty et al. [CBB+12, page 3021] as

[5] `http://www.telekom.com/medien/produkte-fuer-privatkunden/187666`, Text in German, last accessed 21st May 2014.

[6] For more on net neutrality and how QoS conflicts with it see section 4.3.1.2.

well as the literature on point rationing[Tob52, page 530, section 2.3], which will later be used to build a model, it seems to be only partially true. Such a non-linear pay-per-use pricing scheme, where one can only buy services in packages containing more than one unit of a service, alter the utility function. The utility function is transformed from a continuos to a step-function. This change in the utility function will also alter consumption compared to consumption under a continuos utility function. The main reason for that is that the marginal costs are altered dramatically: At many points on the consumption curve the marginal costs are now zero, because the consumer already bought a package which entitles her to consume a given number of service units. For a few other points on the consumption curve the prices have dramatically risen since the consumer needs to buy another package to consume another unit of service.
In tariffication such a tariff is called a two part tariff [Oi71]. The first price is the cap which, for a given price, allows access to the network and a given amount of usage of said network. The second part are the top-ups. Top-ups can either be analogous to pay-per-use tariffs, generally speaking linear pricing, or consist of packages extending network usage beyond the cap, which would be some sort of non-linear pricing. Such two part tariffication can be seen as two-part marginal cost pricing, which tries to keep as many consumers in the market as possible through affordable access fees while still recovering marginal cost of heavy usage [NW74].

Besides this purely economic argument of consumption in packages there are also psychological arguments to it.
The consumer likes to get her moneys worth. So it is possible that she consumes more when forced to consume in packages in order to reduce average costs[7] [JW11]. In other words the consumer tries to recover her investment into another package of data volume. Under the classical economic psychological framework of Richard Thaler [Tha80] one would speak of the sunk cost effect [Tha80, page 47].
So although the ability to purchase additional cap lessens the effects created by the cap, it does not transform the pricing scheme to a pay-per-use tariff. The changes to the utility function as well as the psychological effects such package consumption induces should always be kept in mind when discussing such problems.

4.3. Determinant 3: Policies on which traffic is taken into account

This determinant may be the most interesting for an economist because it can create a real difference to a pay-per-use or flat-rate tariff. Especially Quality of Service techniques are interesting under a data-cap tariff regime.

An emphasis in this section will be on QoS and the economic aspects of such techniques, of course especially with regard to data-cap tariffs.
Another way of differentiating network traffic is not by the kind of traffic but by the time

[7] In economics this phenomenon is called mental accounting. It was introduced by Richard Thaler in 1980 [Tha80], it ist also the reason for the taxi meter effect within the flat-rate effects.

it is sent. Such time based differentiation of network traffic will also be the topic of this section.

4.3.1. Quality of Service

4.3.1.1. The motivation behind QoS

QoS is not necessarily a part of data-cap tariffs, nevertheless it seems rather obvious to use QoS in data-cap tariffs to enable premium services to stay unaffected by the cap. This tendency becomes even stronger when a flat-rate tariff used to be the standard and consumers got accustomed to this pricing scheme.
QoS is at its core a guaranteed service characteristic. This may not always be true although when talking about QoS one usually refers to the preferential treatment of some data packages, technically speaking it can also be the exact opposite. For example a quality of service policy can consist of rules for downgrading non-time-critical traffic [BNW03]. So under such a regime no service would be preferred, but non-time-critical services, such as e-mail or file-sharing, would only get capacity when it is not needed by time critical services, such as VoIP or video-streaming. Such QoS regimes are especially interesting since ISPs often justify QoS regimes with preferential treatment and extra billing with network congestion. A lower effort regime for non-time-critical traffic could also lessen network congestion but is less likely to create enough utility for the consumer to be billed extra. A classical example for QoS in todays Internet would be that it is guaranteed that a video streaming service has at least the bandwidth which is necessary for smooth playback.

4.3.1.2. Net Neutrality

To be able to guarantee a certain quality for a service, it is necessary that the data transmission of the different services can be attributed to them. Or more generally speaking the data packages have to be distinguished. To do that one has to break with one of the Internet's original design principles: the neutral handling of data packages. This principle is often described with the term "net neutrality". In a neutral net all data packages are forwarded in the order they arrive regardless of their source, destination or content. Of course under such a regime it is impossible to offer QoS or even exclude some data from the restrictions of a data-cap. To enable QoS or some exclusion from the data-cap it is necessary to have a look at the meta data of the data packages or even their content.
The topic of net neutrality is controversially discussed around the world. In some countries this discussion even led to some sort of net neutrality law[8]. One major fear concerning abolishment of net neutrality is the abuse of monopolistic power and with it the destruction of the leveled playing field property of the Internet which made it great for innovation. However, these arguments have only been proven in parts in the scientific discourse [CBG11; Sid06; Kru08; Cro07]. There are even voices that the Internet never really has

[8] In Europe most notably a law guaranteeing net neutrality is in place since May 2012 in the Netherlands; `http://arstechnica.com/tech-policy/2012/05/netherlands-becomes-worlds-second-net-neutrality-country/`, last accessed 21st May 2014. Also the European Union as a whole just recently made a legislative push towards net neutrality; `http://www.bbc.com/news/technology-26865869`, last accessed 21st May 2014.

been a leveled playing field [Cro07, page 51].
Net neutrality will not be discussed in detail here since it is a wide topic of its own and need not to be part of the data-cap argument. Nevertheless, it is noteworthy that a data-cap, in contrast to a flat-rate tariff, leads to additional incentives for QoS-agreements. This has a rather big impact since the absence of net neutrality undoubtedly will transform the Internet to a two-sided market [MSW09]. This phenomenon has to be considered when trying to build an economic model for data-cap tariffs.

4.3.2. Time of day based cap

When arguing with congestion as a reason for a shift from flat-rate tariffs to data-cap tariffs, as DTAG did for example[9], it seems logical to apply the cap only to the congestion threatened peak hours. Such an idea seems especially reasonable regarding the fact that the projected increase in traffic is estimated to be particularly large during peak hours [Cis12b, page 5].
So a corresponding policy would divide the day, or some other period of time, into off-peak-times and peak-times[10]. Data-traffic would only be credited against the cap during the peak-times[11]. A problem with such tariffs is that they could easily get complicated because the consumer needs to know which times are peak-times and alter her usage behavior accordingly. As a result such a tariff would probably have none of the flat-rate biases and therefore would need to be substantially cheaper to be able to compete with flat-rate tariffs. Additionally, such a policy could potentially contradict itself as a rational consumer would shift her consumption as much as possible away from the peak-times to the off-peak-times. In extreme cases this would switch the peak- and off-peak-times. Such a dramatic shift seems to be unlikely since peak-hours are mostly determined by factors out of control of the consumer[12].

[9] `http://www.telekom.com/medien/produkte-fuer-privatkunden/184370`, in German, last viewed 21st May 2014.

[10] More complex differentiations are of course possible. They are not regarded here and probably will never be used since such complex differentiations are hard to put into contractual terms and even more importantly hard to explain to the consumer.

[11] An example for such tariffs would be the capped consumer tariffs of webafrica. In their capped tariffs the consumer gets two equally large caps (called anytime Gigs and nighttime Gigs), the first cap can be used anytime, the second only between 12 am and 6 am (off-peak-time, advertised as ”Download Plus”); `http://www.webafrica.co.za/home/adsl/`, last accessed 21st May 2014.

[12] For example there are peak-hours when many people have free time at the same time, such as weekends.

Determinant 1: volume	**Determinant 2: after cap**	**Determinant 3: policies**
• fixed volume	• connection termination – top-ups – no top-ups	• no QoS
• variable volume	• speed reduction – top-ups – no top-ups	• QoS
	• service restriction – top-ups – no top-ups	

Table 4.1.: Overview of the properties of data-cap determinants

5. The current state of broadband tariffication

In the preceding chapters general aspects of data-cap tariffs have been discussed. Now the focus will be on the current state of broadband tariffication around the world. Starting with the current situation in Germany the section will continue with some examples from around the world. After this description of the current state of broadband tariffication, plausible determinants for the design of an economic model for data-cap tariffs are proposed. These determinants will later be used as reference for the model derived in the next part of this thesis.

5.1. Situation in Germany

For a long time no major ISP in Germany has had a data-cap tariff in active use. Still some had and have a "fair use policy" in place, most notably Kabel Deutschland[1]. Such policies are usually no data-caps but may have similar effects as the consumer will be restricted in consumption. 1&1 has, and has had for a long time, a special entry level tariff wich is a data-cap tariff[2]. This tariff is clearly identified as data-cap tariff and marketed in contrast to their flat-rate offerings.

Most notably DTAG made a push towards data-cap tariffs in April 2013[3]. While not immediately offering such tariffs they modified their terms of service for all their broadband Internet tariffs for consumers in such a way that they could be transformed into data-cap tariffs at some unspecific point in the future. DTAG stated that they will enforce data-caps for all contracts under those new terms of service[4] as soon as they see the necessity for it. They justified the necessity for the caps with potential capacity overload in their network and projected them to apply no sooner than 2016. Under this plan the cap should be fixed at 75 GB to 400 GB, depending on the bandwidth-class purchased. After reaching the cap the connection will be throttled to 384 Kbit/s, regardless of the purchased bandwidth. Top-ups will be offered. Also a first QoS service which would not be affected by the cap and not taken into account when calculating the remaining cap was announced. This service was their own IP-TV service.
This plan led to protests against the new tariffs and started a discussion about net neutrality and the right to a basic connection to the Internet, as Internet access is regarded as a vital infrastructure. As a result the federal minister for economic affairs wrote a letter

[1] They state in their terms of service that when a customer downloads more than 10 GB in one day her available bandwidth for file-sharing will be throttled to 100 Kbit/s. Text in German: AGB Kabel Internet und Telefon von 10.2013, Leistungsbeschreibung B.) Internet 2. b.), accessible through `http://s3.kabeldeutschland.de/dl/static/media/AGB_Internet_Telefon.pdf`, last accessed 21st May 2014. In the same document they state other potential network management measures.

[2] `https://dsl.1und1.de/`, text in German, last accessed 21st May 2014.

[3] `http://www.telekom.com/medien/produkte-fuer-privatkunden/184370`, text in German; press release announcing the shift towards data-cap tariffs for fixed-line consumer broadband-access; last accessed 21st May 2014.

[4] For legal reasons in Germany they could not transform the contracts of existing customers. But sooner or later old contracts will be upgraded which will make them subject to the new terms of service.

to DTAG CEO René Obermann criticizing the plans[5]. This letter and the general public outrage built pressure, if nothing else also because the federal government still owns about 30% of DTAG[6]. Under this pressure DTAG modified their initial plans and upgraded the available bandwidth after reaching the cap from 384 Kbit/s to 2 Mbit/s[7], otherwise they left their plans unchanged. This is interesting from a technological point of view as 2 Mbit/s can still be considered as broadband and sufficient for many applications. Hence such a policy will make economic modeling harder as only part of the Internet usage is affected by the restrictions.
Following the announcement of DTAG, O2[8] transformed all its broadband offerings for consumers to data-cap tariffs[9].

5.2. Short overview of data-cap tariffs around the world

Data-cap tariffs are by far not just a German phenomenon. In fact in other parts of the world they have been in use for quite some time now. Still the public perception and the acceptance vary vastly.
A good first example is the situation in South Africa, which was examined in the study "You're Caped!" by Chetty et al. [CBB+12]. At the time of their study in 2010 data-cap tariffs were the predominant tariff type for broadband Internet access and uncapped tariffs were a rather rare and new form of tariff in South Africa [CBB+12, page 3022]. But as the market in South Africa develops, flat-rate tariffs become increasingly common. They came to a coexistence of capped and uncapped broadband tariffs. Most major ISPs such as Telkom, Afrihost and Axxess now offer a variety of tariffs of both types. While there has been a general shift towards uncapped flat-rate tariffs, ISPs argue that their tariff design keeps capped tariffs attractive and they can prove that claim with subscriber numbers[10].

South Africa started out with capped tariffs[11]. Australia and New Zealand also started their broadband area with data-caps, the reason here manly being the huge popularity of US Internet services due to the common language[12]. Later all three countries introduced uncapped flat-rate tariffs. In Germany it is the other way round. So because of the strong psychological effects of a flat-rate tariff[13] such a change in tariffication could be rather

[5] http://www.spiegel.de/politik/deutschland/flatrate-plaene-der-telekom-bundesregierung-kritisiert-Internet-bremse-a-896215.html, text in German, last accessed 21st May 2014.

[6] 14,5 % directly and 17,4 % through KfW, http://www.telekom.com/company/at-a-glance/81660, last accessed 21st May 2014.

[7] http://www.telekom.com/medien/produkte-fuer-privatkunden/187666, text in German, last accessed 21st May 2014.

[8] A German subsidiary of Telefonica.

[9] http://dsl.o2online.de/provider/content/segment/anbieter/produkte/dsl-home/, text in German, last accessed 21st May 2014.

[10] http://mybroadband.co.za/news/adsl/95563-capped-adsl-making-a-comeback.html, last accessed 21st May 2014.

[11] To be precise they started out with capped tariffs with the introduction of broadband due to the stage of development of the infrastructure.

[12] This popularity is a problem since sea cables connecting those two countries to the USA are a scarce resource.

[13] For details about this see section 3.1.

hard to sell. An example for such a shift would be the situation in the USA[14]. Today more than half of US broadband subscribers have caped tariffs and usually no real choice to quit them. This is mostly due to the fact that some of the largest ISPs have either made data-caps an integral part of their service offerings[15] or have experimented with them for quite some time now[16]. This push towards capped tariffs has lead to similar reactions as in Germany a few years later[17].
Despite all that, data-caps seem to spread further. This development is probably due to the fact that broadband Internet access often is a local monopoly and the FCC so far failed to enforce strict net neutrality[18].
The introduction of data-caps in the USA raised awareness for the subject. One result of this is that overview articles about data-caps in the US[19] and around the world[20] can now be found online.

5.3. Proposed determinants for a basic model

To be able to formulate the determinants for a basic model it first needs to be defined what the model should be capable of. A first basic model for data-cap tariffs needs to be as simple as possible while being able to be expanded in the future.
As the ISPs probably did their own modeling to determine their benefit from such tariffs it is probably more interesting to start out from a consumer's perspective. As a first step it is wise to start with a single consumer and not the aggregated view on the whole consumer base of an ISP introducing data-caps.
Consequently, the basic model should be able to describe the impact on consumer behavior through a second budget constraint in form of a data-cap and the resulting welfare shifts.

---

[14] The situation in the USA is different to the situation in Germany in the sense that fixed-line Internet access in almost all regions of the US, is still a local monopoly. Therefore tariff changes are a take-it-or-leave-it arrangement.

[15] AT&T being the he primary example; `http://www.att.com/esupport/Internet/usage.jsp#fbid=Tz7VcjOsqjf`, last acceded 21st May 2014.

[16] A prominent example here would be Comcast, which was the first major ISP in the US to introduce a data-cap and still experiments with such tariffs; `http://customer.comcast.com/help-and-support/Internet/common-questions-excessive-use/`, last accessed 21st May 2014. This tariffication will probably spread when Comcast and Time Warner Cable have merged, forming the by far biggest and near monopoly cable company in the US `http://www.washingtonpost.com/business/economy/comcast-time-warner-agree-to-merge-in-45-billion-deal/2014/02/13/7b778d60-9469-11e3-84e1-27626c5ef5fb_story.html`, last accessed 21st May 2014.

[17] `http://www.fcc.gov/blog/new-rules-open-Internet`, last accessed 21st May 2014, `http://www.publicknowledge.org/letter-to-FCC-on-ATT-Data-Caps`, last accessed 21st May 2014, `http://www.techhive.com/article/2055502/why-your-Internet-service-provider-wants-to-nix-all-you-can-eat-data-buffets.html`, last accessed 21st May 2014.

[18] The FCC has tried to enforce net neutrality right after Comcast's first push towards data-caps, but their rules where successfully challenged in court by broadband providers `http://news.cnet.com/8301-13578_3-57617199-38/appeals-court-strikes-down-fccs-net-neutrality-rules/`, last accessed 21st May 2014.

[19] `http://gigaom.com/2013/11/15/data-cap-2013/`, last accessed 21st May 2014.

[20] `http://arstechnica.com/tech-policy/2011/04/how-Internet-users-are-disciplined-around-the-world/`, last accessed 21st May 2014; `http://gizmodo.com/5813010/how-bad-do-we-really-have-it-bandwidth-caps-around-the-world`, last accessed 21st May 2014.

Determinant 1

A basic model should include all essential determinants but at the same time be as simple as possible.
Determinant 1 was defined as the data volume constituting the cap. It was distinguished between a fixed and a variable cap. For the economic model this differentiation is irrelevant as the data-cap is a strategic variable set by the ISP. Hence the cap has a big influence as it represents a second budget constraint, but will be seen as an exogenous variable in this consumer centered model. As a result the volume of the data-cap is a variable of the model. The setting of the data-cap is a strategic variable of the ISP that is determined ex-ante

Determinant 2

The second determinant was defined as the behavior of the connection after reaching the cap. Since this determinant constitutes the most obvious difference to a flat-rate or pay-per-use tariff it was studied extensively.
Three kinds of behavior were studied: connection termination, connection-speed reduction and service restrictions. For all three variants there could be the option to restore the uncapped service properties or this possibility could not exist[21]. Since all three types exist in real world tariffs it is hard to say which one should be the basis for the model. For reasons of simplicity[22] it seems wise to start out with the simplest case where the connection is terminated after reaching the cap and no top-ups are available for purchase. The option of top-ups could alter the model[23] and should be considered in a second step. Another reason for choosing the simple case of connection termination for a first model is that some cases of connection speed reduction can be seen equivalent to this. This is the case when the remaining connection speed is not sufficient to support the web application the consumer wants to use. So whether a data-cap tariff with connection speed reduction is superior to a data-cap tariff with connection termination is based on the consumer's utility function. If every service that creates utility for the consumer is not usable with the reduced connection speed, her utility at this moment will be zero, as it would be if the connection was terminated[24]. Also universally modeling data-cap tariffs with restricted service properties after reaching the cap could be rather difficult as it heavily relies on the utility function of the consumer. There is a similarity to data-cap tariffs with connection speed reduction as in both one must check for each web service whether it is still useable under the changed connection properties after reaching the cap. This makes it difficult to group web services. As a result most web services cannot be arranged in groups but must be examined individually resulting in a large number of variables. Additionally service restricted models are quite a special tariff and need to have a strong justification, such as

[21] Usually this is the case in order to encourage the consumer to buy a higher level service tire, which is usually more expensive and creates more revenue.

[22] It also seems like this kind of tariff is at the moment the most common data-cap tariff, which could be another argument for choosing this property in a model.

[23] How the possibility of top-ups heavily influences this kind of tariff was discussed in section 4.2.3.

[24] This problem was already discussed in the third paragraph of section 4.2.2.

sea-cable capacity, as such tariffs seriously conflict with the concept of net neutrality[25]. Such a violation of net neutrality probably will not be accepted by both regulators and the public without proper justification.
So in a first step the model should examine data-cap tariffs with connection termination. In a second step it should be tried to have a look at what happens when top-ups are introduced.

Determinant 3

Regarding the third determinant the main focus was on QoS regimes. As QoS regimes are almost inherent to data-caps, it seems a logical economic assumption that they need to be part of the model. Time of day based pricing seems to be rarer in practice and also more complicate to model. Hence it will be left for further research. Also time of day based pricing, besides being part of data-cap tariffs, is closely connected to general economic theory of congestion pricing. Therefore one can be pointed to the general literature on that topic, for example [MMV95].

Summarizing the determinants for the basic model: It will not make any assumptions towards determinant 1, since this is not necessary. Regarding determinant 2 the connection will be terminated when the cap is reached. Top-ups should be considered in a second step. Advanced data-cap tariffs which decrease connection quality will not be part of this first basic model but should be subject to future research. Regarding the third determinant QoS regimes will be an essential part of the model.

[25] See section 4.3.1.2.

Part II.

The economics of data-cap tariffs

6. The economic characteristics of data-caps

As a start it needs to be evaluated in which aspects a data-cap tariff, as defined in section 5.3, differs from the standard model of consumer choice.
A flat-rate tariff, as the point of reference, is pretty straight forward the basic model of consumer choice. The choice of the consumer is only influenced by the prices of the services and their utility. So the consumption is subject to a money budget and in most cases a time budget, which is normally assumed as included in the utility function. Since many Internet services are advertisement supported, the money price for the consumer is often zero. However, the advertisement influences the utility of the web service [DME06, page 182], so the utility function is even more important, as it virtually also includes the money price. Accordingly, choosing a flat-rate tariff the consumer maximizes her utility subject to her money budget.
This model rests on the assumption that consumption is only restricted by the money budget available. But for a data-cap tariff this assumption does not hold. Besides money, data is also restricting consumption. Further consumption will be impossible after reaching the cap because the connection will be terminated. Hence besides the money budget there needs to be a second constraint, a data budget determined by the cap. Consequently the consumer now must maximize her utility subject to her money budget *and* her data budget.

The second core property of the described data-cap tariff is the QoS-scheme. The QoS-scheme should guarantee that certain services are usable irregardless of the cap. So in other words the QoS-scheme enables that the traffic of some services will not be charged against the cap. Consequently services under such a QoS-scheme would be freed of the data restriction, their data price would be zero.
So far a data-cap tariff would only bring one change to the economic model of consumer choice: The consumption will be restricted not only by money but also by data consumed. Additional constraints besides money restrictions are rare in a free market economy and for that reason there is little research on that topic. Normally additional constraints are induced for legal reasons in order to protect someone or something, such as health or environment. A model with such restrictions could be the basis for a model for data-cap tariffs if they are uniformly applicable to a group of services. Such broad restrictions are rare since they are a big intervention in the free market and ultimately peoples' lives. A recent example would be the EU pollution certificates, called EU-allowances. They limit the production within the EU not only by the resources available, which are purchased with money, but also by the the pollution certificates allocated. However this model regards production rather than consumption so it might not be well suited for a data-cap tariff model.
But this example outlines pretty much the core of the changes necessary to handle data-cap tariffs. With a second budget constraint consumption is further limited, hence it introduces an artificial shortage of consumable goods. Such shortages require an allocation mechanism

outside the free market. They are usually necessary during natural or man-made disasters, for example floods or wars. Such a scheme that allocates each individual a consumption limit independently of her money income is called rationing.
The economic theory of rationing was mainly developed during World War 2 and the direct aftermath. Best documented in economic literature today is the rationing system used in the United Kingdom.

7. Introduction to the economic theory of rationing

Although rationing was literally vital to some nations during times of war, especially World War 2, it is a relatively small scientific field. The literature on this topic is probably quite extensive but many models only were published in military and government documents not accessible to the public. The scientific literature published still covers most theoretical aspects and is closely connected to each other. Most of the scientific literature published is from the time during and shortly after World War 2. After the war rationing mostly lost its relevance to the scientific community [NR80, page 25]. In the late 1970s there was a short time of new found interest in the topic of rationing[1]. But more importantly, starting at that time, research based on the basic findings of behavior under external constraints other than money found its way into various new research topics. Later areas of research building on the results of rationing research were general research on distributional efficiency, considering rationing as one alternative, as well as research on credit rationing and time allocation, since time is always a constrained resource. In general one could say that the general topic of scarcity made rationing literature influential to many research areas that have to deal with constraints, this even affects research outside of economics.
In this chapter the basic elements of economic rationing research will be presented in order to be able to later apply them to the data-cap problem.

7.1. The different kinds of rationing

Different kinds of rationing are distinguished according to the number of rationed goods and the basic rationing mechanism used. A good overview of the different types of rationing and other basic aspects of rationing is given by the articles "Theoretical Aspects of Rationing" by H. P. Neisser [Nei43] and "The Political Economy of Consumers' Rationing" by T. de Scitovszky [dS42]. These articles contain more details than what will be reproduced here since only that which will be later needed for modeling data-cap tariffs will be introduced here.
First it is distinguished whether individual items or groups of products and services are rationed. To ration groups of products, instead of individual items, leaves some choice to the consumer since she can choose within the group [Nei43, page 378].
If individual goods or services are rationed the consumption limit for each item can be determined directly. This type of rationing is called specific or straight rationing [Tob52, section 1.2, page 522 f].
If groups of products or services are rationed there needs to be a metric to distinguish the value of the items within the group. This can easily be done by using their money value, such a scheme is called value rationing [dS42, page 116]. However money value could be a bad metric as rationing in most cases is an intentional shift away from the market mechanisms of price determination in order to distribute scarce goods in a socially

[1] Mostly because of the oil crisis which made things such as mobility and energy scarce resources.

acceptable manner [dS42, page 114 f]. Hence in such cases there often is the need for another metric besides money to value the products or services within a group. Such a scheme that establishes a second currency[2] [Car48] is called point rationing [Llo42]. In this scenario each commodity has a price in money as well as points. As there could be different ration groups there also could be more than one point currency [Tob52, page 521]. Hence a consumable item or service has a price in money and the point currency of its group[3].

For point rationing to work some rules on how to form the different ration groups are needed. A simple but effective rule is to arrange substitutes in a group [Nei43, page 389]. Such an approach makes sense as a goal of point rationing is to allow choice while still enforcing a consumption limit.
When almost all or most goods or services available are rationed and in the same ration group the scenario is called general rationing [Kal41][dS42, page 116] or over-all rationing [Nei43, page 379 ff]. The case that all goods are rationed is not considered rationing by a strict definition, as rationing implies that there are unrationed goods[4] [dS42, page 119].

7.2. Fundamentals of rationing theory

7.2.1. General properties of rationing

The essential feature to rationing is that it introduces a ceiling to consumption. As shown in the previous section this can either be done by introducing a limit of consumption for individual products, straight rationing, or by introducing additional currencies, point rationing. In order to be effective the ration income in a point rationing scheme must not be directly related to work or money income [Tob52, page 521], otherwise it would obey to the same logic as money and therefore could not introduce a rationing effect. This is not the only aspect in which ration currencies differ from money: Ration currencies can not be saved [Tob52, page 522]. So either the ration currency is spent within its validity period[5] or it becomes worthless. This has serious consequences in economic terms. It sets the future, as in the next period or later, money value of point currency to zero so it incentivizes to spend it all today[6], meaning within this rationing period. Another difference between the point currencies and money is that money is the only currency that traverses the whole

[2] In the case of straight rationing there is no second currency. Either money is still the only currency and an arbitrary budget constraint is established or rationing coupons completely substitute money as currency.

[3] Of course an item could be in different groups. In this case the consumer could choose which point currency she likes to spend. However it is important to note that even though an item is in different groups it is not possible to make a purchase using money and a combination of different point currencies. So a purchase of an item can only be made using money and *one* kind of point currency.

[4] Rationing, in its economic meaning, by its definition is a temporary restriction to market driven pricing. If all goods are rationed, the pricing is done outside the market. Examples of that would be centrally planned economies in socialist countries.

[5] Actually the ration currency could be invalidated any time. While this in theory is also true for money, it would be much harder and the reaction probably much more severe.

[6] However one can argue that this effect is not that important since consumption is already restricted to a low level and one also still needs money in order to make purchases. Nevertheless, it still is a restriction and such policies also have psychological effects.

economic cycle [Tob52, page 522]. This is true because a ration currency is only applicable for a group of substitutable services, so the provider of that service probably has no need for the ration currency and needs to convert it.

From a welfare perspective there are some indications that rationing harms the consumer less than other means of consumption restriction respectively consumption reduction such as income taxes [Red42]. Although that might not be true for all cases [Har43].

7.2.2. Rationing and budget constraints

Probably the most obvious feature of rationing is that it constrains consumption. In economic terms this means that one or more additional constraints besides the budget constraint are added to the classical economic model. The theory for choosing between two goods under rationing was mainly developed by Worswick [Wor44], Rothschild [Rot45] and Makower [Mak45] as well as Carter [Car48].

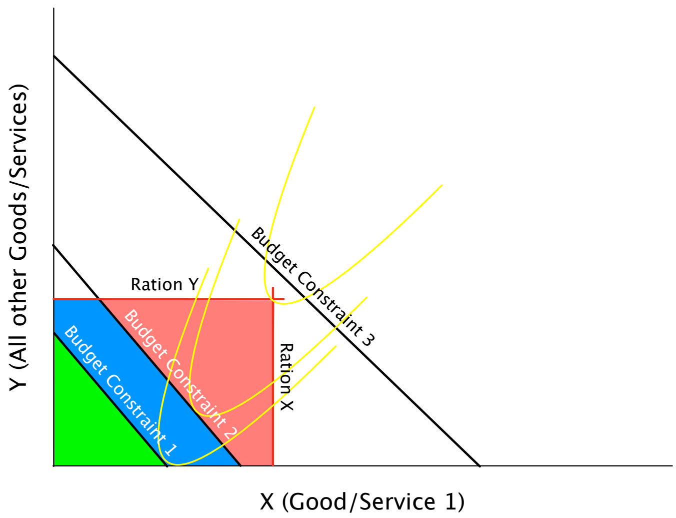


Figure 7.1.: Indifference map in the case of straight rationing.

In the case of single currency rationing the situation is rather simple: the rationing provides an upper limit to consumption. How this limit affects the indifference map is shown in figure 7.1. In the case of budget constraint 1 the rations for both goods in the standard dual good model have no influence, since the consumer is too poor to be affected by rationing. So every point within the green surface can be consumed.

For budget constraint 2 the consumer is still not restricted in her consumption of good X[7], but is not able to spend all her income on good Y. Therefore her possible consumption set completely lies within the blue and green surfaces. Consequently, her budget constraint is altered: The ration line for Y becomes part of the budget constraint wherever the y-value of the budget constraint is bigger than ration. So the consumer is affected by rationing for some goods while for other goods she is not. She is not affected by rationing for some consumption points because her budget restrains her even further.

[7] Of course the other way around, namely a restriction in consumption of good Y would be possible as well.

In the case of budget constraint 3 the consumption is completely subject to the rations since the consumer could buy more from every good without rationing. So the possible consumption set is defined by the rations, consequently they must lie within the red, blue or green surfaces.

In the preceding example it was observable that rationing could or could not affect the possible consumption sets of a consumer. To describe that there is a general term: When a ration alters the consumption set it is called effective [Tob52, page 526]. Ergo if a ration does not affect the consumer it is called ineffective. Consequently, a ration should be effective because otherwise it is of no direct economic use. Nevertheless, it can still serve as a psychological limit which could also influence the consumer[8].

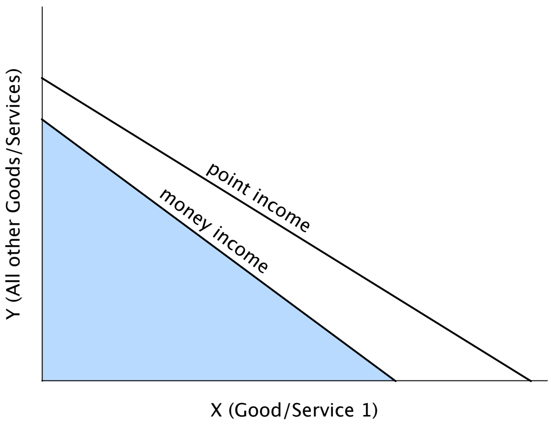
When additional currencies are introduced, in the case of point rationing, the basic logic of restrictions from single currency rationing holds. However, the restriction is of a different nature, so it is worth to discuss it here.
As can be seen in figure 7.2 the introduction of another currency quite literally introduces another budget constraint. Therefore the possible consumption set needs to fulfill both budget constraints. Graphically speaking the possible consumption set is the area between the axes and the parts of both budget constraints nearest to the origin.
There are three cases to distinguish: Money could be the sole constraint, ration points could be the sole constraint or money as well as ration points could be the constraint. These three cases are called money standard, points standard and double standard respectively [Mak45, page 76]. With the money standard only the money ration is effective while the points ration is ineffective. With the points standard it is the other way around. And with the double standard the money as well as the points ration are effective. As one can see in figure 7.2 c, the double standard is a bit more complicated: On the straight defined by A and B it is a money standard and on the straight defined by B and C it is a points standard. Only in the point B both rations are effective and therefore it is a true double standard.

7.3. Applying rationing theory to data-cap tariffs

7.3.1. Parallels between rationing and data-caps

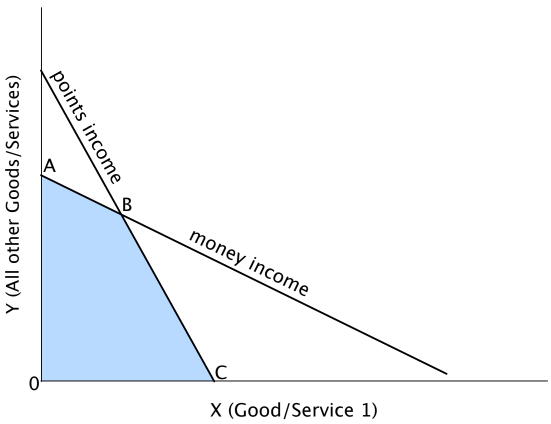
The first question that arises when the economic rationing theory is applied to data-cap tariffs is which kind of rationing should be chosen. A first thought could be to use straight rationing as only the data consumption is rationed. This thought is misleading however, as there only should be an upper limit to data consumption and no limit on the choice. It seems more constructive to look at the problem as a dual currency problem, with money and data as the two currencies. That way point rationing must be chosen to enable a multi-currency-system.
Since it should be a dual currency problem there can only be one rationing group. This raises the question how having only one rationing group for all kind of Internet services can be justified. A video on demand service probably can not be regarded as a suitable

[8] This was shortly discussed in the last paragraph of section 4.2.3.



(a) Money Standard: Only the money budget constraint is binding, there are more points than necessary.

(b) Points Standard: Only the points budget constraint is binding, there is more money than necessary.



(c) Double Standard: Money as well as points can be spent fully at point B. This is also the only point where both constraint are binding at the same time.

Figure 7.2.: How a dual budget constraint restrains the possible consumption points. Reproduced and adapted from [Car48, page 587].

substitute for an online retailer or a website with recipes for Italian dishes. Consequently, the main grouping criteria from rationing theory does not seem to apply here. The argument for a single rationing group here however is a rather practical one. The hypothesis that ration groups should be found on similarity is based on the assumption that only substitutes should have the same ration currency in order to let consumer choose while still restricting consumption for certain categories of goods [Nei43, page 389f]. To be even more concrete the groups were designed to put an upper limit on consumption of basic needs while still allowing choice within basic categories. Even if Internet access nowadays sure can be considered important it is not vital to human beings, so there is no need to restrict access to certain categories of services. While the argument still holds that a video on demand service is no substitute for an online retailer they sure consume the same kind of data[9], in contrast to food and clothing which both are vital but no substitute for each other. With these arguments rationing for data-caps can be seen as a different kind of rationing its goal being not to guarantee minimum quantities of necessities but to prevent excessive use of resources which are limited. As a result it is not necessary to distinguish different rationing groups but a single group can be used since all service use the same scarce factor, namely data transmission capacity.

A problem closely related to the grouping of goods and services is the fixing of point prices. Normally this is a quite complex process since various variables need to be accounted for [Llo42, page 50f] [Nei43, page 392f] [Tob52, page 545 ff]. In the case of data-caps however the point prices are exogenous, they are determined by the data usage of the respective web service. This is another argument for a single rationing group as the choices within the groups can not be controlled through fixing prices, as they are already fixed.

So far data-cap tariffs come down to a point rationing scheme with one rationing group and the point prices for that group are already set by technical means.
An important speciality for Internet services is that many of them are offered for free and revenue is created through advertisement. This is no problem to the model since the money price can be zero. Therefore the choice of such a service is not restricted by the money constraint but by the data constraint. The choice for all services is subject to their utility since the basic economic model states that the consumer maximizes her utility. This is important to state explicitly as an advertisement supported service reduces the money price to zero but the utility of such a service may be lowered due to distracting advertisement. For example there could be two web services which offer the same function to the consumer: service A which costs X money units and Y data units and a service B that costs 0 money units and also Y data units. Service A utilizes no advertisement but service B creates all its revenue form advertisement. As the advertisement might distract the consumer from the original purpose of the service it could lower its utility. To summarize the problem: The data consumption and functionality is the same, the services only differ in utility and money price. Therefore it is possible that indifferent to the difference in money price both services are equally probable to be picked.

[9] At this point it is ignored that the video streaming service needs real time streaming, while the online retailer's service still functions with a certain delay.

For data prices a similar argument can be made in regard to determinant 3 for data-caps[10]: Every web service that is not taken into account regarding the allowed data consumption within the data-cap has a data price of 0. So an analogous example to the one just discussed can be created: Again there are two web services A and B, both services offer the same functionality and have the same money price. However service A is taken into account for calculating the data-cap and therefore has a data price of X, while service B will not be taken into account and therefore has a data price of 0. Now when the consumer maximizes her utility both services are equal because they have the same functionality and therefore the same utility. As a result the decision is made based on the constraints. Since both services have the same money price they are equal in regard to the first constraint. However for the second budget constraint the price of A is higher than that of B. Consequently, the decision is made because of the second budget constraint and service B will be preferred.

The last example illustrates what such a rationing scheme introduces: It enables a second budget. Accordingly, the consumer will now maximize her utility subject to both money price and data consumption of a web service. Hence both prices are a strategic variable to the web service[11].

7.3.2. Important parts of rationing theory not applicable to data-cap tariffs

So far it was discussed how economic rationing theory could be applied to the data-cap problem. But not all parts of the economic rationing theory are applicable and necessary to model data-cap tariffs. In this section it will be explained which parts of economic rationing theory will not be used and why. This description will focus on point rationing, as this kind of rationing was already identified most suitable for the problem at hand.

The first thing which is vital to point rationing [Tob52, section 5.3, page 545 ff] [Llo42] but will not be included in the data-cap model is the setting of the point prices. The reason why this important part of point rationing will not be included is that the producer in this scenario, or for that matter anybody, has only very little influence on the amount of the data price. The data price is the amount of data needed to provide the respective service. Therefore it is fixed by technical means and can only be influenced within quite narrow margins[12].

Another problem that will not be discussed here is the number of currencies [MM56, §4, page 126 ff]. The reason for that is similar as with the setting of prices. By the definition of the problem there is a money and a data price, so those are the only two currencies. Of course there would be no problem to introduce another currency but on the other hand this would go well beyond the data-cap problem. As a result the number of currencies will be fixed and money and data will be the only currencies in the model.

[10] See section 4.3.

[11] The data price is only partially a strategic variable since it is largely influenced by technical means. For more on that see the second paragraph of section 8.2.1.3.

[12] For more on that see the second paragraph of section 8.2.1.3.

Yet another question in rationing research is what the incentive to work will be under a system of rationing [Tob52, section 4, page 538ff]. While this question seems to be very legitimate under a state of general rationing it is a good example how this application differs from the standard model of rationing. General rationing could level the incentive to work since the main incentive to work is the money received in exchange which varies between jobs. It is an incentive because with more money one could buy more things. However under general rationing this no longer is true. For this application such an analogy does not hold because Internet access sure is important but not the only thing one can spend money on.

A last important feature of rationing that will not be discussed here is the possibility to buy ration currency from other individuals [Tob52, section 5.1, page 540 ff][MM56, §5, page 129ff], either for money or for another ration currency. While a black market for ration currency can undermine the goals of rationing, an official way of trading could be an interesting property of a rationing system. Either way such transactions will be probably impossible in this scenario for technical reasons. The data-cap is part of the ISP's billing and therefore it is the only one that can add or subtract data currency from a costumer's account. So far it does not seem to be in the interest of the ISP to allow such transactions. However, ISPs often justify their move from flat-rate tariffs to data-cap tariffs with fairness for consumers who only use relatively small amounts of data. Under such reasoning it seems only fair that such frugal consumers could monetize their unused data-capacity. If the ISP would allow trade the insight gained from rationing literature should be reconsidered to draw first conclusions.

8. Economic rationing models and their relationship to data-caps

So far it was determined what kind of rationing is suitable for the problem at hand and first steps how it can be applied to data-cap tariffication were searched for. As a next step an economic model built on that kind of rationing is needed to start the examination. Since no new model should be build and probably also will not be needed to build an existing model has to be found that could be utilized with the specific kind of point rationing that was identified in the last section. Besides that it needs to be defined which economic aspects should be analyzed and therefore have to be supported by the model. Since an existing model should be utilized, it should be tried to use all aspects the chosen model offers and which are applicable to the problem.
Still it needs to be specified which aspects should be analyzed in order to pick a model. The pricing of Internet access is a problem that involves three parties, the consumer and the producer, as in most classical economic problems, as well as the ISP connecting these two sides. Using a flat-rate tariff as benchmark it could be interesting how a data-cap tariff alters the welfare distribution between these three. This question seems to be one of the biggest and most important ones with a change in tariffication. It is therefore connected to other interesting effects as a shift in welfare could dramatically alter the ways in which individuals on all three sites participate in the market or even if they are in the market at all.

Besides the welfare changes a change in tariffication will also probably alter the consumption points. Since the consumer is in the center of this model this is another one of the most important questions arising from this problem. While the welfare changes are correlated to the consumption points it is important to investigate both individually to see how a change in tariffication affects this correlation.

8.1. Different rationing models to consider

Starting from the general modeling of point rationing as multiple currency system, which was already covered in its basic form, there are different approaches for economic analysis.

First there is a basic distinction between graphical and analytical models, the analytical models being the majority of the published work.

Graphical approaches covering the topic of point rationing are quite rare. Basic graphical considerations concerning rationing in general, on the other hand, are rather common and have already been presented in the section covering general aspects of rationing[1]. However, the diagrams used by Makower [Mak45] and Carter [Car48] for example are for describing the current situation rather than performing an analysis. The only real graphical analysis was done by McManus in his 1954 article ”The Geometry of Point Rationing” [McM54]. McManus's approach is quite interesting but, as most graphic approaches, it gets confusing

[1] See section 7.2.2.

real fast for an increasing number of services and more importantly lacks in detail in welfare analysis.

For the analytical models there are more choices. As already mentioned most of the published models date back to the time during and shortly after the Second World War. Later, during the 1950s and 1960s, interest in the topic of rationing was low. In the mid and late 1970s there was some new found interest in the topic. During this time and in the years up until now new research applying rationing research to other problems starts to appear. This includes, but is not limited to, credit rationing, time management research (time rationing) and rationing in health care. However in almost all cases the research focuses on the scarcity aspect of rationing rather than the multiple currency aspect. Consequently modern approaches based on rationing can not be applied to the data-cap problem[2].
To summarize there is war-rationing literature as well as the late-1970s-rationing literature left for finding a suitable model for the data-cap problem.
In the war-time-rationing literature there are no competing models, it is rather a field of research with only few published authors who build their own work on top of the work of their colleagues. Therefore there are only two models to be considered: The first one is "A survey of the theory of Rationing" by James Tobin from 1952 [Tob52]. This article was supposed to sum up rationing theory, as at that time the interest in rationing theory was diminishing as economic production picked up again and scarcity seemed to be a problem of the past. Thus in this article most insights gained up until this point are included and all noteworthy models developed so far are combined in one article. The second one is the article "Point rationing and the consumer" by M. McManus from 1956 [MM56]. This article was inspired by the additions made to rationing theory in J. S. Henderson's book "Production and Consumption" [Hen52] and it summarizes the mathematical theory of point rationing, as already done by Tobin, and amends it with new findings by Henderson and McManus. Hence this article concludes the war-time-rationing-literature.
As Tobin's article is a survey article it incorporates the work of many other authors. The parts relevant for the data-cap problem are mainly based on the work of Graaf [Gra47], Lloyd [Llo42], Makower [Mak45], Neisser [Nei43], Reder [Red42], Rothbarth [Rot41], Rothschild [Rot45], Scitovszky [dS42], Samuelson [Sam48], Tobie and Houthakker [TH50] and Worswick [Wor44]. The article covers different approaches but strongly advocates the view of rationing as a multi-currency-system. The multi currency parts of the article are highly relevant and applicable to the problem at hand. The analysis conducted in section 2 of the paper fulfills the requirements for the problem defined here, as it is capable of examining the influence of a price change of one good on another good and the welfare shift induced by rationing. Besides that, the article offers a simple, yet powerful notation for the modeling of the multi-currency-rationing-problem.
McManus's article solely focusses on point rationing and also includes the respective findings from Tobin's article. But it goes further and amends this findings with further insights.

[2] This conclusion is drawn to the best knowledge of the author. However newer research based on rationing is quite diverse and even takes place in different disciplines. Therefore there could be approaches rather promising to the problem at hand that were conducted in other disciplines or for some other reason outside the search parameters. It seems quite probable that some newer research based on rationing should be applicable to the problem but despite extensive searching no applicable approach was found.

To do so the notation for the basic multiple currency approach is adapted. The additions include an examination of the effects of a change in the number of currencies and the effects of a market for ration currency. These two additions may be quite interesting but are of no use to the problem at hand since it has a fixed number of currencies and trading data volume seems to be impossible in practice. So McManus's model offers no advantage over Tobin's for the problem at hand but has a more complex notation in order to accommodate the additional possibilities. As a result under those conditions Tobin's model should be preferred.

The newer models from the late 1970s have been the forerunner for the application of rationing to other fields. Most notably are the articles "Rationing, quantity constraints, and consumption theory" by David Howard [How77] and "The theory of household behavior under rationing" by Peter Neary and Kevin Roberts [NR80]. But much like the articles that were inspired by them they mainly focus on scarcity and do not utilize a multi currency model. Therefore they are of no direct use since their models can not be applied to the problem at hand.

8.2. Tobin's survey article about rationing and how it may apply to data-caps

The ideas presented in Tobin's first section about rationing as a multiple-currency-system were already presented here in the sections about rationing in general[3].
The second section about the behavior of an individual consumer under rationing will be the basis for the model development and therefore the subject of the following subsections.
As Tobin's article is a survey article about rationing theory as a whole it also includes analysis for straight rationing. These parts offer no insights for the model development here and therefore will not be covered.
The third section covers the collective demand under rationing. This would be of interest for a model on data-caps, however Tobin's analysis only covers the case of straight rationing. Sections 4 and 5 will not be covered as they cover parts of the rationing research that are not applicable to data cap tariffs[4]. For short these concepts are not applicable to the data-cap problem.

8.2.1. The Basic Model

8.2.1.1. The model

To be able to model a basic point rationing system it is important to understand the interaction of different currencies. That is probably why Tobin described the interaction of the different currencies at the beginning of the section about point rationing in section 2.1.

[3] See section 7.2.2.

[4] For details on that see section 7.3.2.

With this basic understanding the modeling of the constraints for a given problem can start. The first thing needed is a utility function. It is assumed that the consumer will maximize her utility subject to multiple budget constraints. In the case of data-caps the utility function is especially important as there will be many advertisement supported services which are offered for free to the consumer. Therefore they have no money price and are only restricted by their data price. When such a service becomes a managed service, the only way to distinguish different services from each other is their utility. More generally speaking, there are two budget constraints, but in many cases at least one of them is not restricting as the price is zero. This can be the case either for an advertisement supported or managed service. In these cases the utility function is even more important as a differentiator.
However, this is a problem to the model in its current state since the utility for most goods is positive. Hence when all prices are zero and the consumer maximizes her utility she will consume infinite units of that good[5]. To prevent this problem, another currency, besides the two already mentioned constraints of money and data, needs to be added. The prices in this currency need to be greater than zero in all cases, or at least in the cases where data and money price are zero.
Something that always is a natural constraint is the time spend on a given task. And time spend on a given task is always positive. So time would be a good candidate for the third constraint. Time as third constraint has the additional benefit that it is always an issue that should be taken into account when maximizing utility but usually is left out due to complexity considerations or is assumed to be implicitly included in the utility function.

With all that knowledge it is time to define the model. The first equation is the utility function. The model does not make any assumptions about the utility function.

$$U(x_1, x_2, \ldots, x_i) \tag{8.1}$$

Since there are no assumptions about the utility function it is exactly the same as in Tobin's article [Tob52, equation 2.1, page 526]. The only restriction for the choice of utility function is that it needs to be well defined. In this utility function x_i is the quantity consumed of the ith good.

In Tobin's model the maximization of the utility is subject to multiple, namely j, constraints [Tob52, equation 2.2, page 526]:

$$\sum_{i=1}^{n} p_{ji} x_i = R_j \qquad (j = 0, 1, 2, \ldots, m), (m + 1 < n). \tag{8.2}$$

In this equation there are two indices the first is i and runs from 1 to n, this is the index identifying the different goods. The second index is j, running from 0 to m, it identifies the different currencies. Hence there is a restriction of the type of equation 8.2 for every

[5] In real world scenarios utility functions usually have decreasing marginal utility which will lead to a limes for the consumption. Such an assumption would be quite reasonable, but in order to stay as close as possible to the original model it should not be made. So in order to not make any assumptions about the utility function there has to be a currency for which the price can not be zero.

one of the $m+1$ currencies. The other variables in the equation are: p_{ji}, which is the price of the ith good in the jth currency, x_i, which is the quantity consumed of the ith good, and R_j which is the ration for the jth currency. In other words R_j is the upper spending limit for the jth currency. Additionally $m+1<n$ defines that there need to be more services/goods than currencies. This makes sense because otherwise the resulting optimization problem would be insolvable in most cases due to the too large number of restrictions in relation to the variables. This also makes sense on a semantic level since it would be redundant to have more than at most one ration currency per service. For a data-cap scenario this condition should be no problem since there usually is a large number of web services and only three currencies.

This apparatus is rather simple, but as it is designed for j currencies - and the data-cap scenario only needs three - it can be further simplified. The j equations of the type 8.2 can be reduced to the following three constraints. To enable ineffective rations Samuelson [Sam48, page 166, equation 128], in contrast to Tobin, used inequations instead of equations. However, equations are needed for the Lagrange multiplier technique later used since this technique only works with equations and strict inequations. Besides in these cases inequations should be preferred for data-caps, because it is easily possible that a change in a parameter renders a ration ineffective.

$$\sum_{i=1}^{n} f_i x_i \leq F \tag{8.3}$$
$$\sum_{i=1}^{n} d_i x_i \leq D \tag{8.4}$$
$$\sum_{i=1}^{n} t_i x_i \leq T \tag{8.5}$$

In equation 8.3 p_{ji} is replaced with f_i for the money price of the ith service and R_j with F for the amount of money one can or would like to spend on web services. In equation 8.4 p_{ji} is replaced with d_i for the amount of data consumed by the ith service, its data price, and R_j with D for the data-cap. And finally in equation 8.5 p_{ji} is replaced with t_i for the time needed to consume one unit of the ith service and R_j with T for the time budget allocated to using web services.

This is the basic model. Using a Lagrange multiplier approach for maximizing the utility subject to all budget constraints produces i equations of the following form [Tob52, equation 2.3, page 526].

$$u_i - \sum_{j=0}^{m} \lambda_j * p_{ji} = 0 \qquad (i = 0, 1, 2, \ldots, n). \tag{8.6}$$

In equation 8.6 u_i is the partial derivative of equation 8.1 with respect to x_i, in other words the utility of the ith good. λ_j is a Lagrange multiplier describing the marginal utility of another unit of the jth currency. As a result λ is the shadow price of the respective currency and therefore also describes whether a ration is effective or not. When the ration

is effective λ_j is an indicator how much this ration restrains consumption. Consequently if a ration is ineffective $\lambda = 0$, in this case the equality sign in equation 8.2 will not hold [Tob52, page 526]. That seems to be intuitively as another unit of that currency has no utility because not even the available amount of units for this currency is utilized.
Again this general equation can be adjusted for the data-cap scenario. In this scenario there are still i equations, because i identifies the number of goods, but the sum in equation 8.6 can be written explicitly, due to the small number of currencies. Making this changes, the resulting equations have the following form.

$$u_i - (\lambda_f * f_i + \lambda_d * d_i + \lambda_t * t_i) = 0 \qquad (i = 1, 2, \ldots, n) \tag{8.7}$$

In this simplified form it is easier to see how this equation results from maximization. In this equation λ_f, the Lagrange multiplier for money, times f_i, the money price, plus λ_d, the Lagrange multiplier for data, times d_i, the data price plus λ_t, the Lagrange multiplier for time, times t_i, the time price is subtracted from u_i, the marginal utility of the ith good.
Hence the sum of the Lagrange multipliers in all currencies times their prices needs to be equal to the marginal utility for any good. But why must this equation hold? Since the Lagrange multipliers describe the marginal utility of income in their currencies the sum of the Lagrange multipliers times the price of a given good in the respective currency adds up to the utility of the good. This important observation was first published by Scitovsky [dS42, page 118, footnote 11] and used in all following point rationing research. When using a Lagrange multiplier approach it is always important to keep in mind that this approach is only a necessary but not a sufficient condition for an extrema. Also this approach can either yield a maximum or a minimum while only maxima are relevant in this scenario. There are also sufficient conditions for a Lagrange multiplier approach, one of them is presented in section 8.2.2.

The Lagrange multipliers offer a way to measure how the different currencies restrain consumption. They are the shadow prices for each currency. With these shadow prices it is possible to aggregate the prices in all currencies into a single currency, say money. How this is done will be shown in section 8.2.3.

8.2.1.2. Example for the basic model

The model developed so far is a full functioning model for consumer choice under a data-cap tariff, but some aspects important for such a tariff have not been mentioned so far. However every feature needed to answer the research question is already included in this model. How these features are part of the model will be explained in the next section and partially shown in the following paragraphs using an example.
The numbers in this example are not taken from real world scenarios but chosen to illustrate some phenomenons in the simplest way possible.

The first thing in the model is the utility function. In this example the utility function will be

$$U(x_1, x_2., x_3, x_4, x_5) = \\ (3x_1 - 0.25x_1^2) + (4x_2 - 0.25x_2^2) + (4x_3 - 0.25x_3^2) + (3x_4 - 0.25x_4^2) + (1x_5 - 0.25x_5^2).$$

This utility function is chosen randomly, as the model does not make any assumptions about the utility function. However a linear utility function or another unbound function would be highly unrealistic since it conflicts with Gossen's Law. In part this is fixed by introducing a time restriction, however, constant marginal utility remains unrealistic. Probably even worse, it is not possible to calculate Lagrange multipliers for unbound utility functions[6].

In this example there are 5 different services. Besides the utility function there need to be prices in all three currencies as well as the corresponding budgets.
In the model f_i is the money price, d_i the data price and t_i the time price respectively the vectors $\vec{f}$, $\vec{d}$ and $\vec{t}$ specify the money, respectively data or time prices for all available services.
the prices and budgets for the example are fixed as

$$\begin{aligned} \vec{f} &= (1, 2, 0, 2, 0) & F &= 30 \\ \vec{d} &= (2, 1, 2, 0, 0) & D &= 30 \\ \vec{t} &= (2, 2, 3, 3, 5) & T &= 60 \end{aligned}$$

With these parameters several special cases of Internet service pricing are included in this example.
Services 1 and 2 are no special case, they both have money and data prices different from zero.
Service 3 is the first special case since it has a money price of zero. Consequently it must be financed otherwise, probably by advertising. Such scenarios are already known from free newspapers or even standard Internet pricing, where such a model is rather common. The decision would normally be made only in regard to the utility, where the limited resource of time is implicitly included or modeled explicitly as a constraint. The special thing here is that there is another currency that restricts consumption and in this second currency the price is different form zero.
Service 4 again has a money price different from zero, but now the data price is equal to zero. This shows a way managed services, which have QoS-agreements with providers to be excluded from data-cap restrictions in order to have a competitive advantage, can be modeled.
With service 5 the intentions of services 3 and 4 are combined. Thus both money and data prices are zero and the decision for choosing or not choosing this service is again only dependent on the utility function and time price, as it is the case today with flat-rate tariffs and many free, mostly advertising supported, Internet services.

[6] An unbound function does not have any global extrema, however the restrictions make the function de facto bound.

All services have time prices greater than zero because it is impossible for a person to finish a task in zero time. This also is the reason why the time ration was introduced in the last section, so not all prices can be zero and therefore an infinite amount of that service would be consumed due to constant positive marginal utility.

A well-defined problem in this model has a utility function and three budget restrictions. While maximizing there are also i service-utility-equations with which the Lagrange multipliers can be computed. That results to a total of $1 + 3 + i$ equations for any problem. Hence the complete optimization problem for this example has 9 equations. The first four equations are the problem itself, consisting of the utility function and the three budget constraints.

Max

$$U(x_1, x_2., x_3, x_4, x_5) = \qquad \textit{utility function}$$
$$(3x_1 - 0.25x_1^2) + (4x_2 - 0.25x_2^2) + (4x_3 - 0.25x_3^2) + (3x_4 - 0.25x_4^2) + (1x_5 - 0.25x_5^2)$$

subject to

$$\begin{aligned} 1x_1 + 2x_2 + 0x_3 + 2x_4 + 0x_5 &\leq 30 && \textit{money restriction} \\ 2x_1 + 1x_2 + 2x_3 + 0x_4 + 0x_5 &\leq 30 && \textit{data restriction} \\ 2x_1 + 2x_2 + 3x_3 + 3x_4 + 5x_5 &\leq 60 && \textit{time restriction} \end{aligned}$$

The corresponding Lagrange multiplier approach to this optimization problem produces the 5 following equations.

$$\begin{aligned} (3 - 0.5x_1) - (\lambda_f * 1 + \lambda_d * 2 + \lambda_t * 2) &= 0 && \textit{maximization service 1} \\ (4 - 0.5x_2) - (\lambda_f * 2 + \lambda_d * 1 + \lambda_t * 2) &= 0 && \textit{maximization service 2} \\ (4 - 0.5x_3) - (\lambda_f * 0 + \lambda_d * 2 + \lambda_t * 3) &= 0 && \textit{maximization service 3} \\ (3 - 0.5x_4) - (\lambda_f * 2 + \lambda_d * 0 + \lambda_t * 3) &= 0 && \textit{maximization service 4} \\ (1 - 0.5x_5) - (\lambda_f * 0 + \lambda_d * 0 + \lambda_t * 5) &= 0 && \textit{maximization service 5} \end{aligned}$$

Solving this equations yields a consumption point for each of the five services, Lagrange multipliers for all three currencies and the utility resulting from this points. The quantities consumed are $x_1 = 4.933$, $x_2 = 7.333$, $x_3 = 6.4$, $x_4 = 5.2$ and $x_5 = 0.133$. The Lagrange multipliers are $\lambda_f = -0.08$, $\lambda_d = 0.12$ and $\lambda_t = 0.187$. With this consumption points the utility is about 48.934.
These values indicate that all three rations are effective and that an increase in the time budget would increase utility the most.

8.2.1.3. Capabilities of the model so far

The basic model is complete, but what insights can be gained from it so far? Obviously the quantities that will be consumed for given prices and budgets can be computed. Also with that knowledge consumption under a flat-rate and under a data-cap tariff can be compared. The results of such a comparison can be used for pricing data-cap tariffs in respect to flat-rate tariffs and can also be used for first comments on welfare shifts compared to flat-rate tariffs.

Incentives and willingness to pay to become a managed service

It is even more interesting to use this comparison to analyze how restrictions in the data consumption alter consumption behavior and therefore alter the market. For example, this comparison can be used for a first analysis of the incentives to become a managed service. A first approach to do that would be to compare the quantities consumed of a given service with and without a data-cap. If the restriction of data consumption leads to a significant reduction of consumption of this service it could be more profitable for this service to share some of its profits with the ISP[7] and become a managed service.
This could be true in many cases as the operation of web services always requires software which shows strong economies of scale, as long as one is not subject to a per user or per usage based licensing model [SV98, chapter 2, Pricing Information]. To be more explicit: If the software is self-made or purchased without further restrictions, it is essentially sunk cost and therefore every additional user decreases the average costs. This is one of the reasons why many web services show strong network externalities.

Especially interesting in this context could be how freemium services behave under a data-cap tariff. This kind of service tries to attract customers by offering a somehow restricted service for free and encouraging users, after they grew accustom to the service, to subscribe to the premium version of the service, which has some charges [Puj10] [Tee10, page 179]. As the name already suggests, in this business model an important aspect for attracting customers is offering a basic service for free. This aspect would not exist under a data-cap tariff since only one of the prices would be zero and the data price would be greater than zero[8]. Of course a freemium service provider could become a managed service and with that again set both prices (money and data) to zero for their basic service. However, when they do so, their costs of providing the service would rise. Since a freemium business model operates on economies of scale, most customers are free customers. These free customers are necessary to attract the few profitable customers [WBH13] [Tee10, page 179]. As a result, a data-cap could rise the costs for this business model significantly.
Hence the introduction of data-cap tariffs could make the business model of freemium web services less attractive. With the basic model different scenarios can be investigated

[7] Of course this is a huge simplification so far. In a real world scenario there would be a lot more than one ISP, which would be quite different in size and other parameters that lead to negotiation power. Also there would be transaction cost for every negotiation to become a managed service.

[8] The time price can be neglected here since it will always be greater than zero, this is true regardless of the tariff.

and therefore different business models and how they are affected by a restriction in data consumption can be analyzed.

With that technique of comparing different scenarios, it is possible to come back to the general incentives and prices for becoming a managed service.
It is assumed that not only the consumer but also the producer, in this case the provider of a web service, will maximize her utility. In this case the incentive for becoming a managed service is rather simple. The utility of a producer will usually be its profits[9], so the producer wants to maximize its profit. Consequently, the producer always becomes a managed service when it earns higher profits by doing so. Following this logic, the fares an ISP can demand from a web service provider to become a managed service can already be computed. The ISP could reap the difference in profits a web service provider experiences between a flat-rate[10] and a data-cap scenario minus a small sum for the service provider to maintain the incentive.
Of course in practice this computation is a rather hard one to make due to information asymmetry and different tariffs. So the ISP can only make an educated guess and the web service provider will try to use the information asymmetry in his favor while negotiating service charges for becoming a managed service.

Data consumption as a strategic variable

When talking about prices so far the term usually referred to money prices, as the data and time prices where assumed as fixed by technical means. For the time prices this may be true with slight variations between different individuals, since the time consumption of some tasks can be reduced by experience and training.
For the data price however there are technics to reduce it, depending on the type of data, savings could differ. A first step would be making data consumption a factor in web service design. Since today usually unlimited broadband access is assumed when designing a web service [WSC$^+$10, page 6] this could already generate substantial data savings. Another technique is compression which could be considered as implementing a more efficient data usage on a meta level.
The clear incentive to reduce the data price as far as possible is also shown by the success of software compressing data traffic for mobile phones[11].

[9] Of course this is again a huge simplification since there are many web services that are community projects and therefore follow other goals than profit maximization. However profit maximization should be the goal for every commercial website. And as for non-commercial web services it is questionable whether they even got profits that they could share with ISPs.

[10] In this case the flat-rate tariff is the benchmark because it represents unrestricted consumption and therefore the maximum consumption for the given price.

[11] Most notable are the compression Features of Google's mobile Chrome Browser and Opera's Mobile Browsers. The latest development in this trend is Opera's Max technology (`http://www.opera.com/mobile/max`, last accessed 28th May 2014, which compresses all data traffic directed to an Android device and therefore reduces the data price of the used services.

8.2.2. Influence of a change in price of a service on the consumption of another good

The Lagrange multiplier approach presented so far only is a necessary condition for a maximum for an optimization problem. A sufficient secondary condition can be defined with the Hessian matrix of the utility function. For the global maximum the Hessian matrix of the utility function has to be quadratic and negative definite under the m + 1 constraints defined by equation 8.2[12] [Sam48, page 167]. For the presented scenario this is equivalent to the m + 1 + n rowed determinant D [Tob52, page 527]:

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & p_{01} & p_{02} & \cdots & p_{0n} \\ 0 & 0 & \cdots & 0 & p_{11} & p_{12} & \cdots & p_{1n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & p_{m1} & p_{m2} & \cdots & p_{mn} \\ p_{01} & p_{11} & \cdots & p_{m1} & u_{11} & u_{12} & \cdots & u_{1n} \\ p_{02} & p_{12} & \cdots & p_{m2} & u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{0n} & p_{1n} & \cdots & p_{mn} & u_{n1} & u_{n2} & \cdots & u_{nn} \end{bmatrix}$$

Figure 8.1.: m + 1 + n rowed determinant

Using the three currencies and notation of the model developed here, the matrix will look like the following:

$$\begin{bmatrix} 0 & 0 & 0 & f_1 & f_2 & \dots & f_n \\ 0 & 0 & 0 & d_1 & d_2 & \dots & d_n \\ 0 & 0 & 0 & t_1 & t_2 & \dots & t_n \\ f_1 & d_1 & t_1 & u_{11} & u_{12} & \dots & u_{1n} \\ f_2 & d_2 & t_3 & u_{21} & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_n & d_n & t_n & u_{n1} & u_{n2} & \dots & u_{nn} \end{bmatrix}$$

Figure 8.2.: 3 + n rowed determinant relevant to the data-cap scenario

With this matrix the effect on the demand for x_i for a change in p_{jk}, the price of an arbitrary other good, compensated by change in the income R_j, as given in Tobin's equation 2.4 [Tob52, page 527], can be computed to:

$$\left(\frac{\delta x_i}{\delta p_{jk}}\right)_{comp.} = \frac{\delta x_i}{\delta p_{jk}} + x_k \frac{\delta x_i}{\delta R_j} = \frac{\lambda_j D(u_{ik})}{D} \tag{8.8}$$

However, such compensated price changes are rather irrelevant to the research question at hand since both point prices, data as well as time, are seen as exogenous variables. The

[12] Here this would be the three constraints defined by equations 8.3, 8.4 and 8.5.

data price is fixed by technical means and the time price by physical means. Therefore only the money price can be actively altered, so a service provider could see what impact a price change for his service would have on a competing service. So the only case relevant in this model would have the equation:

$$\left(\frac{\delta x_i}{\delta f_k}\right)_{comp.} = \frac{\delta x_i}{\delta f_k} + x_k \frac{\delta x_i}{\delta F} = \frac{\lambda_f D(u_{ik})}{D} \tag{8.9}$$

Hence this equation is more of practical importance and can be used to compute effects of actual price changes. On the other hand it does not offer much additional value to the theoretical model.
Nevertheless an important general insight can be derived from this equation:

$$\frac{\left(\frac{\delta x_i}{\delta p_{jk}}\right)_{comp.}}{\left(\frac{\delta x_i}{\delta p_{gk}}\right)_{comp.}} = \frac{\lambda_j}{\lambda_g} \tag{8.10}$$

In this equation it is easy to see that two commodities must have the same relationship in every currency [Tob52, page 527]. Hence two goods which are complements or substitutes for their money prices are also complement or substitutes for their data prices. This of course does not automatically imply that the relationship between two goods is the same with and without rationing [Tob52, page 527].

8.2.3. Virtual prices and their welfare implications

When additional currencies are introduced, it is an interesting question which prices in a single currency system would lead to the same consumption points. Such virtual prices were introduced by Rothbarth [Rot41] and discussed by Tobin [Tob52, section 2.2.3, page 529].
The method for calculating virtual prices using demand function derivatives was proposed by Graaff [Gra47]. The following equations describe this method as recorded by Tobin but were adapted to the notation and scenario used here. Since the virtual price is always a money price[13], it is the only currency in an unrationed scenario, it will be denoted with f_i'. λ' is the virtual marginal utility of the money income.
The starting point for the derivation of the virtual prices is equation 8.7.

$$u_i = \lambda' f_i' = \lambda_f f_i + \lambda_d d_i + \lambda_t t_i \tag{8.11}$$

Multiplying this equation with x_i and summing up for all i results in:

$$\lambda' F = \lambda_f F + \lambda_d D + \lambda_t T \tag{8.12}$$

Solving this equation for λ' and substituting in the marginal utility function results in:

$$(\lambda_f + \lambda_d \frac{D}{F} + \lambda_t \frac{T}{F}) f_i' = \lambda_f f_i + \lambda_d d_i + \lambda_t t_i \tag{8.13}$$

[13] Mathematically speaking the formula could be easily adapted for any of the currencies.

Solving this equation for f_i' will finally result in:

$$f_i' = \frac{f_i + \frac{\lambda_d}{\lambda_f} d_i + \frac{\lambda_t}{\lambda_f} t_i}{1 + \frac{\lambda_d}{\lambda_f}\frac{D}{F} + \frac{\lambda_t}{\lambda_f}\frac{T}{F}} \tag{8.14}$$

Virtual prices establish an exchange rate for all ration currencies with money. This facilitates the summing up of the money value of all currencies necessary to buy a certain good. With virtual prices it is also possible to start a first welfare analysis from the consumer's point of view. For the consumer's surplus to stay unaltered, the virtual price must be equal to the money price of the service before rationing was introduced. If the virtual price is higher or lower than the money price before the data-cap was introduced, the consumer welfare will be altered in the respective direction.

9. A first analysis based on the model

The focus so far was on the development of a general model for consumer choice under a data-cap tariff. Additionally, it was already discussed in principle what kind of analysis would be possible with this model. The objective of this chapter is to use this model to derive some first trends resulting from the introduction of a data-cap tariff with the help of the newly developed model.

9.1. Framework

For the following analysis the complexity should be reduced as far as possible while still being able to show general trends. Consequently, there needs to be a focus on the features essential to data-cap tariffs, especially when they are in contrast to flat-rate tariffs.

Utility function

The difference between a flat-rate and a data-cap tariff is the number of constraints and the effects of those. The utility function should be as neutral as possible in order to leave the decision about the consumption set to the constraints. The simplest way to achieve that is that each service has the same utility.
What kind of utility function is used does not matter to the model, so the simplest form possible should be chosen. However, when choosing a linear utility function, just the first service with the lowest price or lowest combined price would be chosen in most cases. To prevent that, an utility function with the same decreasing marginal utility for every service should be used. One of the simplest utility functions of that kind is the utility function of the form $u(x) = \alpha * x - x^2$. The parameter α allows to adjust the absorption point. An absorption point can be used to prevent the just mentioned concentration due to a greedy strategy. The framework used here will set the absorption point at 50% of the lowest cap.

Constraints

A first simplification will be that from now on only data and time will be considered as constraints. Since every constraint adds additional complexity the reduction of the number of constraints leads to simplification. This simplification should not influence the trends since money is a constraint for the data-cap case as well as the flat-rate case. On a semantic level this design can be justified with the fact that the predominant business model for consumer web services still is to offer the service to the consumer for free and generate revenue elsewhere, in most cases through advertisement.
The data restriction is essential to the analysis since the addition of the data restriction is the mathematical distinction between a flat-rate tariff and a data-cap tariff.
The time restriction should also be maintained since it allows to use all kinds of utility functions and to model an outside good, which can be useful in some scenarios.
Furthermore a scenario with a time and a data constraint is mathematically equivalent to a scenario with a data and a money constraint. The only mathematical difference between

this two two-constraint-scenarios are the prices which will probably be different. So there might be a problem to analyze scenarios with advertisement supported services, for which the price in the second currency needs to be zero, which can never be the case for time prices. For this scenario the second currency needs to be interpreted as money constraint if the utility function allows such a move or the third constraint needs to be reintroduced.

The rations and prices need to be designed in such a way that all rations are effective. This is necessary because the effect of rationing on consumption should be investigated and because ineffective rations basically are a free market scenario. Despite this a ration could become ineffective due to some changes, for example turning a service into a managed service.

Services

In a real world scenario there are a large number of web services. However, since in this model there are only three parameters for a web service, this large number can be reduced to only a few cases.
The first parameter is the utility, which is an important characteristic of the service but for the reasons just discussed all services in this framework will have the same utility.
Hence the services will be defined by their time and data prices, therefore each service can be represented as a point in the two dimensional service map shown in figure 9.1. As already indicated in the service map, the services can be categorized further. It should be

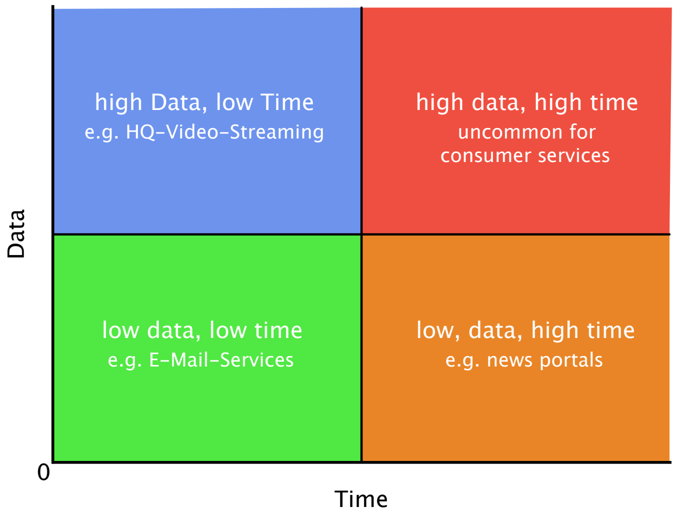


Figure 9.1.: Different kind of services in the analytical framework.

sufficient to define the categories low and high for both dimensions resulting in a total of four services.
Services with a low time and a low data consumption for example can be services for fast information exchange, such as instant messaging and e-mail.
Services with a low data consumption but a high time consumption would be services that mainly transfer text but need a lot of attention. Examples would be some news sites and e-papers.

Services which consume large amounts of data but a relatively small amount of time could be for example all services that transfer high definition images such as HD video streaming and some kinds of online gaming.
The last case with high data and high time consumption does not exist in a consumer setting. The examples are sure debatable. Especially what high and low data consumption is might differ strongly between consumers. This is partly the reasoning behind the IPSs tariff models which offer different connection speeds. So 100 MB of data might be a high volume of data for a consumer with a basic broadband connection with a connection speed of 1 MB/s while it will be a small amount of data to a consumer with a 100 MB/s connection. However, this should not influence the general trends.
Beside these three goods that lie within the service framework it will sometimes be useful to introduce an outside good. Especially in a scenario where there is a data-cap but no managed service or top-ups it can be interesting to see how an outside good behaves in relation to the other goods.

Optimization problem using the framework

Max

$$U(x_1, x_2, x_3, x_4) = \alpha(x_1 + x_2 + x_3 + x_4) - (x_1^2 + x_2^2 + x_3^2 + x_4^2) \tag{9.1}$$

subject to

$$d_l x_1 + d_l x_2 + d_h x_3 + 0x_4 \leq D \tag{9.2}$$
$$t_l x_1 + t_h x_2 + t_l x_3 + 1x_4 \leq T \tag{9.3}$$
$$d_l < d_h \tag{9.4}$$
$$t_l < t_h \tag{9.5}$$

The utility function is neutral since every service has the same utility. d_l and d_h denote low, respectively high data usage analog t_l and t_h denote low and high time prices.
x_4 is the outside good, so it represents all services which do not consume data. The reason to use outside goods is to have an alternative to the services which are studied. When managed services are introduced, it does not make sense to introduce an outside good. The reason behind that is that managed services, which are part of the web services studied, will also have a data price of zero. Consequently a managed service will be indistinguishable from the outside good in this model. This will not even change when the time price is reintroduced since an outside good will also have a time price bigger than zero.

So far high consumption, of data or time, was only defined as being greater than low consumption. But that might not be enough in many cases. It could be that high consumption needs to be greater by at least some magnitude to get a realistic picture.

9.2. A first analysis

To be able to numerically calculate trends, high and low will be identified with a magnitude differentiating them and the caps will be set to a reasonable size in regard to this numbers henceforth. For the start both caps will be set to 100, so $T = 100$ and $D = 100$. Since the absorption point should be 50% of the lowest cap, it needs to be 50, hence $\alpha = 50$. The prices are defined so that a high price is 5 times the low price, resulting in $t_l = d_l = 1$ and $t_h = d_h = 5$. This also leads to services with low consumptions being numéraire goods in their respective currencies. The outside good should be a numéraire good as well and therefore have its price set to 1 as well. It is important to mention that the price that is important for the maximization of the utility is the aggregated price. As a result, the numéraire property in a single currency might not be that important. Such a definition also offers the advantage that a numeral solution to the problems can be computed.

Again it needs to be emphasized that with the introduction of numbers the general trends will not be altered. However the ratio introduced with the numbers is another assumption that is artificially introduced. Therefore, it does not have to be the correct ratio for many cases. When generalizing, the results it is important to remember that this ratio was artificially introduced.

9.2.1. Introduction of a data-cap

9.2.1.1. Situation under a flat-rate tariff

The benchmark for the data-cap is the flat-rate. How will a flat-rate tariff look like in this framework and which utility and consumption points will it yield?
The optimization problem is easy to formulate and will look like this:

Max

$$U(x_1, x_2, x_3, x_4) = (\alpha x_1 - x_1^2) + (\alpha x_2 - x_2^2) + (\alpha x_3 - x_3^2) + (\alpha x_4 - x_4^2) \tag{9.6}$$

subject to

$$1x_1 + 5x_2 + 1x_3 + 1x_4 \leq 100 \qquad \textit{Time} \tag{9.7}$$

The obvious outcome of this optimization is that the same amount of x_1, x_3 and x_4 will be consumed, since they have the same price and same utility. Less of x_2 will be consumed since its price is substantially higher and therefore purchasing x_2 will only maximize utility after the marginal utility of the other three services has already dropped substantially.

Solving this optimization problem results in a utility of about 2143 and $x_1 = x_3 = x_4 \approx 21.43$ and $x_2 \approx 7.14$.
Hence the time ration is effective because otherwise the consumption of each service would

be at the absorption point α. If the utility function allows it[1] a single ration will always be effective, since the consumer wants to consume as much as possible to maximize her utility and a rationing restriction can only limit this consumption. Therefore, as long as there is a marginal utility greater than zero, the consumer will keep consuming until the restriction prevents her from consuming another unit.
This is important to keep in mind as the utility function must not limit consumption. If that is the case, it will not be possible to compare the effects of rationing as consumption is already restrained trough the utility function. For example, if α is changed to 25, the restriction will become ineffective and the same amount of each good will be consumed. This is in fact true for any $\alpha \leq 25$ in this scenario.

Another way for the restriction to become ineffective is if the ration is too high. With the neutral utility function with decreasing marginal utility used here the parameter α fixes the amount which can be consumed at most from each service. So if the ration is higher than or equal to $n * \alpha$ the consumption points will change to $x_i = \alpha$, for $i = 1, \ldots, n$.
Of course the observations concerning effective and ineffective rations become more complicated when additional restrictions are enforced but they will still follow the same logic. In the further analysis it will be important to stay within parameters that lead to effective rationing in order to be able to do an analysis on rationing. However, if a change in any parameter leads to ineffective rationing this also constitutes an important insight.

9.2.1.2. The introduction of the data-cap

How will the parameters change when the data-cap is introduced as a second constraint?

$$1x_1 + 1x_2 + 5x_3 + 0x_4 \leq 100 \qquad\qquad Data \tag{9.8}$$

The consumer will now try to optimize her utility under the combined price for data and time while fulfilling the budget condition.

Situation without managed services

x_4 does not use any data; consequently, it not only is the outside good but also indistinguishable from a managed service in this scenario. So in order to have a simple case with no managed service, for now only goods 1 to 3 will be regarded.
Therefore the problem reduces to:

Max

$$U(x_1, x_2, x_3) = (\alpha x_1 - x_1^2) + (\alpha x_2 - x_2^2) + (\alpha x_3 - x_3^2) \tag{9.9}$$

[1] It seems rather unlikely that the utility function would restrain consumption more than a restriction, because in this scenario the introduction of the constraint would we arbitrary. However it still is a possibility.

subject to

$$1x_1 + 5x_2 + 1x_3 \leq 100 \qquad Time \qquad (9.10)$$

$$1x_1 + 1x_2 + 5x_3 \leq 100 \qquad Data \qquad (9.11)$$

First the optimization for the flat-rate case has to be done again with the modified problem. Obviously still the same amount of x_1 and x_3 and a lower amount of x_2 will be consumed. Since there still is the same budget but the number of goods reduced to three more of each service will be consumed. Nevertheless, the utility will be lower than in the four goods case because of decreasing marginal utility. The concrete values are a utility of about 1666.67 and $x_1 = x_3 \approx 22.22$ and $x_2 \approx 11.11$.
When the data restriction is added those values change to a utility of about 1578.95 and $x_1 \approx 21.05$ and $x_2 = x_3 \approx 13.16$.
The fall in utility makes sense since every additional restriction can only further reduce the possible consumption points. So if positive marginal utility[2] is assumed, an additional restriction can at best leave the utility unchanged and therefore will reduce it in many cases.
It is already a noteworthy insight that the shift from a flat-rate to a data-cap tariff will decrease the consumer's utility in most cases. The consequence in a real world tariff would be that the price of a data-cap tariff must be lower than that of a flat-rate tariff in all cases. If this was not the case a rational consumer would always choose the flat-rate tariff[3].
The change in the consumption points can also easily be explained with the second restriction. The combined prices[4] are 1 for service 1 and 3 for services 2 and 3. According to this modified weights the consumption points already determined make sense since service 1 has the lowest combined price and services 2 and 3 the same higher combined price. It is important to note that the combined price in this case can be simply computed by averaging both prices only because both caps are the same size and the prices are the same, just in different order. In more complex cases it is harder to say at first glance how the different restrictions affect the outcome of the optimization problem. To solve them, the Lagrange multiplier approach or some numerical method to solve the problem must be used. Or the combined price can be computed with the method described in section 8.2.3.

Situation with managed service / outside good

How does the consumption points change when there is a service with no data costs? To investigate that the original problem with 4 services must be examined. Now the utility has sunk to about 2087 and the consumption bundle is $x_1 \approx 20.47$, $x_2 \approx 8.66$, $x_3 \approx 14.17$

[2] This is a reasonable assumption in this scenario since if a good has negative or no marginal utility, the consumer will not consume another unit of this good under free market conditions. This will only change if the consumer is forced to buy some goods even if they have negative marginal utility. Such situations can exist for example in other rationing systems [NR80].

[3] This of course assumes that the ration would be effective, however most ISPs argue that data-caps would not affect most customers, hence the ration would be ineffective.

[4] The combined prices here are a special case of the virtual price.

and $x_4 \approx 22.05$. Every value in the consumption set has changed and the utility has fallen compared to the flat-rate case. At these consumption points the time as well as the data ration are effective, consequently for both restrictions the equality sign holds. Further the trends already observed without a managed service can be confirmed. The usage for the data intensive services has declined while the time intensive service sees an uptake because it is data economical and so the combined price is relatively lower than the time price alone. There also is a slight decline in usage for service 1 despite the fact that it is both low in time and data usage. This decline is in favor of the outside good whose combined price is even lower. In this scenario the managed service is to be seen as an outside good, because it is none of the 3 types of services defined in the framework. It is interesting to see that the outside good has a higher consumption point than any service in the uncapped scenario. However, this can easily be explained because its virtual price is lower than for any service in the uncapped scenario.

9.2.1.3. Becoming a managed service

So far there was no competition. However, it is essential to see how becoming a managed service affects a (direct) competitor. To analyze such scenarios service 4 is eliminated again and the different scenarios where one or more of the three standard type services becomes a managed service are each individually computed and interpreted.

To analyze what effect it has on the other services when a service becomes a managed service, the framework introduced earlier is utilized again. So the optimization problem consists of the utility function from equation 9.9 and the time restriction from equation 9.10 as well as one of the following data restrictions. The result will again be compared to the flat-rate case.

$$0x_1 + 1x_2 + 5x_3 \leq 100 \qquad \textit{Service 1 managed service} \qquad (9.12)$$

$$1x_1 + 0x_2 + 5x_3 \leq 100 \qquad \textit{Service 2 managed service} \qquad (9.13)$$

$$1x_1 + 1x_2 + 0x_3 \leq 100 \qquad \textit{Service 3 managed service} \qquad (9.14)$$

$$0x_1 + 0x_2 + 5x_3 \leq 100 \qquad \textit{Services 1 and 2 managed service} \qquad (9.15)$$

$$0x_1 + 1x_2 + 0x_3 \leq 100 \qquad \textit{Services 1 and 3 managed service} \qquad (9.16)$$

$$1x_1 + 0x_2 + 0x_3 \leq 100 \qquad \textit{Services 2 and 3 managed service} \qquad (9.17)$$

$$0x_1 + 0x_2 + 0x_3 \leq 100 \qquad \textit{All services managed service} \qquad (9.18)$$

Equation 9.18 can be ruled out at first glance, because when all services in a market are managed services it is equal to a flat-rate from a customer's perspective. Equations 9.12 and 9.13 respectively 9.16 and 9.17 seem to be redundant, but that is only as long as the time prices are ignored.

As a reminder here again the consumption points of the flat-rate case: utility of about 1666.67 and $x_1 = x_3 \approx 22.22$ and $x_2 \approx 11.11$. In contrast the consumption points for the data-cap tariff without managed services are $x_1 \approx 21.05$ and $x_2 = x_3 \approx 13.16$, the utility is about 1578.95.

Making service 1 a managed service

Service 1 is a service with a low time and a low data price. At first glance it does not seem to be necessary for this service to become a managed service. However, since becoming a managed service will reduce its virtual price further it will still increase its usage. Restriction 9.12 yields an utility of about 1644.52 and $x_1 \approx 22.59$, $x_2 \approx 11.96$ and $x_3 \approx 17.61$. So becoming a managed service led to an uptake in consumption of service 1. It is notable that the consumption of service 1 now is even higher than in the flat-rate case. This phenomenon can be explained with the fact that the virtual price as managed service is even lower than the flat-rate price. To compensate for the increase in consumption of service 1 there is a decrease in consumption of service 2. Interestingly there also is a notable uptake in consumption of service 3. This is a compensation effect: The decreased consumption for service 2 and the elimination of the data price for service 1 freed up a notably amount of data currency. This way data-capacity became less scarce and the high data price of service 3 has less influence on its virtual price. So making a service with low data consumption a managed service can actually lead to increased consumption of services with high data consumption as long as they have low prices in the other currencies.
Another observation that can be made is an uptake in utility. Since a managed service frees up data currency it seems only logical that it will increase overall consumption and therefore utility under the assumptions made.

Making service 2 a managed service

Service 2 in the framework has a high time price but a low data price. After the findings for service 1 it will be interesting to see how the elimination of the data price will affect the only service with a high time price. Restriction 9.13 yields the following values: a utility of about 1621.62, $x_1 \approx 21.17$, $x_2 \approx 12.61$ and $x_3 \approx 15.77$. Again there is an increase in utility compared to the pure data-cap scenario but it still is lower than in the flat-rate case. Most notably there actually is a decrease in consumption of service 2 while there is an increase in consumption of services 1 and 3. This can be explained by its high time price and the low time prices of the other services. Making service 2 a managed service frees up data currency, which can then be spent on services with low time prices because of the equally distributed utility. The main reason behind that is that by making service 2 a managed service data currency became less scarce than the time currency while they were equally scarce before. This way the time prices gained in influence on the optimization problem and the high time price of the managed service backfires.
So there are scenarios where becoming a managed service can actually decrease the consumption of the managed service.

Making service 3 a managed service

Naturally, service 3 seems to be the best candidate to become a managed service since it is the only service in the framework with a high data consumption.
Using restriction 9.14 yields a utility about 1666.67 and $x_1 = x_3 \approx 22.22$ and $x_2 \approx 11.11$. Hence this optimization problem yields the same values as the flat-rate case. This is

because making service 3 a managed service frees up enough data currency to make the data restriction ineffective. Of course this will not always happen when a service with high data consumption becomes a managed service. However, this illustrates a phenomenon that could easily arise when managed services are introduced and data-caps are kept constant. In such a case it is quite possible that the data ration becomes ineffective. Part of this phenomenon is the observation from the last paragraph that under such conditions becoming a managed service can even lower the consumption of the respective service and increase consumption of some other services.

If the ration does not become ineffective, service 3 will strongly profit from becoming managed service. But the other services will also profit largely as even if the data restriction does not become ineffective, large amounts of data currency will be freed up. This can be seen for example when in the given scenario the data-cap is reduced to 20 and everything else is kept constant.

Making services 1 and 2 managed services

Following the observations made in the last two paragraphs it seems most probable that turning two services out of three into managed services will render the data restriction ineffective. However, since services 1 and 2 both have a low data consumption, this must not be necessarily true. But while the data ration may not become ineffective, the absolute value of λ_d will surely drop and move closer to 0.
Using restriction 9.15 yields a utility of about 1661.54 and $x_1 \approx 22.31$, $x_2 \approx 11.54$ and $x_3 = 20$. This values are close to the values that result from the flat-rate case. Hence the presumption that λ_d will fall holds and making service 1 and 2 managed services almost renders the data restriction ineffective. In this scenario service 3 profits strongly from service 1 and 2 becoming managed services while service 1 is relatively unaffected since it only saw a relatively small decrease through the introduction of the data-cap. Service 2 also profits strongly because its consumption point dropped significantly with the introduction of the data-cap. While this insight again strongly relies on the values set for this analysis the trend is independent of that: The more services become managed services, the more data currency is freed up, which will eventually lead to the data restriction becoming ineffective. If services with high data consumption become managed services, this process will of course need fewer services.

Making services 1 and 3 managed services

With the observations of the last paragraph it seems certain that whenever service 3 and another service become managed services more than enough data currency is freed up to render the data restriction ineffective. And therefore returning to the flat-rate scenario.
Using restriction 9.16 yields a utility about 1666.67 and $x_1 = x_3 \approx 22.22$ and $x_2 \approx 11.11$. Hence the presumption is right and the data restriction becomes ineffective and the consumption is back to flat-rate levels.

Making services 2 and 3 managed services

With the knowledge of the last paragraphs it is certain that the data constraint will become ineffective in this scenario.
Using restriction 9.17 yields a utility of 1666.67 and $x_1 = x_3 \approx 22.22$ and $x_2 \approx 11.11$. Hence the presumption was right again and the data constraint is now ineffective.

Effects of managed services

The first insight is not surprising but important to restate: Whenever a managed service is introduced the utility for the consumer will rise compared to the pure data-cap scenario. However utility has still to be lower than in the flat-rate scenario, except when the managed services renders the data restriction ineffective and therefore restores the flat-rate scenario.
The next insight is that making a service with low data usage a managed service can actually lead to an increase in consumption of another service with high data consumption. The reason for that is that the service with the high data consumption might offer high marginal utility, but further consumption of this service is not possible until further data currency is freed up. That can only happen through making another service, even one with low data consumption, a managed service. More generally speaking, making a service a managed service will always free up data currency and therefore might also help competitors.
Not only can becoming a managed service benefit another services, it can actually decrease consumption of the service becoming managed. This will be the case if the service becoming managed has a low data price but a high price in the other currencies. The reason behind that is that the managed service frees up data currency that could be spend on other services. And as more data currency is available to distribute, λ_d will drop while the other currencies compensate for that and see a rise in their Lagrange multipliers. As a result, the influence of these prices on the aggregated price increases.
Turning some services into managed services while maintaining the same data-cap as before will always help the unmanaged services since it frees up data currency which should be a scarce good under a data-cap tariff. If enough data currency is freed up, the data restriction will eventually become ineffective. After the data restriction has become ineffective, it is a flat-rate scenario again. If services with high data consumption become managed services, it will free up more data and therefore lead to a de facto uncapped scenario faster if the cap remains unchanged.

9.2.2. Data-caps and service competition

The analysis in the last section looked at the effects the introduction of a data-cap and managed services would have on consumption. The analysis was conducted using three prototypic services. Therefore, the research design so far had no competition implemented. Nevertheless, it is especially interesting how a competing service will be affected when a service becomes a managed service.

9.2.2.1. Expanded framework

To be able to examine such a scenario the same framework can be used but it needs to be expanded to include two competing providers for each service. To model that, equations 9.9 to 9.11 need to be expanded to 6 services.

Max

$$U(x_1, x_2, x_3, x_4, x_5, x_6) = \sum_{i=1}^{6} (\alpha x_i - x_i^2) \tag{9.19}$$

subject to

$$1x_1 + 5x_2 + 1x_3 + 1_4 + 5x_5 + 1x_6 \leq 200 \qquad \textit{Time} \tag{9.20}$$

$$1x_1 + 1x_2 + 5x_3 + 1x_4 + 1x_5 + 5x_6 \leq 200 \qquad \textit{Data} \tag{9.21}$$

In this set of equations services 1 and 4, 2 and 5 as well as 3 and 6 are the same services just from different providers. Therefore they are identical in any parameter, until one of them becomes a managed service. Alpha will still be 50 in order to see the same utility for each service as in the three-services-case. For the same reason both rations are increased to 200, so the consumer will consume the same amount of every service as in the three-service-case for a flat-rate or pure data-cap scenario.

9.2.2.2. Competition analysis

Flat-rate and pure data-cap scenario

For the flat-rate and the pure data-cap scenarios there are no changes since they are exactly the same, just expanded.
The flat-rate scenario yields a utility of about 3333 and $x_1 = x_4 \approx 22.22$ and $x_2 = x_3 = x_5 = x_6 \approx 11.11$. After introducing the data-cap utility is about 3157.90 and $x_1 = x_4 \approx 21.05$ and $x_2 = x_3 = x_5 = x_6 \approx 13.16$.

Simple competition scenarios

The simplest question under this competition scenario is how it would affect a service if its direct competitor were to become a managed service.
Regarding that the first scenario would be how service 4 would be affected when service 1 becomes a managed service. The maximization for this problem yields a utility of about 3234 and the following values $x_1 \approx 22.80$, $x_2 = x_5 \approx 12.52$, $x_3 = x_6 \approx 15.37$ and $x_4 \approx 21.31$. The utility is quite analogous to the scenarios examined in the last section. Services 2 and 5 as well as 3 and 6 yield the same values. Service 1 has increased in consumption and is now above the level in the previous section due to a larger time budget to redistribute. For the same reason it is even above the value in the flat-rate scenario. Since services 1 and 4 have a low data and a low time price, their combined

price also is relatively low and therefore when service 1 becomes a managed service some of the freed up data currency will be spent on service 4. Also services 2 and 3 and with them their identical competing services 5 and 6 see compensation effects like in the last section. But as some of the freed data currency is already spent on service 4 the increase in consumption for services 3 and 6 is relatively small. To compensate the increased time consumption of the so far mentioned services the consumption of services 2 and 5, with their high time prices, has to fall.
Hence an important insight gained so far is that becoming a managed service for a service with low prices in both currencies will not only benefit this service but also the direct competitor.

When service 2 becomes a managed service it yields a utility of about 3206 and $x_1 = x_4 \approx 21.10$, $x_2 \approx 13.68$, $x_3 = x_6 \approx 14.57$ and $x_5 \approx 12.05$. So the results are quite analogous with the important exception that consumption of the direct competitor service does not increase but rather decreases. The reason behind that is its high time price. Since making service 2 a managed service makes data less scarce and therefore time more scarce[5] the high time price of service 5 leads to a reduction in consumption to maximize utility. The same logic is behind the increase in consumption of services 3 and 6 and the relatively strong increase for services 1 and 4. The strong increase in consumption of services 1 and 4 is despite their high time price since time is not that scarce in this scenario because the managed service has a high time and a low data price. The high time price for services 2 and 5 does not only lead to the decrease of consumption of service 5, but makes the increase, due to becoming a managed service, for service 2 relatively small. In the monopolistic scenario discussed earlier this even led to a decrease in consumption. This is not the case here since with the doubling of the number of services the data-cap was also doubled and this in turn led to the currencies being less scarce.
What can be taken from this scenario is that if a service has a low data price and a high price in the other currency, it will only profit little from becoming a managed service. However, it will profit indirectly since becoming a managed service will increase the shadow prices of the other currency[6] and therefore hurt its direct competitor.

So what will happen if service 3, the only service in the framework with a high data price, becomes a managed service? It yields a utility of about 3333. That is the utility of the flat-rate scenario. So making service 3 a managed service has made the data restriction ineffective, as already observed in the non-competing-scenario. While this will not happen every time a service with a high data usage becomes a managed service, it always is true that the higher the data consumption of a service is the more data currency is freed up when it becomes a managed service. Every non-managed service available will profit from this additionaly available data currency , even the direct competitor.
An important insight to take from this is that a data intensive service could easily make the data ration ineffective by becoming a managed service and therefore again level the playing

[5] This direct correlation of course will only hold as long as both rations are effective.

[6] This only is true as long as all rations are effective. Only the shadow prices of effective rations are correlated.

field for direct competitors and other services. This again emphasizes the importance of the size of the data-cap as a strategic variable of the ISP.
In order to examine how service 6 is affected when service 3 becomes a managed service the data-cap must be drastically reduced while everything else is kept constant. When this is done, the results are in line with the findings so far. The compensation effects will lead to a very slight increase for services 1 and 4 and a decrease for services 2 and 5. Service 3 will profit tremendously from becoming a managed service while its competitor, service 6, also profits but on a much smaller scale but still relatively more than the other services[7].

Lessons learned

How becoming a managed service affects direct competitors is related to the own prices in all currencies. These prices, and not only the data price, do also influence how much a service will profit from becoming a managed service. A service with a high data price will profit more from becoming a managed service than a service with a low data price. If the service to become a managed service has a high price in a currency other than data the increase would be smaller than if it had low prices in all other currencies[8]. Another important insight is that by becoming a managed service the shadow price for data is lowered, which will increase consumption of the other services, in most scenarios even that of direct competitors.
Again it is important to note that these effects only hold as long as the size of the data-cap is not changed and all rations are effective. If a ration is not effective, its price has no influence on the optimization.
Consider a real life example with money and data. If money is no issue it will not restrict the optimization and therefore, when a service becomes a managed service the consumer will consume the maximum amount of that service which is supported by the utility function. Also the effect of the data currency freed up by the managed service can be compensated by adjusting the size of the data-cap[9].

9.2.3. The influence of cap size

So far the size of the data-cap was assumed to be constant in order to be able to conduct an analysis of the changes in quantities consumed by the individual services. However, the size of the data-cap is one of the few strategic variables an ISP has. Hence it is worth analyzing how changes in the size of the data-cap affect the consumer.

Some basic insights about the importance of the size of the data-cap where already gathered. Up to this point the data-cap was always assumed to be effective. The notion of effectiveness was taken from Tobin's rationing model and means that the restriction must

[7] The shifts in this scenario are tremendous. The scenario is as given for this section except the data-cap is reduced to 100. The shifts compared to the pure data-cap scenario are +1% for services 1 and 4, -16% for service 2 and 5, +685% for service 3 and +130% for service 6.

[8] This again assumes that all rations are effective. The scarcer the currency is in which a service has a high price the less it will profit from becoming a managed service.

[9] However, it is questionable how frequent changes of the size of the data-cap could be realized in a real world scenario considering contractual obligations and regulatory oversight.

be binding. This makes sense as a restriction that is not binding is no restriction at all. With this knowledge, the possible size of the data-cap can already be tremendously narrowed down. Furthermore, a negative data-cap does not make any sense since there is no such thing as negative data. Also to be able to consume anything, the data-cap must be greater than zero. So the lower, not included, bound for a data-cap is zero. The upper bound is the biggest amount of data for which the data-cap is still effective. Any bigger cap is perceived as the same since the data ration has become ineffective and therefore is not even noticed by the consumer. Consequently, this situation is no different from a flat-rate tariff in the eyes of the consumer[10].

Whether a given data-cap constitutes an effective ration or not depends on the prices of all services in all currencies and the utility function as well as the rations in all currencies. For effective rations those factors also determine the shadow prices for all currencies as discussed in section 8.2.1.1. Because of this interdependency general statements are rather hard to make.
The only general statement the can be made beyond what was already mentioned is that if all rations are binding and the size of one drops, the influence of the price in this currency on the quantities consumed will increase. Because of this effect the decrease of one ration can render other rations ineffective.

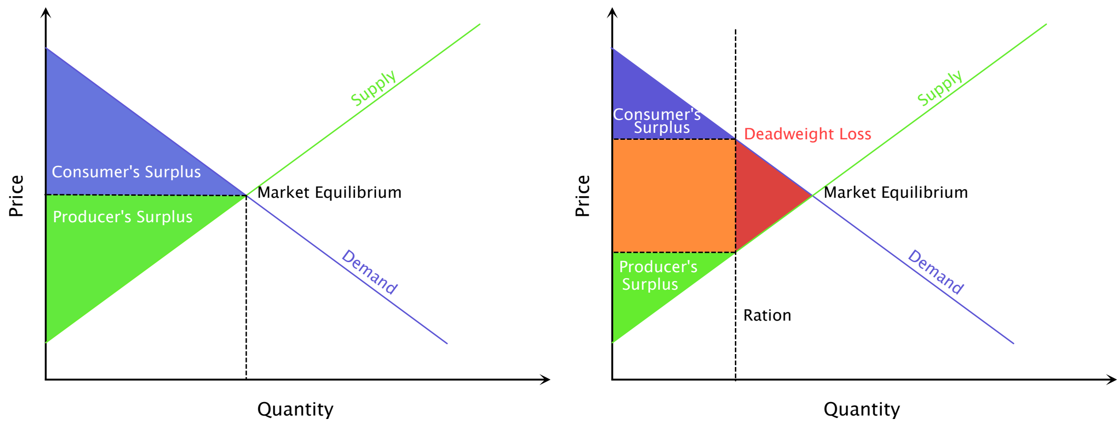
9.2.4. Welfare effects of data-caps

9.2.4.1. General welfare effects

To compute exact welfare effects is again impossible without a concrete scenario or at least some assumptions about the relation of the different variables. However, as only the consumer's side of the problem is investigated here, the producer's surplus can not be computed anyways. Hence the variable of interest is the consumer's surplus. Its value depends on the difference between the reservation price and the actual price spend on a good. This however again depends on supply and demand. Since another constraint can only further restrict consumption or at best leave it unaffected, the welfare effects of rationing and consequently of data-caps can at best be neutral. As a result the consumer's surplus will be unchanged whenever the data ration is ineffective, leaving it at the flat-rate stage. If the data ration is effective, it will restrict the supply and as a result there will be a deadweight loss.
Such effects were studied in basic economic welfare theory for simple rationing [Var10, page 265] [CJ09, pages 495 ff]. The effect of quantity rationing on the standard free market welfare distribution can be observed in figure 9.2.

In this figure it is clear to see that an effective ration will always lead to a deadweight loss. Another thing to see is that in such a scenario there is no market determined equilibrium, but the price is subject to bargaining. The price can be any point on the ration line restricted by the demand and supply curves. Hence the orange area in figure 9.2 (b) will be distributed according to the bargained price. Yet it is questionable whether these

[10] As mentioned earlier it still might make a difference on a psychological level.



(a) Standard welfare model with demand and supply curve.
(b) Changed welfare distribution due to rationing.

Figure 9.2.: Welfare effects due to increases in prices induced by rationing.

insights can be directly applied to a data-cap tariff coming from a flat-rate tariff, which induces a completely different supply and demand structure. On the other hand this is a good example of how additional restrictions produce a deadweight loss. While the mathematic structure for the loss is different for the switch from a flat-rate to a data-cap tariff, the principle is the same.

The welfare distribution as depicted in figure 9.2 assumes the standard case of pay-per-use. For the flat-rate case not only the pricing but also with it the welfare distribution is different. The supply curve in the flat-rate case is a horizontal line since the consumer gets unlimited data for a fixed price. However, due to flat-rate effects, that does not mean that the producer's surplus is zero because some consumers are willing to pay an irrationally high price. For a data-cap the situation is quite similar with the difference that the supply curve has a defined end point at the quantity of the cap. This should benefit the aggregated producer's surplus since some consumers get their consumer's surplus from extensive consumption which will be restricted in this case.

9.2.4.2. Virtual prices and resulting price changes

A basic concept to compute the welfare effects of rationing are virtual prices. A virtual price in this context is the money price that produces the same consumption as the rationed problem. How virtual prices are computed was discussed in section 8.2.3. Since virtual prices mostly depend on the prices of the different services offered, it is rather hard to make general statements. To derive some general trends, the framework introduced earlier can also be used here. As a small modification the time price will here be used as money price[11].

The results are loosely correlated with the changes in the consumption points. The correlation works in such a way that the virtual price compares to the original money price

[11] As already mentioned in the virtual price section there is no mathematical difference between the different restrictions with the only exception that the time price can never be zero.

as the new consumption point of a service does to the original one. If the consumption drops due to the introduction of a data-cap, the virtual price of the respective good will rise accordingly. On the other hand a rise in consumption means a decrease of the virtual price. This trend holds in almost all cases but not in all since this observation neglects the compensation effects. Because of the compensation effects of the utility function it is also hard to see a correlation in magnitude.

In the simplest case, such as discussed in section 9.2.1.2, the virtual price is the average of money and data price. Looking at the formula for the virtual price simple averaging of the two prices can be used when both Lagrange multipliers and both rations are equal. In this case every rise in prices is compensated for by a price drop and vice versa. Regarding that one could think that in such a scenario the welfare effects would be neutral[12]. However, the data-cap does still restrict consumption and therefore despite the neutral price changes the overall consumption and with it the welfare drops.

General statements for more complex scenarios[13] are even harder to make. The only thing that can be said for sure is the basic correlation described in the second paragraph of this section. To illustrate that, it is helpful to compute the virtual prices for the basic example from section 9.2.1.3. Because service 1 has become a managed service, while both caps are still the same size, it must be $\lambda_f \neq \lambda_d$. Making the necessary computations will yield $\lambda_f \approx 4.817$ and $\lambda_d \approx 1.993$. So the money ration is notably more restricting than the data ration. This makes sense since this example comes from a scenario with equally restricting rations where the data price for one service has been reduced to 0. With this Lagrange multipliers the virtual prices are $f_1' \approx 0.707$, $f_2' \approx 3.829$ and $f_3' \approx 1.463$. This results in a price change of -29.300% for service 1, -23.420% for service 2 and +46.300% for service 3. The new consumption points resulting from the introduction of this data-cap with the managed service, as compared to the flat-rate case, produce changes of + 1.661% for service 1, +7.641% for service 2 and -20.763% for service 3. Looking at those percentages the correlation regarding the general trend is evident. However, a correlation regarding the magnitude can not be found because of the compensation effects.

[12] This thought only makes sense as long as the utility function is neutral.

[13] For example with more than two currencies, caps with different sizes or unequally restricting caps.

10. Conclusion

10.1. Summary

This thesis is divided into two parts. The first part tried to analyze the concept of data-cap tariffs in order to be able to build a model.

First, in section 3.1, it was investigated whether and how the effects constituting a flat-rate bias are also present for data-cap tariffs. The basic result was that most forms of data-cap tariffs can be seen as a hybrid between a pay-per-use and a flat-rate tariff. Accordingly, flat-rate biases are partially applicable to most types of data-cap tariffs. This insight already leads to the most important chapter of the first part, the identification of the different determinants of data-cap tariffs and their different properties in chapter 4. Three basic determinants were identified. The first determinant was how the volume of the data-cap is determined, it could be either a fixed or a relative value. The second and probably most interesting determinant is what happens when the cap is reached. Three scenarios can be distinguished here: The first one is the simplest, when the cap is reached the connection will be terminated. In the second scenario the consequences are less severe since the connection will not be terminated but the connection speed will be reduced. And finally in the third scenario, after reaching the cap there will be some restrictions on the accessible resources, for example only domestic web services will be accessible after reaching the cap. In all three scenarios the respective restrictions can be fixed for the remainder of the billing cycle or there could be the option to buy additional data volume. The third and last determinant are policies on what services will be counted when calculating the already used volume. The basic distinction here is between QoS and non-QoS tariffs. An overview for the different determinants and properties of data-cap tariffs can be found in table 4.1. Data-cap tariffs containing these elements can be found in practice around the world and other properties for these three determinants are imaginable.

Based on this determinants a simple and realistic data-cap tariff could be made up of the following elements: connection termination, a QoS-scheme and in a first step no top-ups. Such a scenario was the basis for the second part of this thesis.
Rationing theory was found to be very fitting and similar to the situation created by a data-cap tariff. Any constraint to consumption, other than the utility function, can be seen as a point rationing ration. Therefore a data-cap tariff can be seen as a multi-currency system with the three currencies money, data and time.
From the available rationing models Tobin's "A Survey of the Theory of Rationing" [Tob52] has been found most suitable as a starting point for a model for data-cap tariffs. Based on the model presented in this article the consumer choice under a data-cap was modeled as an utility function which the consumer maximizes subject to the three constraints of money, data and time. This maximization problem can be solved with a Lagrange multiplier approach, where the Lagrange multipliers measure how much the individual currency limits the choices within the maximization. With this information it is also possible to calculate prices in a single currency that produce the same consumption points

as this multi-currency system. These prices are therefore called virtual prices and can be used to compare how prices have to be adjusted to produce the same outcome as the addition of a certain constraint. So in this case virtual prices are a means of calculating hidden price changes induced by the introduction of a data-cap.

Based on this model, a first analysis of some basic scenarios produced some insights about data-cap tariffs. For this analysis a simple framework was used. This framework assumes that all services are offered for free, so the restriction of money can be ignored. Further it is assumed that the prices in data and time can be divided into the two categories, high and low price. Lastly it is assumed that there are no services with high data and high time price and that both currencies have an equally big cap. Hence without a managed service both restrictions are equally restricting.
The first insight gained was that a data-cap tariff can at best leave the consumer welfare unchanged. Since consumer welfare is probably decreased, this must be compensated for by a lower price for the data-cap tariff as compared to a flat-rate tariff.
Besides that, adding a data-cap is just adding an additional constraint. Hence the optimization is now subject to an additional constraints. However besides the further restricted set of possible consumption points nothing changes here.
This changes when QoS is introduced, as it is an interesting question if a service should become a managed service. A first insight here is that managed services will increase the utility compared to a pure data-cap scenarios without QoS since a managed service frees up data currency that can then be used on other services. As a result, the overall consumption and with it the utility will be higher. This also implies that becoming a managed service could help the competition.
A very important insight is that becoming a managed service can actually decrease the consumption of the respective service. This is the case when a service has a low data price but a high price in the other currencies. In such a scenario becoming a managed service will increase the virtual price of the service and therefore decrease its consumption.
On the other hand becoming a managed service will in most cases benefit services with a high data price and a relatively low price in the other currencies[1]. In this case however there is always the immanent risk that making a high-data-price-service a managed service will render the data restriction ineffective for some consumers. This in turn will benefit all services, even competitors of the managed service.
All insights in this paragraph are ceteris paribus, therefore everything else is unchanged, this especially applies to the size of the data-cap.

When competing services are introduced more insights regarding managed services can be obtained. As a simplification it was assumed that direct competitors give the same utility with the same prices and therefore are interchangeable in any quantity at any time[2]. Most of the insights gained when a single service becomes a managed service are still valid when service competition is introduced. But there are also some differences and additional

[1] It does not need to be low in all cases, in most cases it should be enough if it is not relatively high, so not higher as for the other services.

[2] In a real world scenario this would probably be different as a consumer often has to decide between competing web service as a mixed usage of them does not make sense for practical reasons.

effects.
When a service with low prices in all currencies becomes a managed service, it will only profit little from it but also its direct competitor will profit because some of the freed data will be spend on the competitor.
In a competition scenario, contrary to the monopolistic scenario, it could be worthwhile for a service with a low data price and and high prices in the other currencies to become a managed service. The service will still face the same problems as in the monopolistic scenario but so will his direct competitor and without being a managed service the effects will be severer for the competition. It is a classical case of raising competitors costs.
And lastly for the competition analysis it is noteworthy that a service with a high data price will profit strongly from becoming a managed service. How strongly will depend on its prices in the other currencies and the respective caps. However, as this might free up large amounts of data currency such a move might render the data restriction ineffective. And even if that is not the case, a competing service with a high data price will also profit strongly from the large amount of data currency now available again.

Regarding welfare it is even harder to make general statements. Part of the reason behind that is that welfare for flat-rate tariffs already depends on actual usage. Hence the baseline is already a variable. However, effective constraints will restrain consumption and with that usually comes a decrease in welfare since the restriction must be beyond the market equilibrium. Another aspect regarding welfare are the prices of the individual goods. They change relatively with the introduction of additional currencies. As a general rule it can be noted that the virtual price will rise if a service has a high price in a scarce currency.

10.2. Implications

The most important implication of this thesis is probably the model itself. It allows for an analysis of many different forms of data-caps. Regarding consumer choice and the associated welfare effects this model produces consumption points which can be compared with other tariffication schemes.

Besides this, the basic analysis gives some interesting result, mostly regarding managed services.
The first important implication of this model is that with a data-cap there will always be a loss in consumer's surplus as long as the cap is effective. Hence if an ISP offers a flat-rate tariff as well as a data-cap tariff, all data-cap tariffs must be cheaper than comparable[3] flat-rate tariffs. Or else a rational consumer would always choose the flat-rate tariff.
Assuming that a web service provider wants to maximize the usage of its service there is a immanent need to become a managed service. This effect is due to the fact that an effective data-cap must always restrain consumption.
But not every service should become a managed service, sometimes becoming a managed service can even help competitors. Therefore the decision to become a managed service has to be made for every individual case, taking all variables into account. A main reason

[3] Regarding connection speed and other technical properties besides consumed data volume.

behind that are the compensation effects. Having multiple currencies with a budget for each creates interdependencies between the different services when optimizing the utility because the utility needs to be maximized while staying within several budgets. The general rule is that a service should consider becoming a managed service if it has a high data price and data is a relatively scarce resource. Additionally, the prices in the other currencies, especially if they are scarce, should be relatively low. It also needs to be taken into account that becoming a managed service will reduce the scarcity of the data currency. This will on the one hand also help all other available services, even direct competition, and on the other hand makes the decision to become a managed service less preferable in retrospective.

10.3. Limitations

Since this is the first iteration of a basic model, there are obviously a lot of limitations which should be investigated in future research.
By design this research only looks at the consumer's side of the data-cap problem. For a complete picture, the consequences on the ISPs and the service providers must also be taken into account and therefore thoroughly investigated. Associated with that the problem was simplified as much as possible whenever the producers' side was involved. To obtain results for more complex scenarios, as they are sometimes found in real world scenarios, additional complexity sometimes needs to be added in this area. For example the incentive for a web service provider to become a managed service was simplified to the statement that it will become a managed service whenenever it will increase its revenue by doing so. That might be an oversimplification in many cases.

Further analysis must also include more complex utility functions. Since the utility function is the object of the maximization, its form can have a huge impact on the results of the analysis. For example, it was assumed that competing services are perfect substitutes for each other and can be consumed in any combination to fulfill the need for such a service. Due to customer lock-in and similar effects it seems to be more realistic that the services are perfect substitutes but a consumer will usually only consume one of them. And even perfect substitutes are rare in real world scenarios.
Also the model so far assumes that after reaching the cap the connection will just be terminated. Nonetheless for all kinds of data-caps, as discussed in the first part, there is the option of top-ups. To investigate this option will be an important task for future research, as some important psychological traits are associated with this. The same is true for a connection speed reduction after reaching the cap. It seems as this kind of data-cap is a midway between a pay-per-use and a flat-rate tariff. Therefore for economical as for psychological reasons it is hard to model. Nevertheless it seems to have great potential which implies a need to investigate it further.

The most important limitation of this analysis probably is not the focus on the consumer but the focus on the individual rather than on the aggregated level. Since this is a first approach for a model, it makes sense to start with the individual consumer as a basis. Nevertheless, telecommunications with its high sunk cost is a mass business and therefore the aggregated level is the relevant one. Hence for tariff design it is rather irrelevant whether a data-cap is effective for an individual user. The same is true for the decision to become a managed service. What is relevant in both cases is the aggregated level.

10.4. Outlook

The topic of data-cap tariffication is closely connected with managed services and that in turn is closely connected with net neutrality. Recent developments in this area have somehow be contradicting as the EU and Brazil have adopted relatively strict net neutrality rules[4], while the US has net neutrality laws but allows managed services of almost all kinds[5]. So it is worthwhile to further investigate data-cap tariffs. Since the work here were first steps towards a model, most of the next steps are already defined by the limitations of this basic model.

It will be necessary to investigate effects on ISPs and web service providers to get the whole picture. To achieve that it also is critical to develop methods to compute the aggregated behavior of consumers as well as ISPs and service providers.

It must also be a matter of future discussion how top-ups and data-caps with connection speed reduction influence data-cap tariffs as they are a important characteristics that position data-cap tariffs in-between pay-per-use and flat-rate tariffs. As a result the investigation of top-ups will have general economical as well as behavioral economical aspects.

[4] `http://thehill.com/blogs/congress-blog/technology/206137-net-neutrality-eu-brazil-and-us`, last accessed 30th May 2014.

[5] `http://www.politico.com/story/2014/05/fcc-approve-net-neutrality-proposal-open-Internet-106720.html`, last accessed 30th May 2014.

Declaration

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 14.09.2014 ________________________

Ferdinand Gantert

Bibliography

[BNW03] Roland Bless, Kathleen Nichols, and Klaus Wehrle. A lower effort per-domain behavior (pdb) for differentiated services. RFC 3662, The Internet Society - Network Working Group, 2003.

[BSE+13] Jeremy Blackburn, Rade Stanojevic, Vijay Erramilli, Adriana Iamnitchi, and Konstantina Papagiannaki. Last call for the buffet: economics of cellular networks. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 111–122. ACM, 2013.

[CADS10] Jay Chen, Saleema Amershi, Aditya Dhananjay, and Lakshmi Subramanian. Comparing web interaction models in developing regions. In *Proceedings of the First ACM Symposium on Computing for Development*, pages 6:1–6:9. ACM, 2010.

[Car48] C.F. Carter. The dual-currency problem. *The Economic Journal*, 58(232):586–594, 1948.

[CBB+12] Marshini Chetty, Richard Banks, A.J. Bernheim Brush, Jonathan Donner, and Rebecca Grinter. You're capped: understanding the effects of bandwidth caps on broadband use in the home. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 3021–3030. ACM, 2012.

[CBG11] Hsing Kenneth Cheng, Subhajyoti Bandyopadhyay, and Hong Guo. The debate on net neutrality: A policy perspective. *Information Systems Research*, 22(1):60–82, 2011.

[Cis12a] Cisco visual networking index: Forecast and methodology, 2012–2017. Technical report, Cisco Systems, 2012. URL: `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf` [cited 1st June 2014].

[Cis12b] The zettabyte era - trends and analysis. Technical report, Cisco Systems, 2012. URL: `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf` [cited 1st June 2014].

[CJ09] Richard N. Cooper and Andrew John. *Economics: Theory Through Applications*. McGraw Hill Higher Education, 2009.

[Cro07] Jon Crowcroft. Net neutrality: the technical side of the debate: a white paper. *ACM SIGCOMM Computer Communication Review*, 37(1):49–56, 2007.

[DME06] Peter J Danaher, Guy W Mullarkey, and Skander Essegaier. Factors affecting web site visit duration: a cross-domain analysis. *Journal of Marketing Research*, 43(2):182–194, 2006.

[dS42] T. de Scitovszky. The political economy of consumers' rationing. *The Review of Economic Statistics*, 24(3):114–124, 1942.

[GR12] John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. Technical report, 2012. URL: `http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf` [cited 1st June 2014].

[Gra47] J. De V. Graaff. Rothbarth's "virtual price system" and the slutsky equation. *The Review of Economic Studies*, 15(2):91–95, 1947.

[Har43] W.C. Haraldson. Welfare economics and rationing. *The Quarterly Journal of Economics*, 58(1):146–148, 1943.

[Hen52] John Steele Henderson. *Production and Consumption*. University of Alabama Press, 1952.

[How77] David H Howard. Rationing, quantity constraints, and consumption theory. *Econometrica: Journal of the Econometric Society*, 45(2):399–412, 1977.

[JW11] David R Just and Brian Wansink. The flat-rate pricing paradox: conflicting effects of "all-you-can-eat" buffet pricing. *The Review of Economics and Statistics*, 93(1):193–200, 2011.

[Kal41] Michał Kalecki. General rationing. *Oxford Institute of Statistics Bulletin*, 3(1):1–6, 1941.

[Kru08] Jörn Kruse. Internetüberlast, netzneutralität und service-qualität. *Wirtschaftsdienst*, 88(3):188–194, 2008. Text in German.

[KW10] Jan Krämer and Lukas Wiewiorra. Consumer perception of mobile telephony tariffs with cost caps. In *Telecommunications Internet and Media Techno Economics (CTTE), 2010 9th Conference on*, pages 1–8. IEEE, 2010.

[Llo42] E. M. H. Lloyd. Some notes on point rationing. *The Review of Economic Statistics*, 24(2):49–52, 1942.

[LS06] Anja Lambrecht and Bernd Skiera. Paying too much and being happy about it: Existence, causes and consequences of tariff-choice biases. *Journal of Marketing Research*, 43:212–223, 2006.

[Mak45] Helen Makower. Rationing and value theory. *The Review of Economic Studies*, 13(2):75–80, 1945.

[McM54] M. McManus. The geometry of point rationing. *The Review of Economic Studies*, 22(1):1–14, 1954.

[MM56] M. Mc Manus. Points rationing and the consumer. *Metroeconomica*, 8(2):118–134, 1956.

[MMV95] Jeffrey K MacKie-Mason and Hal R Varian. Pricing congestible network resources. *Selected Areas in Communications, IEEE Journal on*, 13(7):1141–1149, 1995.

[MSW09] John Musacchio, Galina Schwartz, and Jean Walrand. A two-sided market analysis of provider investment incentives with an application to the net-neutrality issue. *Review of Network Economics*, 8(1), 2009.

[Nei43] Hans P Neisser. Theoretical aspects of rationing. *The Quarterly Journal of Economics*, 57(3):378–397, 1943.

[NR80] J Peter Neary and Kevin WS Roberts. The theory of household behaviour under rationing. *European economic review*, 13(1):25–42, 1980.

[NW74] Yew-Kwang Ng and Mendel Weisser. Optimal pricing with a budget constraint—the case of the two-part tariff. *The Review of Economic Studies*, 41(3):337–345, 1974.

[Oi71] Walter Y Oi. A disneyland dilemma: Two-part tariffs for a mickey mouse monopoly. *The Quarterly Journal of Economics*, 85(1):77–96, 1971.

[Puj10] Nicolas Pujol. Freemium: attributes of an emerging business model. 2010. URL: `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1718663` [cited 1st June 2014].

[Red42] Melvin W Reder. Welfare economics and rationing. *The Quarterly Journal of Economics*, 57(1):153–159, 1942.

[Rot41] Erwin Rothbarth. The measurement of changes in real income under conditions of rationing. *The Review of Economic Studies*, 8(2):100–107, 1941.

[Rot45] K.W. Rothschild. Rationing and the consumer. *Oxford Economic Papers*, 7(1):67–82, 1945.

[Sam48] P. A. Samuelson. *Foundations of Economic Analysis*. Harvard University Press, 2. pr. edition, 1948.

[Sid06] J Gregory Sidak. A consumer-welfare approach to network neutrality regulation of the internet. *Journal of Competition Law and Economics*, 2(3):349–474, 2006.

[SV98] Carl Shapiro and Hal Varian. *Information rules*. Harvard Business Press, 1998.

[TBAA89] Kenneth E Train, Moshe Ben-Akiva, and Terry Atherton. Consumption patterns and self-selecting tariffs. *The Review of Economics and Statistics*, 71(1):62–73, 1989.

[Tee10] David J Teece. Business models, business strategy and innovation. *Long range planning*, 43(2):172–194, 2010.

[TH50] James Tobie and Hendrik S Houthakker. The effects of rationing on demand elasticities. *The Review of Economic Studies*, 18(3):140–153, 1950.

[Tha80] Richard Thaler. Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1):39–60, 1980.

[Tob52] James Tobin. A survey of the theory of rationing. *Econometrica: Journal of the Econometric Society*, 20(4):521–553, 1952.

[Var10] Hal R. Varian. *Intermediate Micro Economics - A modern approach.* W. W. Norton & Company, 8. edition, 2010.

[WBH13] Thomas M Wagner, Alexander Benlian, and Thomas Hess. The advertising effect of free–do free basic versions promote premium versions within the freemium business model of music services? In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 2928–2937. IEEE, 2013.

[Wor44] George N.D. Worswick. Points, prices and consumers' choice. *Bulletin of the Oxford University Institute of Economics & Statistics*, 6(3):33–39, 1944.

[WSC+10] Susan P. Wyche, Thomas N. Smyth, Marshini Chetty, Paul M. Aoki, and Rebecca E. Grinter. Deliberate interactions: Characterizing technology use in nairobi, kenya. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2593–2602. ACM, 2010.