Network Architecture and Content Provision: An Economic Analysis*

J. MacKie-Mason
Univ. of Michigan
jmm@umich.edu

S. Shenker
Xerox PARC
shenker@parc.xerox.com

H. R. Varian
Univ. of Calif., Berkeley
hal@sims.berkeley.edu

June 29, 1996

Abstract

There are at least two competing visions for the future National Information Infrastructure. One model is based on the application-blind architecture of the Internet; the other is based on the application-aware architecture of cable TV systems and online services. Among application-aware architectures, some are content-aware and some are content-blind. In this paper we examine the consequences of these different network architectures for content provision.

*This work was supported by the National Science Foundation grant SES-93-20481.

1 Introduction

The formation of the National Information Infrastructure is often depicted as a battle between various economic entities such as the cable TV companies and the regional Bell operating companies. More recently, the Internet – with its enigmatic origins, anarchic organization, and startling growth – has gained widespread visibility and is now seen as a fundamental component of the NII. In addition, the various online services (AOL, Prodigy, CompuServe, MSN, etc.), direct broadcast TV, wireless telephony, and other emerging technologies will also likely be important parts of the NII.

These various components display a wide range of characteristics. The physical transmission medium runs the gamut from satellite transmission to coax cables to fiber optics. Some components are broadcast in nature (everyone in the system receives every signal) and others are switched (signals are directed to specific receivers). The corporate entities deploying these component technologies are diverse — from startups to established concerns, and from local monopolies to international competitors — and they face disparate regulatory restraints. These physical, corporate, and regulatory issues have been thoroughly discussed in the literature, and we do not address them here. We focus instead on another issue that, to date, has received considerably less attention.

We begin with the observation that these various components of the NII have radically different *architectures*. By architecture we are not referring to the actual physical implementation of the network. Rather, we are referring to the nature of the service provided. In this paper, we concentrate on one crucial feature of service architecture: *transparency*. We call the entity that transports the bits to consumers the *network provider*. In some architectures, the network provider is aware of the content of transported bits (e.g., movie, point-to-point video, music, etc.), and in others the content is completely opaque to the network provider.[1] See [5] for a brief and speculative discussion of the implications of this distinction.

We then ask the question: how does this difference in architecture affect the nature of content provided to consumers? Does this difference in the network provider's awareness affect what services and content are offered in these systems?

This question deals with issues too complex to yield a simple and definitive answer. Our purpose here is to provide some initial intuition about the effect of architecture on content provision. To that end, we analyze this question in

[1] In Section 2 we clarify that there are gradations of awareness, from awareness of the applications (e.g., teleconference vs. file transfer) to awareness of the actual content (e.g., the specific movie being transmitted).

the context of some very simple models; we use these to identify some of the major issues, and illustrate them through examples. In particular, throughout our discussion we restrict ourselves to the case of a single network provider that serves customers who choose to connect; our present analysis does not address the more complicated, and realistic, case of multiple coexisting and competing architectures.

We focus on two economic characteristics of service architecture:

- the ability to differentiate transport prices for different goods (Section 3); and,
- various cost elements that affect the economics of service architecture (Section 4).

In the category of cost elements we consider costs of delivering content (technological and institutional), and costs of consuming (finding and processing) content. Both of these sections consider already created content. We also consider the impact of architecture on the *creation* of content (Section 5).

2 Architectures

The architectural distinctions we make here concern the extent to which the network provider distinguishes between the bits it conveys. The Internet, telephony, and cable TV occupy very different places along this spectrum; they are representatives of the three basic architectural choices that we describe below.

2.1 Architectural Choices

2.1.1 Application-blind networks

One of the Internet's central design principles is that the network provides only bit transportation; it is up to the end hosts to construct higher-level applications on top of this raw transport mechanism. This architecture has the great advantage that it need not be modified as new applications arise, because applications are implemented entirely at the end hosts and no centralized authority need approve such applications. Consequently, the Internet has seen a proliferation of applications, such as electronic mail and the WWW, that were not envisioned when the IP protocol was originally designed.

We will use the term *application-blind* to refer to architectures where a general interface is made available to end users, who then implement their applications on top of this interface. For purposes of our analysis we assume that application-blind networks operate as common carriers. They offer a single, nondiscriminatory price for transport and accept any and all traffic at that price. The operational relevance of this architecture is that prices cannot be based on what is *inside* the packets, but only on raw flow measures and various connection fees.

2.1.2 Application-aware networks

In contrast, telecommunication infrastructures developed by private enterprise such as cable TV and telephony are more tightly coupled to specific applications. The underlying transport function deep within these network may be general but the interface presented to users is highly restricted. In this type of network the architecture is aware of the type of application that is being used.

For example, telephone networks are designed around voice traffic and cable TV protocols are designed for video traffic. Similarly, online service providers such as Compuserve and AOL generally know what kinds of applications (voice, data, images, etc.) are being transmitted.

By an *application-aware* architecture we mean that the network service provider can identify the general type of application being invoked (e.g., e-mail, audio playback, interactive video). For our purposes, the critical point about this distinction is that the price for transport can be set knowing the application being used on the network. This means that the network provider can charge different prices for different applications, and consequently also exclude certain applications.

2.1.3 Content-aware networks

Besides being able to determine the general applications customers use, some networks can also monitor and even control the content that is transported over some applications. For example, cable TV can distinguish basic from premium channels and video-on-demand systems know what movie has been requested. Online services also know what sort of content is being requested: airline prices, bibliographic references, cartoons, etc. We define a *content-aware* architecture as one in which the network provider can identify the network content.

It is not the case that all *application*-aware networks are also *content*-aware. For example, telephony is content-blind, as is electronic mail or WWW browsing on an online network. When a network is content aware, pricing can be differentiated not just according to the general application, but by the content itself.

2.2 Architectural Implications for Content Provision

Our fundamental point is that network architecture can have important implications for the nature of goods available, and for the pricing of those goods. Anecdotal evidence is certainly consistent with this view. For instance, the application-blind Internet supports a diverse and rapidly growing set of applications such as electronic mail, file transfer, teleconferencing and WWW. The application-aware telephone system supports only a narrow range of services (basically FAX, low-speed modems and telephony), but provides access to varied content in the form of 900 numbers. Cable television, which is content-aware, offers only one application and rather limited content. Are these differences related to architectural differences?

This question has two different components. First, given a set of content goods already created, what effect does architecture have on the set of goods offered to consumers? Since the network provider is the entity actually delivering the goods to consumers, the pricing and access policies of the network provider will affect which goods actually are purchased by the consumers; we investigate the role of architecture in shaping those pricing and access decisions. Second, how does the choice of architecture affect the creation of content by determining which goods get produced? It is clear that the behavior of the network provider will affect the incentives seen by content providers.

We consider the two components of our basic question separately. To focus on the role of architecture, we simplify or eliminate most other relevant factors. For instance, we assume the network provider is a monopoly. This provider is free to maximize profits without competitive (or regulatory) pressures. Moreover, we assume that the network provider is a monopsonist and can set transport prices without negotiating with the upstream content providers. Lastly, we assume that content provision is competitive, with a large number of content providers, and incurs no marginal cost (we can easily incorporate a finite marginal cost into the formalism at the expense of notational simplicity).[2]

Our modeling of goods and consumers is also quite simple. In most of what follows, we label separate goods by an index i, although in some examples it is

[2] By making these assumptions, we have avoided the question of *bundling* or vertical integration between the network provider and content providers.

more convenient to consider a continuum of goods labeled by $x \in [0, 1]$. These goods all have equivalent bandwidth requirements. We model the consumers as consuming at most one unit of each good, with a reservation value v_i^α for consumer α. Initially, we assume this willingness-to-pay is independent of the consumption of other goods (later we will modify this assumption slightly in Sections 4.2 and 4.3), and so a consumer's satisfaction is merely the sum of the consumed v_i^α minus the price paid for the goods.

In a content-blind but application-aware architecture, different goods denote different applications. In a content-aware architecture, different goods can refer to different content as well as different applications. Given this different definition of good in the content-aware and application-aware architectures, our analysis need only distinguish between aware and blind architectures. The blind architecture cannot distinguish between goods, and the aware architecture can. This key architectural distinction has several different implications for content selection from already created goods. We discuss the foremost of these, price differentiation, in the Section 3; various cost implications for content provision are in Section 4. Content creation is considered in 5.

3 Price Differentiation

The network provider charges consumers a price p_i to deliver good i. In a blind architecture, since the bits being transported are opaque to the network provider, charges can only be based on the bandwidth characteristics which, for convenience, we have assumed to be equivalent. Thus, in a blind architecture, all prices must be the same: $p_i = p$ for all i. In contrast, in an aware architecture the network provider can differentiate between different goods, and charge accordingly. The ability to price differentiate is the most important architectural distinction between the aware and blind architectures.

The issue of price differentiation is rather similar to the much studied issue of price discrimination (see [7]). Price discrimination refers to the ability to charge different consumers differently, whereas price differentiation refers to the ability to charge differently for different goods. In most contexts it is assumed that different goods can be priced differently, so the blind architecture produces a pricing structure much less flexible than typically considered in the literature; not only are all consumers charged the same prices, but all goods are also identically priced. The aware architecture is merely embodies the ability, automatically assumed in most pricing contexts, to distinguish among goods in the pricing scheme. We do not consider the natural last stage of this progression; an aware architecture able to price discriminate so that different consumers face different prices.

Before embarking on our analysis, we first establish some notation. Let $R_i(p_i)$ denote the revenue generated by good i at price p_i, and let $p_i^{max} = \arg\max R_i(p_i)$ and $R_i^{max} = R_i(p_i^{max})$. In addition, let R denote the total revenue, and S the total consumer surplus. We use $R(p)$ to denote the total revenue when all $p_i = p$: $R(p) = \sum_i R_i(p)$. Let $p^{max} = \arg\max R(p)$ and $R^{max} = R(p^{max})$.

We have assumed that the network provider is a monopolist, and thus can set prices to maximize revenues. The aware architecture, in the absence of any other effects (which we consider later in Section 4), sets each p_i to maximize the revenue raised from that good, which is p_i^{max}. While the aware architecture extracts the maximal revenue from each good, the blind architecture sets the single price to maximize overall revenue $p = p^{max}$. Goods for which $R_i(p^{max}) = 0$ are not offered on a blind network, since there is no demand for them at the price charged by the network provider. Thus, the most obvious impact on the provision of goods is that in the aware architecture all goods are offered, whereas in the blind architecture some goods are not offered; the goods not offered are those whose maximal-willingness-to-pay is relatively low. Below, we consider two other aspects of price differentiation.

3.1 Welfare and Revenue Effects

Price differentiation affects the cumulative demand, revenue, consumer surplus, and total welfare (consumer surplus plus revenue). There is a substantial literature on the output and welfare effects of price discrimination; see [7] for references. As in the price discrimination literature, we find that there are few unambiguous results comparing the welfare implications of the blind and aware architectures; while revenue is always larger in the aware architecture, the total welfare and consumer surplus can either be bigger or smaller in the aware architecture.[3]

To clarify the source of this ambiguity, it is helpful to observe that there are two sources of heterogeneity in our model: (1) heterogeneity in content valuations by a single consumer (i.e., where $v_i^\alpha \neq v_j^\alpha$), and (2) heterogeneity in the valuations of a particular good by different consumers (i.e., where $v_i^\alpha \neq v_i^\beta$). When we consider models where the first form of heterogeneity dominates, the picture is

[3]Roughly speaking, the only other general result is that allowing price differentiation will increase welfare only if output is increased.

much clearer.

To isolate the effect of heterogeneity in goods, we could start with the case where all consumers have identical valuations of each good: $v_i^\alpha = v_i^\beta$ for all i and all α, β. However, it turns out that we can incorporate slightly more consumer heterogeneity without changing the results. We assume, in what we will call the canonical example, that for each consumer α and each good i, either $v_i^\alpha = v_i$ or $v_i^\alpha = 0$; either a consumer desires good i, and is willing to pay an amount v_i for it, or doesn't care for it at all. The fraction of users who are willing to pay for good i is denoted by f_i; for convenience we normalize the population size to one.

The aware architecture sets $p_i = v_i$ for each i and can extract the entire consumer surplus and raise revenue $R = \sum_i f_i v_i$. In contrast, the blind architecture sets p to maximize $R = p \sum_i f_i \delta(v_i \geq p)$.[4] The remaining consumer surplus $S = \sum_i f_i (v_i - p)_+$ (where we use the notation $z_+ \equiv \max[0, z]$) can be positive. Goods with $v_i < p$ are not offered, since no consumers would purchase them.

The consumer surplus is always zero in the aware architecture but can be nonzero in the blind architecture. The revenue and the total welfare in the aware architecture are never less than they are the blind architecture, as is the usual case with first-degree price discrimination.

Thus, all other considerations aside, network providers would always prefer the aware architecture since price differentiation allows them to increase their revenue. In contrast, given the monopolistic nature of the network provider, consumers would prefer the blind architecture, since it can lead to nonzero consumer surplus.

However, once we add even the slightest heterogeneity into the positive consumer valuations, these certainties no longer hold (although the maximizing of revenue in the aware architecture does). To illustrate this, we examine the effects of price differentiation in three simple examples, each with two goods and two consumers. If $v_1^\alpha = 10$, $v_1^\beta = 8$, $v_2^\alpha = 3$, $v_2^\beta = 2$, then the consumer surplus and the total welfare are maximized in the aware architecture. If $v_1^\alpha = 9$, $v_1^\beta = 4$, $v_2^\alpha = 4$, $v_2^\beta = 1$, then the consumer surplus and the total welfare are maximized in the blind architecture. If $v_1^\alpha = 10$, $v_1^\beta = 8$, $v_2^\alpha = 6$, $v_2^\beta = 4$, then the consumer surplus is maximized in the blind architecture, but the total welfare is maximized in the aware architecture. Thus, even in such simple two-good two-consumer models we can exhibit all possible combinations of outcomes consistent with revenue being maximized in the aware architecture.

[4] The δ function equals one when its argument is true, and zero otherwise.

3.2 Revealed Profit Maximization

While we cannot in general predict the welfare and surplus effects of price differentiation, we can, in some cases, link the prices in the aware architecture to that in the blind architecture. It turns out that "revealed profit maximization" puts some restrictions on what patterns of prices can be observed. For instance, consider the case where the valuations of the various goods are nonoverlapping. This models the situation where the intrinsic values of the various goods are well separated, such as video telephony, personal electronic mail, and electronic junk mail. To make this precise, define $I_i = [\min_\alpha v_i^\alpha, \max_\alpha v_i^\alpha]$ and assume that $I_i \cap I_j = \emptyset$ for $i \neq j$. Moreover, assume that, without loss of generality, $v_1^\alpha \geq v_i^\alpha$. Obviously, $p_i^{max} \in I_i$. However, we can also conclude that (1) $p^{max} \in I_i$ for some i, (2) if $p^{max} \in I_i$ then $p^{max} \geq p_i^{max}$, and (3) if $p^{max} \in I_1$ then $p^{max} = p_1^{max}$. This follows trivially from inspection of the demand curves.

4 Cost Implications for Content Provision

We now consider cost-related effects of the architectural difference between blind and aware networks. There are two important types of cost. The first type of costs are *network costs*; these arise at the network level, and are not attributable directly to individual decisions. We consider the impact of costs directly arising from delivering goods in Section 4.1; we consider costs borne by users but due to network provider decisions in Section 4.2. The second type of costs are *user costs*, which can be traced to decisions made by individuals. We consider such costs in Section 4.3.

The effect of different types of cost on content provision depends on the network architecture. An aware network — with its greater control — is advantaged when there are significant *network costs*. That is, the selection of which goods to provide is more efficient. There is no such advantage with *user costs* since these can be allocated efficiently by the actions of individual users without network intervention.

The different locus of control also implies a difference in the selection order for content. An aware network controls content selection, and thus orders choices by the profits each good generates. (We make our notion of "ordering" precise below.) In a blind network, with its single transport price p, customers control content selection. Any good with positive demand at price p will be purchased by some consumer, so goods are generally ordered by maximal willingness-to-pay.

This contrast between ordering by profitability versus maximal willingness-to-pay has interesting implications for the content diversity on various networks. Consider, for example, two very different kinds of goods: low-value *mass-market* goods, which have low maximal willingness-to-pay but high total revenue, and high-value *niche* goods, which generate relatively little revenue but have high maximal willingness-to-pay. Aware architectures will tend to favor low-value mass-market goods, while blind architectures will favor high-value niche goods. We discuss this more fully in Section 4.1. This is consistent with what we observe when comparing the offerings of the Internet with cable television.[5]

We now make these observations more concrete by considering several cost factors that affect the provision of goods.

4.1 Liability and Gateway Effects

On a blind network, the net provider incurs no cost when a new application or good is offered. All applications use the common, existing network layer protocol (*e.g.*, IP in the Internet) so additional programming is not required. Assuming that blind network providers are treated as common carriers, they bear no liability for any content made available. In contrast, aware networks typically incur service, liability and reputation costs when they offer new applications or goods; we describe these costs more fully below.

Content-aware networks are unlikely to be treated as common carriers by the courts, and thus face some liability for their content. The recent libel suit (see [3]) against Prodigy is a reminder of this. Similarly, the Church of Scientology has filed suit against a network provider over copyrighted documents presented to the Internet by one of the provider's users (see [1]). Further, an aware network may have a reputation for content screening that would be costly to let deteriorate. Internet service providers are not held responsible for the poor quality of available information, but presumably AOL and other online service users do judge their provider on this basis. Thus, content-aware networks may face a potential cost, either in liability or reputation, for each new piece of content they offer.

There is a similar cost to adding new applications to an application-aware network, but for very different reasons. Application-aware networks are typically built using application-level gateways or interfaces. Adding a new application

[5]This is not to imply a causal link; there are many other differences between these two systems that could explain this observation. For instance, the broadcast nature of cable television would favor mass-market goods while the switched nature of the Internet does not, although the advent of multicast is changing that somewhat.

requires modification to the network to change these gateways/interfaces, and thus the network faces an incremental cost for each new application offered. For instance, Web access in the various online services took a great deal of effort to establish (and was substantially delayed). Application-aware networks may also have a reputation to uphold, but here it hinges on the quality of the application rather than the quality of the content itself (e.g., in a video-on-demand system, it would be reasonable to hold the network accountable for the fidelity of the signal, but not for the quality of the content itself.).

The liability and gateway effects in the aware architecture can be modeled by introducing a cost C_i for offering a new good i. These are *network costs*, which are not attributable to specific users. Incurring a cost C_i when offering a new good (a new application in an application-aware architecture, or new content in a content-aware architecture) influences which goods are brought to market. In a blind architecture, all goods with $R_i(p^{max}) > 0$ are offered. In an aware architecture, only those goods with $R_i^{max} > C_i$ are offered. Thus, as we informally argued before on more general grounds, the two architectures have radically criteria for deciding whether or not a good is offered; the aware architecture criterion depends on the maximal total profit whereas the blind architecture criterion depends on whether or not any consumers are willing to buy the good at price p^{max}.

To see the qualitative effect this can have on the nature of goods offered, we return to the canonical example and consider the simplifying case where $C_i = C$ for all i and goods fall into a few categories: goods are either high value or low value ($v_i = v^h$ or $v_i = v^l$) and either mass-market or niche ($f_i = f^m$ or $f_i = f^n$). We make the following assumptions: $p^{max} = v^h$ (that is, the revenue is maximized in the blind case when the high value price is charged) and $f^m v^h > f^m v^l > C > f^n v^h > f^n v^l$. Then, both architectures offer mass-market high-value goods, and neither offers niche low-value goods. However, the blind architecture offers high-value niche goods and not low-value mass-market goods, while the aware architecture does exactly the reverse.

Without the costs C_i, in the canonical example the network provider always preferred the aware architecture, the consumer surplus was greater in the blind architecture, and total welfare was maximized in the aware architecture. This simple story no longer holds when we introduce the gateway/liability costs. If we let n^{mh}, n^{ml}, n^{nh}, n^{nl} denote the number of goods in each of the the various categories, we see that the revenue (and the total welfare)[6] of the blind and aware

[6]Because we have assumed that $p^{max} = v^h$ there is no consumer surplus in either architecture and so the total welfare is the same as the revenue.

architecture are, respectively, $(n^{mh} f^m + n^{nh} f^n) v^h$ and $(n^{mh} v^h + n^{ml} v^l) f^m - (n^{mh} + n^{ml}) C$. The difference is given by: $n^{nh} f^n v^h - n^{ml} f^m v^l + (n^{mh} + n^{ml}) C$. Evidently the blind architecture may generate more profit and higher total welfare. Indeed, the blind architecture might serve a *larger* market in value terms than the aware architecture, even though it favors niche over mass market applications. This can occur if either of the following are true:

- high-value niche applications are valued much more than low-value mass market applications; or,
- there is such a diversity and number of niche applications relative to mass market applications that almost the same or even more people end up served by at least *some* applications.

We can further explore this issue by looking at two special cases. First, in the case where all v_i are the same, but the f_i vary, then every good would be served under the blind architecture, without incurring any gateway costs. The aware architecture will serve the mass market goods, but still may not serve the niche goods if $C > f^n v$, so the blind architecture saves on gateway costs and may offer more goods: it is clearly preferred by the network provider (since the provider can collect all of the surplus in either case). Second, suppose all $f_i = 1$ but the v_i are uniformly distributed between 0 and 1. In the blind architecture, $p = 1/2$, $R = 1/4$, and the consumer surplus S is $1/8$. In the aware architecture, $R = (1 - C)^2/2$, and there is no consumer surplus. The total welfare is greater in the aware architecture whenever $C < 1 - \sqrt{3}/2 \approx .134$, and the revenue is greater in the aware architecture whenever $C < 1 - 1/\sqrt{2} \approx .293$. For $C < 1 - \sqrt{3}/2$, the total welfare and the revenue are maximized in the aware architecture. For $C > 1 - 1/\sqrt{2}$ the total welfare and the revenue are maximized in the blind architecture. In the intermediate case $1 - \sqrt{3}/2 < C < 1 - 1/\sqrt{2}$, total welfare is maximized in the blind architecture yet revenue is maximized in the aware architecture. Note, however, that in all cases the consumer surplus is higher in the blind architecture. Therefore, the economics of service architecture suggest both possible inefficiencies and political economy conflicts.

Appendix A describes, through an example, how the gateway/liability effect not only inhibits the introduction of niche goods in aware architectures, but also inhibits the introduction of marginally improved goods.

4.2 Clutter Effects

When there are many goods available, the *clutter* that results can decrease the value to consumers (where, as before, goods refer to content in a content-aware architecture and applications in an application-aware architecture). In the content case, it becomes harder (and slower) to locate the desired material; in the application case, the interface becomes cluttered and harder to use.

Content-aware architectures can control the content available to users and play an editorial role; in fact, this is already standard practice in the moderated discussion groups of many online services. Content-blind architectures can only restrict the provision through a single price cutoff. Similarly, application-aware architectures can control the applications available to users (again, already done on the online services). The concern with clutter in blind architectures is evidenced by the huge popularity of Web indexing services (`http://www.yahoo.com/` receives about 1.5 million visits per week) and the demand for them from commercial providers (for example, AOL recently purchased WebCrawler). Readers of Usenet newsgroups and Internet mailing lists are also familiar with clutter costs.

As with gateway costs, clutter costs are *network costs* – they depend on the total number of applications or goods available, not the choices made by particular individuals. However, these costs arise not as expenses directly incurred by the network in providing a service, but in reduced user utility as more goods are offered. We model the clutter effects by decreasing an individual's utility as more goods are offered (see [6] for a discussion of similar congestion effects for electronic mail). For instance, if the network offers a set of goods G and a consumer α has purchased a set G_α of goods, then the total valuation by that consumer is $\sum_{i\in G_\alpha} v_i^\alpha - F_c(|G|)$ where F_c is some nondecreasing function. If $\sum_{i\in G_\alpha} v_i^\alpha - F_c(|G|) < 0$ we assume the consumer leaves the network. Let $\mathcal{C}$ denote the set of connected consumers, those for whom $0 \leq (\sum_{i\in G}(v_i^\alpha - p_i)_+ - F_c(|G|))$.

In the presence of clutter in an aware architecture, the network provider can decide which goods to make available. The problem is choosing G and p_i to maximize the total revenue, which is given by the sum $\sum_{\alpha\in\mathcal{C}} \sum_i p_i \delta(v_i^\alpha \geq p_i)$. This maximization problem depends on the details of the distribution of consumer preferences; in general, the goods are ordered by revenue, but the revenue must be evaluated with respect to those consumers still connected to the network, given the other goods offered.

We can formalize the notion of profit ordering by using a necessary condition for offering good j, conditional on the set of other goods that are offered at an optimum. The incremental good j is sold to some subsect of connected users, yielding revenue $R_j(p_j)$. However, some users detach from the network as a result of the extra clutter. Let $\mathcal{D}_j$ denote the set of users who detach, that is, those for whom $F_c(|G_{-j}|) - F_c(|G|) - v_j^\alpha \delta(v_j^\alpha > p_j) > S_{-j}^\alpha$, where G_{-j} is the set of offered goods excluding good j, and S_{-j}^α is the surplus obtained by user α from goods G_{-j}. Then, we can say that goods are ordered by revenue in the sense that all offered goods satisfy $R_j(p_j) > \sum_{\alpha \in \mathcal{D}_j} \sum_{i \in G_\alpha} p_i$, and all other goods do not.

It is interesting to note that the clutter effect may tend to favor mass market over niche goods. According to the detachment rule above, the users who consider detaching when a new good adds to the clutter are those who have rather low surplus: less than the incremental increase in clutter cost from one new good. Call these the *marginally attached* users; per such user the revenue required from a potential new good will be low. However, if there are many marginal users, the total amount of revenue required from a new good will be relatively high. Thus, new goods will tend to fare better if they are mass market, or, for niche goods, when they have high value per person.[7]

A network with a blind architecture cannot directly control which goods are offered; the network can merely set the overall price p. The content provider will offer the good i if and only if $R_i(p) > 0$; however, this revenue must be computed over those consumers still attached to the network. There are two relevant aspects that differentiate this from the aware case. First, the goods are ordered differently, generally by maximal willingness-to-pay rather than maximal profit; that is, any good for which $\max_{\alpha \in \mathcal{C}}[v_j^\alpha] > 0$ will be offered. Second, content providers will decide to enter the market as long as there is any demand; the resulting clutter will tend to reduce the consumer surplus. This is the typical tragedy-of-the-commons phenomena present in many congestion models; see [4] for an analysis of congestible resources.[8]

We can illustrate these two aspects of clutter with the following example. There is a continuum of goods labeled by x, with $x \in [0, 1]$. Each individual consumer α values each of these goods an amount v^α, and these values v^α are uniformly

[7] Although surely not due to clutter alone, the reports that about one-quarter of users of online services like AOL detach within a year suggests that this bias in selecting new content may be important.

[8] Another aspect, which we have not explored here (in fact, it is precluded by our particular modeling of the clutter effect), is that if different goods experience different amounts of clutter (i.e., are devalued differently), then goods that are less susceptible to clutter can displace those that are more susceptible to clutter, even if their intrinsic value is less. This is akin to the effect that in a computer network congestion tolerant traffic can squeeze out congestion intolerant traffic, even though the congestion intolerant traffic may intrinsically be much more valuable.

distributed between 0 and 1 throughout the population. We assume that the clutter effect function F_c is of the form: $F_c(z) = \lambda z^2$.

Let p be the transport price of these goods (which are priced the same even in the aware architecture since they are essentially identical). Suppose a network offers x goods; then a connected user α receives utility

$$U_\alpha = \int_{1-x}^{1} (v^\alpha - p)ds - F_c(x)$$

By solving for $U_\alpha = 0$ we find that all users with $v^\alpha > p + \lambda x$ are connected, and thus, the fraction of connected users and the total demand for each offered good is $1 - p - \lambda x$. Total profit is calculated as demand times price times the number of goods, or $\pi = (1 - p - \lambda x)px$.

An aware network can set both p and x to maximize profit; the maximizing values are $p = \frac{1}{3}$ and $x = \frac{1}{3\lambda}$, yielding a demand of $\frac{1}{3}$, a revenue of $\frac{1}{9}$ and a consumer surplus of $\frac{1}{6}$ per unit good. Goods are offered as long as they yield revenue at least $\frac{1}{9}$; at $x = \frac{1}{3\lambda}$, any additional goods would yield less revenue. Although the example is stylized, in that each good is essentially the same, it illustrates the result that goods are offered based on a profit ordering in an aware network.

A network with a blind architecture can set p but not x; content providers will enter the market as long as some customers are willing to buy. For any p, as long as the demand per good $1-p-\lambda x$ is positive new goods will enter, ultimately driving the demand to zero. Thus, in the blind architecture no stable equilibrium can have positive demand, revenue, or consumer surplus. Given the set of offered goods, a new good will be purchased (and profit will be positive, so it will be offered) as long as at least one customer values it more than the transport price; that is, if $\max_\alpha v^\alpha > p$. Since the optimal price is $p = \frac{1-\lambda x}{2} < 1$ and $\max_\alpha v^\alpha = 1$, all goods are offered, until congestion crowds out all users. This illustrates both that goods are ordered by maximal user value in a blind architecture, and that clutter creates a congestion externality.

In Appendices B and C we present two more examples illustrating aspects of the clutter phenomenon. The example in Appendix B shows how consumer surplus and total welfare can increase with increasing p in a blind architecture. The example in Appendix C provides a particularly clear illustration of the externality aspects of clutter.

4.3 Attention Effects

In the previous subsection we explored the fact that when many goods are purchased by an individual consumer, her satisfaction may be significantly less than the sum of the v_i^α's. We assumed that the whole may be less than the sum of the parts due to clutter: the more goods or applications available, the harder it is for the user to find what she wants. Another possibility is that users have limited attention, and their enjoyment of a given good is decreased when they are also consuming other goods. There has been considerable recent discussion of the extent to which we are moving from a service economy to an *attention* economy (see, e.g., [2]). For example, a subscription to HBO has reduced value if the user also subscribes to additional movie channels. This is a special case of the goods being (imperfect) substitutes for each other. Note that clutter depends on the goods offered, while attention depends on the goods consumed.

We model the attention effect quite similarly to the way we modeled clutter: if consumer α has purchased a set G_α of goods, then the total valuation is $\sum_{i \in G_\alpha} v_i^\alpha - F_a(|G_\alpha|)$ where F_a is some nondecreasing function. Each consumer then chooses G_α to maximize the sum $\sum_{i \in G_\alpha}(v_i - p_i) - F_a(|G_\alpha|)$.

It might appear that attention effects are a modest variant on clutter effects, and not worth the bother of further modeling. Indeed, we initially conjectured that the same hypotheses we held for clutter effects would apply:

- **Externality**: Content value is reduced as the demands on a user's attention increase. An aware network provider can serve a positive editorial function, increasing the value of the net by controlling the offerings.

- **Content Ordering**: Since the aware net provider controls the offerings, content selection would be ordered willingness-to-pay under a blind architecture, and by profit under an aware architecture.

In fact, both of these conjectures turn out to be wrong. With a little thought, it may be obvious why they are wrong. However, we think it is worthwhile to explore how attention effects differ from clutter effects. For one thing, demand interactions of this sort surely *are* relevant for an aware net provider selecting content (though not in the way we originally guessed). In addition, the economics of service architecture is a novel topic, and it is instructive to understand how very slight differences in modeling assumptions can lead to quite different results.

4.3.1 No Externality

The first thing to notice is that the attention effect *does not* create an externality. The $F_a(\cdot)$ term represents interdependence in a single user's utility function; it is unaffected by what goods other users consume. Thus, there is nothing intrinsic about the attention affect that favors the aware architecture over the blind.

We use a simple example with homogeneous users and heterogeneous goods to illustrate how the users and the network respond to attention effects. Consider a continuum of goods labeled $x \in [0, 1]$, where x is also the value of the good to all consumers. We suppose that the attention cost function is given by $F_a(z) = \frac{1}{2}(z - \frac{1}{8})_+$. In a blind architecture, a user will purchase all goods that have value $x > q \geq p$. The value cutoff q will generally be greater than the price p because for each good consumed the user pays both p and an incremental attention cost, F'_a. Therefore, surplus is

$$U = \int_q^1 (v - p) dv - F_a(1 - q).$$

The consumer chooses to purchase all goods with value greater than $q = \frac{1}{2} + p$, where the $\frac{1}{2}$ is simply the marginal attention cost. Substituting back into the surplus function and taking the derivative with respect to p we find that

$$\frac{\partial U}{\partial p} = \begin{cases} -\frac{1}{2} + p, & \text{if } p \leq \frac{3}{8} \\ -1 + p, & \text{if } \frac{3}{8} < p \leq 1 \end{cases} .$$

For all feasible values of the transport price, surplus is decreasing in price. There is no externality, and reducing the set of offered goods (by raising price) can never make consumers better off.

4.3.2 Ordering by value or profit?

The attention effect is a function of consumed goods, not offered goods, and users choose the goods consumed. As a result, it turns out in both architectures that the goods are generally ordered by maximal willingness-to-pay rather than total revenue. Why is this result different from the ordering by revenue for the aware architecture with gateway or clutter effects? Both the gateway and the clutter cost are a per good cost borne by the network as a whole. The first is a direct cost that must be paid to make an application available to anyone on the network; the second is an indirect cost (in the form of reduced consumer surplus available for extraction) that again arises from making an application available to anyone. The attention cost is a per user cost per good consumed by that user. Since there is no network-wide cost to be recovered, the network offers goods based on individual consumer valuations, not total revenue.

As before, the result is trivial for a blind network: the net provider merely sets the uniform transport price, p, and all goods that some consumer values more than p are offered (consumer valuation is net of the marginal attention cost induced by consuming the good). For an aware network, the result is also straightforward. The net provider should offer all goods for which $\max_\alpha[v_i^\alpha - F_a'(|G_\alpha|)] > 0$.[9] To see why this is so, suppose not for good j: if the network adds j to its offerings, then at least one user would experience higher utility by purchasing it (and there are no external effects on other users), and thus the net could charge a $p_j > 0$ and increase its profits.

4.3.3 Attention Effects and Social Welfare

It is easy to determine the welfare-maximizing outcome. Each user orders all of the goods from highest v_i^α to lowest. The user then adds the goods to her consumed set in order until $v_j^\alpha < F_a'(j)$. The union of the goods desired by all users is the optimal offered set.

Neither architecture will achieve the socially optimal provision of content or applications if the net provider is a monopoly. In an aware network, the provider chooses prices $p_i > 0$. Now consumers order the goods by $v_i^\alpha - p_i$. Although the ordering may be different, goods are now consumed only until $v_j^\alpha - p < F_a'$, so fewer goods are consumed (and offered). The same argument applies directly to a blind network (except that the ordering is the same).

Although network cost recovery (using the simple pricing mechanisms we consider here) lowers social welfare, there is no market failure as there was with the clutter effect. The result is simply the usual monopoly result: the net provider restricts the output in order to raise prices above marginal cost and earn supracompetitive profits. The editorial role in an aware network is not an intrinsic advantage in dealing with the attention effect: private consumers internalize the problem and solve it themselves.

This is not to say that there are no differences between the outcomes of aware and blind architectures in a model with attention effects. The aware monopolist can charge a different price for each good or application; the blind monopolist

[9] We abuse the notation slightly by referring to the incremental change in attention costs as the derivative F_a'. In fact, the argument changes by integer values, and the incremental change is a first difference.

chooses only one price. As a result, if we ignore the cost factors considered in Section 4.1, profits will be higher and more goods will be offered in an aware network. To see the latter result, recall that in an aware network all goods are offered for which $\max_\alpha[v_i^\alpha - F_a'] > 0$, whereas in the blind net only goods for which $\max_\alpha[v_i^\alpha - F_a'] > p > 0$ are offered. However, though more different goods are offered in an aware net, $p_i < p$ for some i, $p_j > p$ for some j, and the net effect on social welfare is ambiguous.[10] In any case, these characterizations are not special to a model with attention effects: they are the typical results for monopoly price differentiation when there are no external costs.

5 Implications for Content Creation

In Sections 3 and 4 we considered goods that had already been created and investigated how the offered set of goods depended on the choice of architecture. In this section, we ask how architecture affects decisions to create content and applications in the first place. We have identified two relevant effects, which we discuss in separate sections below.

5.1 Initial Profits as an Incentive to Enter

We have assumed that the market for content provision is competitive, and that already created content is provided at marginal cost (which we have assumed to be zero for convenience). What would induce a content provider to enter such a market? The typical story is that the early entrants to the market do not face perfect competition and so are able to extract significant profits before perfect competition sets in. This story of initial profits must be applied carefully in our case, since we have assumed that the network provider is not only a monopolist but also a monopsonist. Consequently, an aware network provider may be able to set the price for content of each good (i.e., what the content provider is paid) at the marginal rate, even in the absence of content competition. The incentive to enter new content markets in an aware architecture may be small. However, in the blind architecture, the network provider can only set a single transport price p that applies to all goods. The initial content provider can extract a profit $D_i(p_i^{max})(p_i^{max} - p)$ from the market (where D_i is the demand function for good i). Thus, the blind architecture offers significantly greater incentives to develop goods for high-value (i.e., greater than p) content markets.

[10]That is, more different goods are offered, but less is consumed of some of the goods.

5.2 Gateway on the Market

With an aware architecture, the network provider decides which content will be offered; in a blind architecture the network cannot discriminate between offerings, and so the content creators decide what will be available. As we discussed earlier, this difference affects the *selection* of available content: maximal profit in an aware network and maximal value to a user in a blind network. There is another obvious difference that is likely to affect content *creation*: a single firm (the network provider) decides in an aware network, while multiple firms each make independent decisions in a blind network. We expect that in a blind network there will be more experimentation with content creation, and collectively more risk-taking (not because of differences in risk aversion, but because of differences in beliefs about the likely success of projects). We have not yet developed a complete model of this phenomenon, but we illustrate the principle with a simple example.

As before, assume there are no fixed or variable costs to offering a content good, once created. However, there is a cost of creation I_i. Therefore, before creating the good the condition $E[R_i(p) > I_i]$ is evaluated by every potential content provider in a blind network, whereas the condition $E[R_i^{max} > C_i + I_i]$ is evaluated only by the network provider if the architecture is aware.[11] When there is no uncertainty about demand and cost conditions, the in *who* does the evaluation difference is immaterial; when there is significant uncertainty then the difference can be crucial.

To see this consider an extreme but illustrative situation with K content providers. Each good has probability $0 < \epsilon < 1$ of having high value ($R_i(p) > I_i$ and $R_i^{max} > C_i + I_i$), and probability $(1 - \epsilon)$ of having low value ($R_i(p) < I_i$ and $R_i^{max} < C_i + I_i$). Moreover, the content providers and the network provider have independent, identically distributed beliefs about each good; with probability $0 \leq \delta < 1$ they believe that the good has high value (in the sense just mentioned) and with probability $(1 - \delta)$ they believe that the good has low value.

The probability of any particular good being offered in the blind architecture is $1 - (1 - \delta)^K$. However, in the aware architecture the good is offered only with probability δ. Thus, the ratio of successful goods offered in the two architectures in this example is $\frac{1-(1-\delta)^K}{\delta}$ which, for small δ and large K, is very large.

This example indicates that if the perceived probability of success is low, then the blind architecture is much more likely to offer goods; some content provider will eventually experiment with it. In the aware architecture, since

[11] Recall that C_i represents liability and gateway costs borne by the network provider for each good or application offered.

the network provider is the only one making this evaluation, there is much less experimentation. What is the effect on welfare? The answer is not simple because the greater number of successful goods in the blind architecture are offset by the associated greater expenditure on developing unsuccessful goods. However, for the content providers' decisions to be rational in expectation, it must be that on average they are sufficiently compensated from successful goods to compensate for the risk of losses on unsuccessful goods. If consumers are willing to pay enough to support the investment in both the successful and unsuccessful goods, then the basic welfare effect should be clear: the greater number of successful goods created in the blind architecture generate greater total surplus.[12]

6 Discussion

One key distinction between various competing visions for the NII is the extent to which the network provider is aware of the content of the bits it is conveying to consumers. Aware architectures are aware of the content, while blind architectures are not. Our focus, in this paper, is the impact this architectural distinction has on the provision of content.

The most obvious, and important, difference between the blind and aware architectures is that of price differentiation; aware architectures can price based on content (or application-type); blind architecture cannot. Other important differences include gateway/liability effects, which are costs borne by the aware network provider for each offered good, and clutter and attention effects, where costs are imposed on consumers for each good offered or consumed.

The general impact of architecture on the provision of goods is complicated, and depends on the detailed distribution of consumer preferences. We can, however, describe more precisely the impact in several limiting cases. When the variation in v_i^α is mostly in i, so most users have the same preferences but their valuations of various goods differs widely (i.e., $v_i^\alpha \approx v_i^\beta$ but $v_i^\alpha \not\approx v_j^\alpha$) then the aware architecture produces higher revenue and higher total welfare; the blind architecture produces higher consumer surplus. In contrast, if the total potential welfare is dominated by niche goods and gateway/liability costs are significant, then the blind architecture offers higher revenue, total welfare, and consumer surplus. Thus, in both of these cases consumers would prefer the blind architecture. However, if clutter is the dominant effect, then the aware architecture can pre-

[12] We are working on a more complete analysis of this phenomenon.

vent the tragedy-of-the-commons overprovision of goods that plagues the blind architecture. In this case, revenue, total welfare, and consumer surplus are all maximized in the aware architecture. The analysis of the attention affect emphasizes the importance of carefully modeling the way in which costs arise: although this effect bears striking resemblance to the clutter effect, there is no commons externality, and welfare may be higher under either architecture, depending on the distribution of preferences.

Architecture also has important implications for the creation of goods. Our analysis suggests that the blind architecture may encourage more vigorous creation of content than the aware architecture for two reasons. First, the blind architecture leaves more profits for early entrants to markets, thereby encouraging the rapid filling of unmet consumer needs. Second, in the presence of significant uncertainty about the profitability of various goods, networks with a blind architecture benefit from the diversity of beliefs among content creators; taken as a collective whole, the community of content creators experiments widely with new goods in the blind architecture. The increased content creation in the blind architecture not only increases consumer surplus, but can also lead to a higher level of revenue for the network provider because of the increased total demand for goods.

If we ignore the liability/gateway effect, one might argue that since the aware architecture can set its prices to mimic the blind prices, aware architectures can foster an equivalent level of content creation. However, note that the blind architecture has committed to a common carriage access policy and a flat pricing policy which leaves profits for early entrants to content markets. On the other had, the aware architecture can always revert back to access and pricing policies that maximize its short-term revenue (i.e., the revenue it can collect from the present set of goods), and this might discourage content providers from even entering the market. Thus, perhaps the most important aspect of the blind architecture is that it commits the network provider to leave some profits on the table for the content creators.

References

[1] U.S. District Court. Religious Technologies Center v. Arnaldo Pagliarina, et al., 1995. E.D. Virginia, Civil Action No. 95-110.

[2] Richard A. Lanham. *The Electronic Word: Democracy, Technology, and the Arts*. University of Chicago Press, Chicago, 1993.

[3] Peter H. Lewis. Judge allows libel lawsuit against Prodigy to proceed. *New York Times*, May 26:C4, 1995.

[4] Jeffrey K. MacKie-Mason and Hal R. Varian. Pricing congestible network resources. *IEEE Journal of Selected Areas in Communications*, 13(7), 1995. Available at URL:[13].

[5] Scott Shenker, David D. Clark, and Lixia Zhang. Services or infrastructure: Why we need a network service model. In *Proceedings of Workshop on Community Networking*, pages 145–149, 1994.

[6] Daniel Shiman. When e-mail becomes junk mail: The welfare implications of the advancement of communications technology. Technical report, State University of New York-Oswego, 1995.

[7] Hal R. Varian. Price discrimination. In Richard Schmalensee and Robert D. Willig, editors, *Handbook of Industrial Organization*, volume I of *Handbooks in Economics*, chapter 10, pages 597–654. Elsevier Science Publishing, New York, 1989.

A Imperfect Substitutes and Liability/Gateway Effects

Previously we have only considered goods whose valuations were additive. IN this example we explore an effect due to imperfect substitutability of goods. Consider the case where there is a single *generic* good, say good # 1, with broad and uniform appeal: $v_1^\alpha = 1$ for all α. Assume there are K groups of consumers, each comprising an equal share of the population, and for each such group there is a "custom" good whose value is slightly better than that of the generic good (and groups do not prefer other group's custom good to the generic good). Let $\epsilon < 1$ be the incremental increase in willingness-to-pay. That is, $v_2^\alpha = 1 + \epsilon$ if good 2 is the custom good for consumer α. However, because these goods are imperfect substitutes for each other, the user's total utility is merely the maximum valuation of the consumed goods. We set the gateway/liability costs of all goods to be $C < 1$.

[13] ftp: //gopher. econ. lsa.umich.edu/pub/Papers/pricing-congestible.ps.Z

In the blind case, all of these custom goods are offered and the generic good is not. The additional welfare generated by the customizing of goods is ϵ (normalizing the size of the population to 1).

In the aware architecture, if $(K-1)C < \epsilon$ then the custom goods are offered and the additional welfare generated by the customizing of goods is $\epsilon - (K-1)C$. If $(K-1)C > \epsilon$ then the custom goods are not offered, only the generic one is, and there is no additional welfare generated by the customizing of goods. In both cases the loss of welfare (compared to the blind case) is $\min[\epsilon, (K-1)C]$. This example suggests that aware architectures would may inhibit small incremental increases in quality through customization or slight differentiation. It also suggests that if high degrees of customization of a basic generic good is the dominant effect, then the blind architecture will produce higher total welfare.

B Clutter and Welfare

Consider an example with a continuum of goods labeled by x, with $x \in [0, 1]$. Each individual consumer values each of these goods an amount x (i.e., the x'th good is valued an amount x). The clutter effect function F_c is: $F_c(z) = \frac{1}{2}(z - \frac{3}{8})_+$. In the aware architecture, the entire consumer welfare can be extracted, and so if x is the minimal level of good provided (i.e., all goods with values greater than x are offered) then the total revenue is: $R = \frac{1-x^2}{2} - \frac{1}{2}(1 - x - \frac{3}{8})_+$. This is maximized when $x = \frac{1}{2}$ and $R = \frac{5}{16}$. In the blind architecture, all goods have the same price p and all goods with $x \geq p$ are offered. Since $R = p(1-p)$, revenue is maximized at $p = \frac{1}{2}$ and the resulting revenue is $R = \frac{1}{4}$, which is less than the aware revenue. The consumer surplus is $S = \frac{1}{2}(1-p)^2 - \frac{1}{2}(1 - p - \frac{3}{8})_+$, which takes on the value $\frac{1}{16}$ when $p = \frac{1}{2}$. For $0 < p < \frac{1}{2}$ R increases with p but S decreases with p. When $\frac{1}{2} < p < \frac{5}{8}$ the roles are reversed and S increases while R decreases with p; in this range increasing the price decreases the clutter effect more than it increases the revenue extracted. In fact, the consumers would be better off with higher prices than the network provider wants to charge! For $p > \frac{5}{8}$ there is no more clutter and both R and S are decreasing with p.

Note, that in this example even though the consumer surplus increases with increasing p in the region $\frac{1}{2} < p < \frac{5}{8}$, the total surplus (revenue plus consumer surplus) decreases in this region. However, if we modify the example slightly, so that $F_c(z) = \frac{1}{2}(z - \frac{1}{2} + \epsilon)_+^{\frac{1}{2}}$ then, for sufficiently small $\epsilon > 0$, the total welfare and the consumer surplus are increasing in the region $\frac{1}{2} < p < \frac{1}{2} + \epsilon$, even though the blind revenue is maximized at $p = \frac{1}{2}$. This last statement can be verified by noting that the price derivative of the total welfare is $D'(p)(p - F'(D(p))$ where $D(p)$ is the demand function. Thus, the total welfare increases with p if and only if $F'(D(p)) > p$.

C Clutter as an Externality

One key aspect of the clutter effect is that the offering of a good affects all consumers negatively (actually, it only affects those consumers who remain connected), but may only offer some users benefit. This negative effect can be controlled by the aware architecture, but not in the blind architecture.

Suppose there are many possible goods, and that each good appeals to an entirely separate group of like customers. That is, for each good i, there are f_i potential customers, each of whom value the good at v_i, and no consumer wants more than one good. Order the index numbers for the goods so that $v_i > v_{i+1}$. Everyone bears the same clutter costs, which depend solely on the number of goods offered: $F(N)$.

In a blind architecture, the network provider can set a single transport price. Customers will participate in the network only if their surplus is positive, so we obtain an individual rationality (IR) constraint for each group of the form $v_i - F(N) - p > 0$, given N. Given the ordering of the goods, $N = \arg\min_i [v_i - F(N)]$, so the IR constraint can be binding for the least valued good, at most. A profit-maximizing network will thus set $p = v_N - F(N)$, conditional on its choice of N. That is, the price will be set to extract all of the surplus of the group with the lowest surplus. The network then chooses N to maximize its profits where the profit function is $\pi(N) = [v_N - F(N)] \sum_{i=1}^{N} f_i$.

Now we can consider some welfare consequences of the clutter externality. In general, we would expect that because clutter imposes an externality, there will be *too many* goods in equilibrium. The first way to investigate this conjecture is to ask whether welfare increases if the price were increased. Consumer surplus at price $p(N-j)$ is

$$\begin{aligned} S(N-j) &= \sum_{i=1}^{N-j} f_i[v_i - F(N-j) - p(N-j)] \\ &= \sum_{i=1}^{N-j-1} f_i[v_i - v_{N-j}] \end{aligned}$$

where the simplification results from substituting in $p(N-j) = v_{N-j} - F(N-j)$.

The change in consumer's surplus from raising price from the profit-maximizing level $p(N)$ to any higher price that is profit-maximizing for the smaller number of offered goods is:

$$\begin{aligned} S(N-j) - S(N) &= \sum_{i=1}^{N-j-1} f_i[v_i - v_{N-j}] - \sum_{i=1}^{N-1} f_i[v_i - v_N] \\ &= -\sum_{i=1}^{N-j-1} f_i[v_{N-j} - v_N] - \sum_{i=N-j}^{N-1} f_i[v_i - v_N] < 0. \end{aligned}$$

Thus, forcing the network provider to charge a higher price (say, by imposing a tax) discouraging some low-value goods from being offered will in fact lower consumer surplus, despite the clutter externality. Since we already know that a higher price will also lower profits, it follows immediately that a higher price will also lower total welfare.

This does not mean that the optimal set of goods is being offered, however. Suppose it were possible to pick and choose which goods would be offered, which is precisely what an aware network can do. How would total welfare be changed by eliminating a single good? Let G_{-j} refer to the set of goods when the jth good is removed from the profit-maximizing set. The change in total welfare from removing the good is

$$\begin{aligned} W(|G_{-j}|) - W(N) &= \sum_{i \in G_{-j}} f_i[v_i - F(N-1)] - \sum_{i=1}^{N} f_i[v_i - F(N)] \\ &= -f_j[v_j - F(N)] + [F(N) - F(N-1)] \sum_{i \in G_{-j}} f_i \end{aligned}$$

The first term reflects the loss of surplus from excluding the jth group of customers; the summation reflects the reduction in clutter costs that offering the jth good imposes on all other customers. When the clutter savings are large relative to the surplus from the jth good, welfare would increase by excluding the good. This will tend to be the case for niche goods (small f_j small relative to $\sum_{i \in \tilde{G}} f_i$), and, of course, when the marginal clutter cost $F(N) - F(N-1)$ is large.

Does an aware network necessarily deal better with the congestion problem? Yes: aware profits are maximized by extracting the full consumer's surplus with $p_j = v_j - F(N)$, so the first term in the expression is the negative revenue from good j, and the aware network drops goods with the lowest revenue until this foregone revenue exceeds the additional revenue (surplus) that can be extracted from other users as clutter decreases. Thus, the aware network completely solves the externality problem, and the welfare-maximizing set of goods is offered.

As we have seen, when there are clutter effects an aware network will order content selection by maximal profit, while a blind network will order by maximal willingness-to-pay, just as with the liability/gateway effect. However, with a clutter externality, too many goods will be offered in a blind architecture, particularly too many niche goods. Forcing the network provider to raise its transport price may not solve the clutter problem, and in fact may reduce both consumer surplus and total welfare. Adopting an aware architecture will internalize the clutter externality, and in the special case in which the network can extract all surplus, will even result in the socially optimal set of goods.