Spiders on Real Estate Platforms*

Dries De Smet† and Patrick Van Cayseele‡

March 15, 2010

Abstract

Spiders that automatically collect information on properties for sale on broker's websites, are both an opportunity and a threat for traditional real estate marketing platforms. When the traditional platform decides to allow the spider to copy his content, he takes into account two adverse effects. First, the spider might redirect buyers that would otherwise not have visited the platform. Second, buyers interested in properties for sale by brokers now are sent directly to the broker's website, skipping the platform.

In a simple theoretical model, we find that the platform will allow the spider if his own familiarity is low and if the share of houses sold by brokers is low. Dynamic effects on these parameters make the platform more cautious. We illustrate this model with two case studies on real estate platforms, in the Netherlands and Belgium.

1 Introduction

The rise of the world wide web has changed many business, for the better or the worse. For most companies, the internet means an interesting way to save time, by the ease of sending and retrieving information. But the internet has probably the biggest impact, apart from the effect on those who facilitate information technologies (such as software and application developers, hardware constructors), on commerce and advertising.

The internet is very suited for e-commerce. On the seller side, it eliminates the costs of a physical store. On the buyer side, it eliminates search costs and transportation costs. The internet also proved to be very important as an advertisement tool, which is what these paper is all about. It is not hard to develop a website which is easily accessible for everyone from everywhere. Besides, given the enormous growth of leisure time spent on the internet, it is also an attention-attracting device where advertisers can get a piece of the attention-pie.

One particular sector we investigate in this article, is the real estate sector. In most countries, there are two ways to sell a house. First, an owner can try to sell a house by himself. Or, second, one can enlist a broker's help for the advertisements, the visits and the whole selling process.

*We thank Frank Verboven and Thomas Provoost for helpful discussions and comments.

†dries.desmet@econ.kuleuven.be, Center of Economic Studies, K.U.Leuven (corresponding author)

‡patrick.vancayseele@econ.kuleuven.be, Center of Economic Studies, K.U.Leuven; Faculty of Economics and Econometrics, University of Amsterdam; LICOS Centre for Institutions and Economic Performance, K.U.Leuven

The rise of the internet created some opportunities, both for brokers and private sellers. Brokers can set up a website where they announce the houses for sale. In principle, a private seller can do that too, but given the non-recurring nature of selling a house, this will be either too costly or impossible to appear on the top of the search results of search engines. Besides, classified ads platforms where knocked up, quickly taking over the role of traditional classified ads in (free) newspapers. These platforms exist on a very general level (such as craigslist) but also more specific, such as job boards, dating ads and real estate websites. Especially for private sellers, such a platform is an interesting way to announce the sale. Because these platforms attract many potential buyers, it also becomes interesting for brokers. These platforms earn money either by charging the classified ads or by placing banners or other advertisement types.

A recent trend in real estate platforms, is the emergence of spiders. These programs, also known as bots, crawlers, harvesters or scrapers, collect information from websites in an automatic way. The most known are pricebots or shopbots: programs that collect product information and prices for specific products from different distributors. Mostly this information is presented on a spider website, also known as a metasite. In the context of real estate, a spider can be used to collect information on houses sold from different broker's web pages. It automatically notices updates on websites and provides an up-to-date overview of properties for sale. This overview can be map-based (such as real estate on Google Maps), or list-based (with possibilities to narrow the search results).

For traditional non-spider platforms, these spiders might be both a threat and an opportunity. On the buyer side, the spider website can attract new potential buyers and refer them to the traditional platform. But the spider might also lure away potential buyers from the traditional platform, either by sending them to competing platforms or by eliminating the need to search on the traditional platform (by providing enough spidered information). On the seller side, allowing a spider to copy content might give the private sellers a higher value proposition in case of an increased amount of potential buyers. Though, a platform might loose on the brokers, since their websites are automatically included on the spider website. Especially if the spider website becomes the preferred site to start searching for real estate, it is likely that the spider links property for sale by a broker directly to the broker's website. This eliminates the need for the broker to be on the traditional classified ads platform.

The decision of traditional classified ads platform whether to allow a spider to collect and represent information lies at the heart of this paper. We investigate the decision of the traditional platform, when confronted with a spider and explain the mechanics that can influence this decision.

The remainder of this paper is organized as follows. The next section provides the necessary background and a partial survey of the relevant previous contributions and lists some crucial insights. In section 3, the model is set up and the results are analysed. Section 4 discusses some extensions to the model, by endogenizing some parameters or adapting the initial model. Section 5 confronts the theoretical results with empirical observations for Belgium and the Netherlands. A sixth section concludes with what we consider promising avenues for further research and implications for policy making.

2 Background and Related Literature

At the heart of this article lies the desirability of spiders, as seen from the viewpoint of both traditional platforms.There is quite some literature about spiders, though most articles are published in the research fields of information technology and legal analysis. Especially recent legal battles on spiders might prove interesting, given the different argumentations of courts to forbid or allow spidering. Though, before analyzing the research in law and the jurisdiction, we first analyze the potential threats and opportunities of spiders. Afterwards, we also have a look at the economic concepts that might help to understand the problem at hand.

2.1 Opportunities and Threats of Spiders

Spiders can be seen as an opportunity or a threat for traditional websites. We list the opportunities (O1) and threats (T1-T4) below.

O1-Traffic The existence of a spider is an opportunity, since it provides users with an overview of all existing products/websites, resulting in a click-through to the website. In this sense, the spider website is a traffic generator for the traditional website. Users who otherwise would not have found the product, service or website are now relegated by links to the website.

T1-Traffic It can also be the case that the spiders steals traffic. Since the spider aggregates information from different sites, it will be more interesting for users to surf to the spider in stead of the traditional website. If visitors of the spider do not click through to the traditional website, then the net effect will be a loss in users. As a consequence, this might also lead to a loss in advertisement revenues for the website.

T2-Competition If information is easily available for consumers, then search costs are reduced. The likely effect of reduced search costs is lower prices (Stahl 1989), i.e. increased competition.

T3-Control Company websites can promote their products or services by showing them in their most attractive way. If the information is aggregated by a spiders, it might be the case that this representation ignores the unique selling point of the products or services. Even worse, the information might be deteriorated by the spider. In sum, the company looses control over their own content.

T4-Overload The technique of spidering might have a detrimental effect on the capacity of a website, leading to a devaluation of the quality for other users. In the case eBay vs Bidders Edge[1], Bidders Edge admitted that it sent some 80,000 to 100,000 requests to Ebay's computer systems per day. eBay claimed that this takes unnecessary bandwidth and capacity, while Bidders Edge said that its searches represent a negligible load. Anyhow, certainly if many spiders are active on the net, this can deprive too much capacity and lower the speed for normal surfers.

[1]See Fischer (2001) for a discussion and further references.

2.2 Legal Actions against Spiders

If the disadvantages of spiders outweigh the advantages, the websites can try to protect themselves against spiders.

The easiest way to stop a spider is to send a 'cease and desist' letter. This is a unilateral order to stop the activity, or else face legal action. In some cases this might not be sufficient and only legal steps can force the spider to halt its activity.

Jennings & Yates (2009) summarize some arguments that courts used to forbid spiders.

Spidering might be a copyright infringement. Websites contain a lot of elements protected by copyright, such as texts and coding (literary works), photographs and animations (artistic works), music and audio content (musical works) and lay-out. In the case Shetland Times vs Wills[2], the Scottish court argued in 1997 that copying the news headlines, even with linking to the relevant article, can be considered as a copyright infringement. In the case of real estate web sites NVM vs Zoekallehuizen.nl[3], the Dutch court argued in a preliminary statement that copying the address, price and 1 to 1.5 rule description are not considered as a copyright infringement, but as a quote. The same holds for pictures, given that their size is sufficiently reduced.

Similar to the argument of copyright infringement, some countries also have a database regulation in order to protect the content of database holders. The British regulations for example, stipulate that "a person infringes the database right if, without the consent of the owner of the right, he extracts or re-utilizes all, or a substantial part, of the contents of the database" (Jennings & Yates 2009). This argument was used in the German case Stepstone vs Ofir and the French case Cadremploi.fr vs Keljob[4].

Another argument is related to the terms of use that a website puts on its own website. The auction site eBay puts in its terms of use "You agree that you will not use any robot, spider, scraper, or other automated means to access the sites for any purpose without our express handwritten permission."[5] In the case of real estate associations Canadian Real Estate Associations vs Sutton Quebec Real Estate Services[6], the courts argued that Sutton should have been familiar with the terms of use. The question is whether these terms of use are enforceable. In the Dutch case NVM vs Zoekallehuizen.nl, the court argued that such a declaration does not bind the users. Since the real estate information is made publicly available, other parties can use this information as long as this does not infringe copyright or other rights.

Probably the most used argument is trespass to chattels, i.e. the unauthorized use, dispossession or interference with the tangible property of another. While this early common law required a physical touching of another's chattel, modern interpretation stipulates that indirect touching is sufficient. In recent history, trespass to chattels has been used in many technological cases, such as telephone lines, radio, television broadcastings and e-mail spam (Fischer 2001). This claim was used in the case eBay vs Bidders Edge[7] and

[2]See Jennings & Yates (2009) for further references.

[3]Arrest nr AV5236, Court Arnhem, 16-03-2006. Available on `www.rechtspraak.nl`.

[4]Idem footnote 2.

[5]eBay, Your User Agreement, accessed at the 11th of March, 2010, `http://pages.ebay.com/help/policies/user-agreement.html`.

[6]Idem footnote 2

[7]Idem footnote 1.

Oyster Software vs Forms Processing[8].

2.3 Other Actions against Spiders

Besides legal actions against spiders, a site can also implement technical protection measures. An easy way is blocking the IP address of the spider, denying him access to the website. This might prove not sufficient since it is easy to circumvent. Another common way is putting a robot.txt on the web space. This text document indicates which subsites should not be accessed by the robot. This standard is easy to implement but similarly to the terms of use, the question arises whether spiders are bound by this exclusion.

If there is a net disadvantage from spidering for traditional websites, then a middle course might be that the spider makes side payments to compensate the losses of traditional websites. This scheme exists already for Google Fast Flip, a website from Google that shows snap shots from news websites such as New York Times, Washington Post and Newsweek. Google Fast Flip shows advertisements on it web site and shares the revenues with the news websites[9]. Note that the side payments can also go in the other direction, i.e. traditional websites paying the spider to be spidered. If there is a net advantage from spidering, then traditional websites will be willing to pay for the service. For a further discussion on side payments, see also section 4.

Note that some legal actions might also be inspired by the aim for a settlement on side payments. In the Belgian case Copiepresse vs Google, the journalist Danny Sullivan[10] argued that is was perfectly possible to stop Google accessing the news pages of newspapers by simply putting a robots.txt file on the website. It is likely that Google would have complied voluntarily with this request. Sullivan poses:

> This case was never about getting content out. It was about trying to blackmail Google into including content. Now the newspapers may get a large fine coming their way, but whether Google will feel it still wants to cut a deal with them remains unseen.

2.4 Related Literature

Most scientific articles on spiders are written either in the field of information technology or law. The former mostly stresses the technical details to build a spider, or the impact of spiders on website performance. Law articles focus on the arguments used by courts (e.g. Jennings & Yates (2009), Fischer (2001)), though some articles investigate the economic efficiency of spiders (e.g. Rosenfeld (2002), Short (2004)). Our article joins this literature by examining the desirability of spiders. Besides, this article is related to many economic concepts, which we describe below.

First, the platforms under investigation are two-sided markets, as described by the seminal contributions of Rochet & Tirole (2003), Armstrong (2006) and Caillaud & Jullien (2003). The main problem of platforms is to get both sides of the market on board. In

[8] Oyster Software vs Forms Processing, 2001 WL 1736382 (N.D.Cal. 2001, United States.

[9] See the article Read news fast with Google Fast Flip on the official Google Blog, `http://googleblog.blogspot.com/2009/09/read-news-fast-with-google-fast-flip.html`.

[10] Sullivan, Google Loses In Belgium Newspaper Case, 13-02-2007, `http://searchengineland.com/google-loses-in-belgium-newspaper-case-10500`

the context of real estate platforms: a platform without potential buyers would be useless for sellers and vice versa. We use these so-called cross-side network externalities when we develop the utility functions of both sides of the market (buyers and sellers). Related with two-sided markets are search and matching models, because they also focus on bringing together two or more distinctly different demands. Hendel, Nevo & Ortalo-Magné (2009) apply the matching model of Coles & Muthoo (1998) to real estate marketing platforms. While the platform choice is based on a stock-and-flow model, their model has some similarities with ours since they also incorporate limited knowledge on the existence of platforms. The choice for sellers is between a well-known but more expensive platform and a less-known and more illiquid platform. Interestingly, they also test this claim empirically by investigating a broker's platform and a 'for sale by owner' platform in Madison, Winconsin, (United States). Unfortunately, there is no active spider in this market.

Second, within this framework of two-sided markets, spiders can be seen as an application of vertical separation. Vertical separation means normally that the manufacturer does not sell its products directly to the consumers, but sells to a retailer who sells in turn to consumers (Bonanno & Vickers 1988). In two-sided markets, a platform normally connects one side (e.g. buyers) with the other side (e.g. sellers). Vertical separation can then be seen as two connected platforms, where one platform focuses on one side, whereas the other focuses on the other side. Consider again the market for real estate. A traditional platform aims at connecting both buyers and sellers. If a spider platform comes in the market, it can be seen as vertical separation, since the spider platform particularly aims at reaching buyers, while the traditional platform is now focused more on the seller side.

Third, in stead of the vertical separation model, a spider can also seen as a platform next to the other traditional websites, and therefore a direct competitor. If a spider copies the information of a traditional website, then it can be interpreted as the construction of an (asymmetric) adapter, in the tradition of Katz & Shapiro (1985). Spidering makes platforms compatible, since a surfer can also use the spider platform to access the content of the traditional platform. The adapter is asymmetric since the content of the spider cannot be accessed from the traditional platform.

Note that the discussion on the economics of copyright infringements by spiders is related to the discussion on digital music and movies a couple of years ago. Similar to the spidered websites, it has been shown that a potential copyright infringement for digital music should not necessary harm music distributors, due to the sampling effect (Peitz & Waelbroeck 2006). A similar discussion is whether the copying of pages from books by the Google Books project can be considered as fair use. Travis (2006) argues that "like the samples on iTunes and similar digital music services, the primary utility of Google Book Search will be to enable Internet users to preview works about which they lack the information to make a purchasing decision. [...] Google is salvaging entire libraries full of dusty, crumbling books while creating a highly efficient marketing platform for authors."

Our contribution clarifies the impact of spiders on traditional platforms and investigates whether these spiders might turn out beneficial for these platforms. We explicitly model the channels through which spiders might be (un)helpful for traditional classified ads platforms.

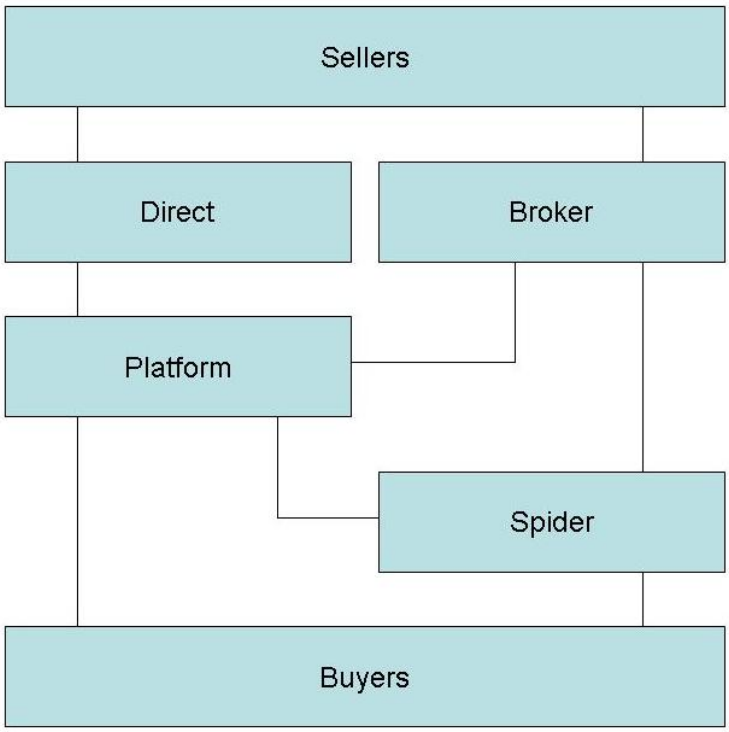

Figure 1: General Structure of Real Estate Market

3 Theoretical Model

In this section we present a theoretical model that helps to understand the decision of traditional real estate marketing platform. He decides whether to allow the spider to copy the content or not. We first explain the general structure of the market. Then we expound the assumptions and the relevant equations of our framework. The results are presented and interpreted in the third subsection.

3.1 General Framework

In the real estate market, the two most important groups of interest are the sellers and the buyers. Sellers bring their house on the market, buyers plan to acquire a house. Between the groups, there exist platforms that connect both sides. Figure 1 clarifies the structure of the market. Sellers can choose to sell their house privately, or they can enlist the help of a broker. The broker announces the 'house for sale' on his webpage. A spider can automatically retrieve the information on these webpages and present it on its spider site or metasite. We assume that brokers will never resist that their content is spidered, although the opportunities and threats for platforms may also hold for brokers (see section 2.1). Though, given that they want to announce their portfolio of houses on a traditional platform, it is also likely that they want to be spidered for free. Private sellers can only announce on a traditional classified ads platform who is a monopolist. Brokers can also use this platform to announce their properties for sale. The traditional platform charges the sellers, while the spider gets revenues from other sources (such as banners). Potential

buyers can surf to the platform and/or the spider. This depends on whether they know the platform or the spider, on the content (and how unique it is) and on the magnitude of their opportunity cost. As is typical in two-sided markets, the platform or the spider becomes more attractive the more sellers joined.

3.2 Set-Up

The traditional real estate marking platform, henceforth the platform[11], charges sellers willing to advertise on a pay per click basis. Buyers can access the website for free. We make abstraction from any fixed and variable cost. Therefore the platform's profit function reads

$$\Pi = p(S^p B^p + S^m B^m) \tag{1}$$

where p is the pay-per-click price, S^p,S^m is the amount of sellers who sell their house privately p respectively with a broker m[12]. B^p,B^m is the amount of buyers who click on the advertisements of private respectively brokers. Since we investigate the decision of the platform whether to allow the spider or not, we present two profit functions. In one regime spidering is allowed (indexed s), in the other regime spidering is not allowed (indexed n).

$$\begin{aligned} \Pi_s &= p(S_s^p B_s^p + S_s^m B_s^m) \\ \Pi_n &= p(S_n^p B_n^p + S_n^m B_n^m) \end{aligned} \tag{2}$$

The spider does not charge the content it spiders and is financed by other sources, e.g. banners. Therefore it aims at reaching as much potential buyers as possible (see the discussion on vertical separation, subsection 2.4). We do not model the profit function of the spider, i.e. we treat the spider as a mechanic player that spiders all the content of the brokers.

Sellers are charged on a pay-per-click basis. Since a spider does not charge for the content, all brokers will automatically join the spider. The fraction of sellers that sell their house with a broker is determined exogenously, i.e. it is determined by factors outside the model. This fraction is labelled μ. The rest $(1-\mu)$ sells its house privately. We assume that sellers derive a profit π from each buyer that sees the advertisement. Sellers are heterogeneous in this profit, π is distributed uniformly over the 0-1 space. Whether the sellers join or not, is independently of the decision of the platform to allow spidering, since sellers are charged on a pay-per-click basis. The platform cannot apply third degree price discrimination, therefore the price p is equal for all potential sellers. The demand functions read:

$$\begin{aligned} S_s^p = S_n^p &= (1-\mu)(1-p) \\ S_s^m = S_n^m &= \mu(1-p) \end{aligned} \tag{3}$$

Buyers are searching for real estate. They have an opportunity cost k for each website they open. This cost is weighted against the chance of finding a good house. This chance

[11] If we mention 'platform', then we refer to the traditional classified ads platform. While the spider is technically speaking also a platform, we refer to it as 'spider', never as 'platform'.

[12] We use m to denote broker because we want to avoid confusion with B = buyers. m comes from mediator.

increases with the number of properties announced on the website. If we assume that there is substantial competition in the market for brokers, then the number of houses sold on the website of a broker will be limited and therefore will attract only few visitors. We assume that buyers do not access the websites of brokers directly. A fraction λ_1 of the buyers is familiar with the platform, a fraction λ_2 is familiar with the spider. Therefore, buyers who are familiar with one website, will use it if the number of sellers is higher than their opportunity cost. The utility of a potential buyer (U^B) is equal to $U^B = S - k$. If they know both websites, then it depends on whether the platform allows spidering. If spidering is allowed, then these buyers always use the spider website, since this website also connects them to the content of the platform. Even if brokers advertise on the platform, then the spider does not link to the platform but it links directly to the website of the broker. If spidering is not allowed, then in the case buyers know both websites, they first look randomly at the spider or the platform. Second they'll also visit the other website given it contains enough unique information (i.e. advertisements that are not on the website they visited first). Buyers are heterogeneous in their opportunity cost k, which is distributed uniformly over the 0-1 space. The demand functions read:

$$\begin{aligned} B_s^p &= \lambda_2(S_s^p + \mu) + \lambda_1(1-\lambda_2)(S_s^p + S_s^m) \\ B_s^m &= \lambda_1(1-\lambda_2)(S_s^p + S_s^m) \\ B_n^p &= \lambda_1\lambda_2(S_n^p + \frac{1}{2}S_n^m) + \lambda_1(1-\lambda_2)(S_n^p + S_n^m) \\ B_n^m &= \lambda_1\lambda_2(S_n^p + \frac{1}{2}S_n^m) + \lambda_1(1-\lambda_2)(S_n^p + S_n^m) \end{aligned} \tag{4}$$

Note that these quantities are the number of buyers that actually reach the advertisements of private sellers and brokers; e.g. the buyers who access the spider see the private sellers on the platform but not the brokers.

3.3 Results

The simple model presented above has one solution for each regime:

$$\begin{aligned} p_s^* &= \frac{2\lambda_1(\lambda_2-1) - \lambda_2(2-3\mu+\mu^2) + \sqrt{\lambda_1^2(\lambda_2-1)^2 - \lambda_1(\lambda_2-1)\lambda_2(2-3\mu+\mu^2) + \lambda_2^2(\mu-1)^2(1-\mu+\mu^2)}}{3\lambda_1(\lambda_2-1) - 3\lambda_2(\mu-1)^2} \\ p_n^* &= \frac{1}{3} \end{aligned} \tag{5}$$

Within the parameter range λ_1,λ_2,$\mu \in [0,1]$, it always holds that $p_s^* \geq p_n^*$, though that thus not imply that profits are always higher in the spider case. This can be seen in figure 2. In the left panel, the profits are shown in function of μ (keeping λ_1 and λ_2 fixed), in the right panel, the profits are show in function of λ_1 (keeping μ and λ_2 fixed). This leads to the following proposition.

Proposition 3.1 *The platform allows spidering if the fraction of selling by brokers μ is sufficiently low and if the familiarity of the platform λ_1 is sufficiently low. The familiarity of the spider λ_2 has no influence on this decision.*

In his decision, the spider compares profits Π_s^* and Π_n^*. Whether $\Pi_s^* \geq \Pi_n^*$ depends on λ_1 and μ. For given parameter values, there exist critical values $\hat{\mu}$ and $\hat{\lambda_1}$ above which

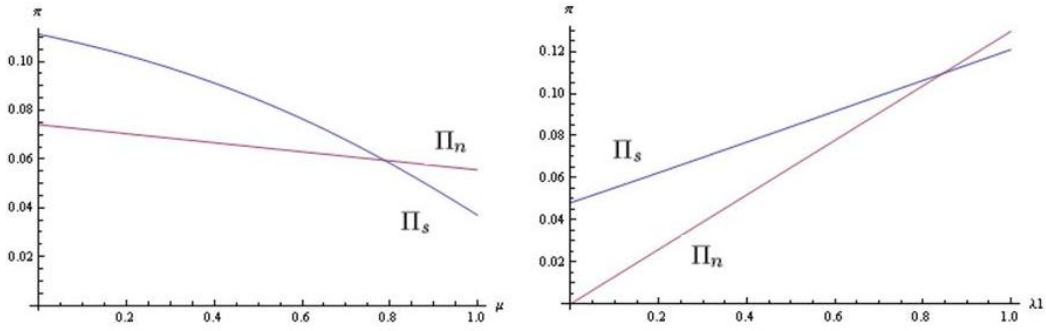

Figure 2: Profit in function of μ (left, $\lambda_1 = \lambda_2 = \frac{1}{2}$) and of λ_1 (right, $\mu = \lambda_2 = \frac{1}{2}$)

the platform forbids spidering. Related to this proposition, we can pose the following corollary.

Corollary 3.2 *The higher the familiarity of the platform λ_1, the lower the critical μ above which the platform prohibits spidering. Vice versa, the higher the fraction of selling by brokers μ, the lower the critical λ_1 above which the platform prohibits spidering. λ_2 has no influence on these critical values.*

In sum, the decision for a platform to allow a spider is inspired by the effect of a spider on the amount of buyers visiting the platform. On the one hand, the platform may gain buyers from being spidered for two reasons. First, some buyers with a high opportunity cost k would not have visited any of the two websites, but are willing to visit the spider because it contains most available houses. Second, those people who are familiar with the spider but not with the platform will also visit the platform indirectly since the spider refers to the platform. On the other hand, the platform also loses buyers since the spider does not refer to the broker's content on the platform, since it refers directly to the broker's website. Some of these buyers would have visited the content of the broker as well on the platform in the case where spidering is not allowed. As a consequence, the number of buyers that access the private sellers on the platform is always higher if spidering is allowed, the number of buyers that access the broker's content is always lower. The net effect determines whether the platform will allow spidering or not. On the seller side, it holds that $p_s^* \geq p_n^*$, therefore $S_s^p \leq S_n^p$ and $S_s^m \leq S_n^m$. Though, since sellers join on a pay per click basis, they are not driving the decision of a platform to allow the spider.

4 Extensions to the Model

In this section, we discuss some potential extensions of the model presented in the previous section. First, we propose to include side payments. Second and third, we endogenize the relevant parameters. We end with a discussion on the existence of many platforms (in stead of a monopolist platform).

Side Payments Hitherto, we did not include the possibility of side payments. Similar to the analysis of Katz & Shapiro (1985), if the total profits for all companies are larger

with spidering, then side payments might create a mutual beneficial situation.

The reason why side payments are not modelled is that we approached the spider as a mechanical player who did not take strategic decisions. One of the reasons why we abstain from making the spider strategic is that there exist many different mechanisms in the market to earn revenues. Moreover, most of them rely on factors that are not included in the model.

The most common model nowadays on the internet is an advertisement-supported model, where the website attracts readers (whatever the content may be) and advertisers pay to have their part of the attention pie (whether it is related to the content or not). If we would include an advertisement-supported model, then the spider would gain from having as much potential buyers as possible (which it cannot influence) and the magnitude of potential side payments would depend on the size of the advertisement market. Therefore it would not add much to the model, since the real issues stay exogenous.

Note also that it should not always be the spider who compensates the platform for potential losses. It can also be the other way around. If spidering is beneficial for the platform but is loss-making for the spider, then the platform can pay the spider to be in the market. This can be the case when there is a fixed cost that should be covered (with an advertisement market that is too small). Again total profits of spidering should be higher than under non-spidering, otherwise there is not enough to redistribute.

Endogenizing λ The parameters λ_1 and λ_2 capture the brand awareness of the websites. There are two interesting possibilities to endogenize this familiarity parameter: by making an investment in advertising and by making the familiarity dependent on the sellers or buyers in the previous periods.

If we add to our model strategic investment in brand familiarity, then we bump into the same problems as with the note on side payments. If the spider does not make profit, then it cannot make a strategic investment in advertisement. Of course, we can investigate the effect of an investment of the platform in λ_1 on his profit. Though another question might be more interesting. Suppose that the platform is on the market first and it cannot stop a spider from spidering. If the spider incurs a fixed cost, then it might be possible to deter the spider from the market. Similar to the analysis of Fudenberg & Tirole (1984), it is interesting to look whether the incumbent will over- or underinvest in order to deter; or whether he will accommodate entry. It might even be the case that the platform is willing to invest in the familiarity of the spider. However, to analyse this strategic decision, we should include a profit and cost function for the spider.

Another way to endogenize λ_1 and λ_2 is to make it dependent on the number of buyers and/or sellers in the previous period. Since these parameters have a positive effect on the profit, platforms will use the number of buyers and/or sellers also for strategic reasons. This is similar to a model where buyers and sellers decide sequential whether to join the platform, i.e. each side decides on the number of buyer or sellers in the previous period (see De Smet (2008)). A likely effect from this extension is that allowing a spider can be beneficial today, but in the long run it might reduce λ_1 or increase λ_2. The latter has no effect on the decision whether to spider or not, though a reduction of λ_1 makes the platform even more dependent on the spider, while reducing profits (see figure 2). Therefore, the platform may deny access to the spider, even though it is beneficial on the

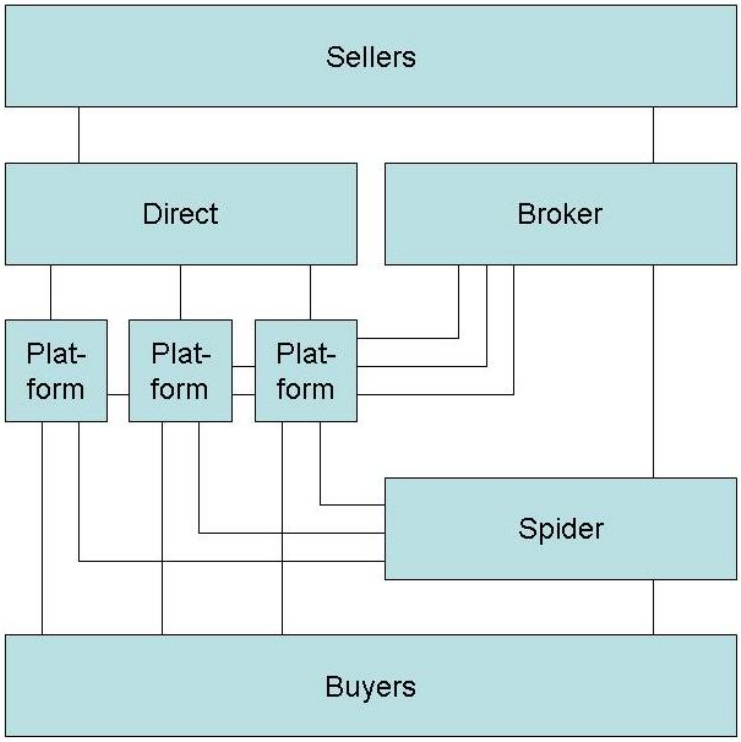

Figure 3: General Structure of Real Estate Market with n Platforms

short run, because of the adverse effects on the long run.

Endogenizing μ A similar reasoning holds for an endogenous μ. Hitherto we assumed that this fraction is determined by factors outside our model, though the rise of a spider might be beneficial for brokers and make it cheaper for them to sell houses. Given their better performance, more house sellers will be attracted to a broker, increasing the share of μ. As can be seen in figure 2, this reduces the profit of the platform and makes him less willing to allow the spider to copy content. Making the share of brokers endogenous, result in another concern for the traditional platform. He will be willing to give up short term profit for long term gains, by denying access to the spider.

n Platforms Up until now, we assumed that there is only one active platform in the market, though in many countries there are many platform operating. Suppose there is not one platform, but n platforms in the market, as shown in figure 3. Now each platform has to decide whether it allows the spider or not. If the results of section 3 can be extended to the n case, then we can conclude the following. The share of brokers μ will affect all the platforms similarly: the higher μ the less likely they allow the spider. Though the familiarity parameter λ_1 might be differentiated for the platforms. All things equal, the lower the parameter λ_1, the more likely a platform is to allow spiders. Therefore there might exist an equilibrium where small platform allow the spider while large platforms deny access. Besides, if many small platforms join, then the spider might be even more attractive for buyers as it becomes the largest platform which puts even more pressure

Table 1: The Belgian Online Market for Real Estate Classified Ads (Top 10)

	Platform	Ads 2009	Ads 2008	Users 2009	Start
1	Immoweb.be	113111	105624	118289	Dec-96
2	Zimmo.be	108762	na	1467*	Oct-06
3	Vlan.be	68524	80500	46127	Jun-96
4	Immo4free.com	34500	22400	369*	na
5	Logic-Immo.be	31548	na	7595*	Nov-04
6	Hebbes.be	31480	43777	25465	Mar-01
7	Kapaza.be	18350	22532	209085	Apr-03
8	Koopjeskrant.be	17030	9238	27641	Sep-96
9	2dehands.be	8806	10126	174538	Jun-00
10	eBay.be	8354	29714	354052	Nov-01

Source: Ads: respective websites, Users: CIM metriweb (or statbrain.com if indicated with *), Start: dns.be
Data retrieved on 04/07/2008 and 04/12/2009

on the larger platforms that deny access. In the end, a possible equilibrium is that all buyers turn first to the spider which redirects to many traditional platform. Similarly, many spiders might be active in the market, all offering a fairly complete overview of the houses.

5 Discussion of Case Studies

5.1 Belgian Case

Belgium is a heavily populated country (355 inhabitants per squared kilometre) with nearly 11 million inhabitants. Since building lots become more sparsely (also because of tighter regulations), there is a reasonably large supply and demand for houses. A large share of these houses is directly sold by the owner, without the intermediation of a broker (though the intermediation of a notary is obligatory in Belgium). Therefore, there is a need for platforms aggregating the supply of houses. This role was traditionally granted to (specialized) newspapers, but has been to a large extent taken over by the internet. Table 1 lists the ten most popular sites measured by the number of Belgian properties for sale[13].

Immoweb was one of the early movers in the market in 1998 and still has the largest share of properties for sale. Vlan and Hebbes are general classified ads sites, so these platforms are not exclusively reserved for real estate. Note that this biases the usage figures upwards, since we cannot split the visitors of real estate classified ads and other ads. Both platforms are owned by a large press group. Roularta, co-owner of Vlan, is specialized in magazines and regional advertising press; the other owner (Rossel) publishes newspapers in the French speaking market. Concentra, owner of Hebbes, publishes three national newspapers: Gazet van Antwerpen, Het Belang van Limburg and Metro (for free, also owned by Rossel). The other major press groups, de Persgroep and Corelio

[13]For detailed information on these data, see the Appendix.

stopped the activities of their platforms, Immonet respectively Spotter. Immo4free is privately owned. Logic-immo.be is a website that displays only houses sold by brokers; private sellers cannot advertise their houses on this platform. Kapaza, 2dehands and Koopjeskrant are specialized in classified ads, not exclusively limited to real estate. Their basic services are for free. eBay is active in Belgium since 2001 and is widely known as an auction platform. It started with classified ads in 2006, and offers this service for free. The introduction seemed relatively successful, though the number of ads dropped with 72% in the period 2008-2009.

Note that we haven't discussed number 2, Zimmo.be. This is a spider website who doesn't have own content but spiders it from the broker's websites and other platforms. Zimmo.be is developed by Webplications, a Belgian software company that also sells software to brokers. Since May 2009, the press group Corelio took a 30% participation in Webplications. Corelio, owner of the traditional platform Spotter, decided to stop the real estate activities of Spotter though it did not communicate about this stop.

Remarkably nearly some platforms allow Zimmo to spider their content, others don't. The largest one Immoweb refuses access, while smaller ones (Vlan, Hebbes, Kapaza) allow the spider. This is in line with the extensions discussed in section 4. In these extensions, we argued that platforms with a higher λ deny access to the spider, while platforms with a low visibility allow the spider. Moreover, the denial of Immoweb might also be inspired by dynamic effects, such as the lower visibility in later periods, or a higher share of sales by brokers. Given the objective function of Zimmo, it is not unlikely that they aim to promote brokers, since they also sell software for brokers. In their search results, Zimmo always shows the properties of brokers first. Houses sold by private sellers appear at the end, only accessible by pressing on a scroll bar several times.

5.2 Dutch Case

The Dutch real estate market has many similarities with the Belgian market, but it also has some important differences. Similar to Belgium, the Netherlands has a high density ($400/km^2$), but has more inhabitants (16.5 million). Though the largest difference is that in the Netherlands most houses are sold by the intermediation of a broker. It is even quite common to have two brokers: one for the seller and one for the buyer. Therefore, there is a smaller need for a traditional platform that allows private sellers to announce their property for sale. The main Dutch real estate websites are summarized in 2, for a discussion of the sources, see the Appendix.

The largest and most known website is Funda.nl. This website is owned by the Dutch Real Estate Brokers Society (Nederlandse Vereniging van Makelaars (NVM) which groups the largest share of brokers in the Netherlands. Initially, Funda nearly had a monopoly on the market and displayed only houses from their own society. Things changed when Jaap.nl and Zoekallehuizen.nl came in the market. Both platforms are/were associated with a bank (respectively DSB (bankrupt since October 2009) and Rabobank). These banks value potential buyers because it allows them to sell mortgages. Real estate and mortgage brokers are separated, but there is a clear connection. As Yost (2008) puts it, "mortgage brokers provide the means for consumers to shop for real estate, while brokers market the real estate products that underlie the demand for mortgage products". Given

Table 2: The Dutch Online Market for Real Estate Classified Ads (Top 10)

	Platform	**Type**	**Ads 2009**	**Users 2009**	**Start**
1	Jaap	S	340635	10410	Sep-99
2	Funda	S	274530	145458	Jun-00
3	Zuka	S	235852	5913	Dec-06
4	Miljoenhuizen	S	157124	3647	Jan-06
5	Huislijn	S	150000	8355	Jul-96
6	Dimo	S	104251	757	Aug-98
7	Marktplaats	P	102633	1752232	May-99
8	Huizenkrant	P	28261	330	Jul-99
9	Marktnet	P	25673	19247	Nov-99
10	VBO	P	23623	3541	Jun-98
	Zoek Alle Huizen	S	na	4771	Jul-05

Explanation: S stands for spider, P for platform.
Source: Ads & Type: respective websites, Users: statbrain.com, Start: sidn.nl
Data retrieved on 13/12/2009

the complementarities, a real estate spider might seem a valuable advertising tool for mortgages. Jaap.nl and Zoekallehuizen.nl operate as spiders, and they also automatically include the content of the NVM brokers. Funda.nl, the website of NVM saw this as a threat of their position on the market and took this spidering issue to the court. In August 2007, the court judged that the copying of Jaap.nl was a copyright infringement since they had no agreement with the owners of the copyright. Though, the court argued also that it is allowed to copy small chunks of texts (maximum 155 symbols) and copy pictures (maximum 194x145 pixels). These limited texts and pictures are considered as quote, therefore explicit agreement is not necessary. This citation right was earlier decided in a similar lawsuit in 2006, where Funda took the spider website Zoekallehuizen.nl (or ZAH.nl) to the court. Due to these judgments, Funda.nl decided to become a spider as well, and present also properties for sale from brokers which are no member of NVM. Since September 2007, Funda shows also advertisements of other brokers, though only at the end of the search results.

If we want to reconcile this with the insights from our theoretical model, then Funda.nl should be considered as a traditional platform, and the NVM brokers as private sellers. It is now the society NVM who decides whether other platforms might copy this information. They have considered the potential gains for NVM brokers from being spidered, but also the dynamic effect where the fraction μ (which can be translated as non-NVM brokers might increase due to the spidering). The difficulty for Funda is that they cannot forbid the spiders, though NVM brokers can take technical protection measures on their websites, eliminating the threat of being spidered. This was discussed in the case NVM vs Zoekallehuizen.nl. Zoekallehuizen.nl said to respect these protection measures. It might be the case that there is a free rider problem concerning these protections. Forbidding spidering is good for the NVM brokers as a whole, though being spidered while others are not might be beneficial for an individual broker.

6 Conclusion

In their book *Spidering Hacks*, Hemenway & Calishain (2003) present some tricks and applications for screenscrapers, though they first warn the readers. "Whatever your spider does, it needs to do it in the spirit of keeping the site from which it draws information healthy. if you write a spider that sucks away all information from advertising-supported sites, and they can't sell any more advertising, what happens? The site dies. You lose the site, and your program doesn't work anymore."

Given the many possibilities for websites to halt the activities of spiders, it is necessary that spidering is mutually beneficial. In a simple theoretical model, we show that a spider has two effects on traditional real estate marketing platforms. On the one hand, the platform attracts potential buyers that would otherwise not have visited the platform. On the other hand, the platform loses revenues from the brokers since the spider links directly to the broker's webpages.

It is more likely that the traditional platform allows spidering when the share of brokers is small, or when the visibility of the platform is low. We also argue that dynamic effects on the visibility or the broker's fraction may make the platform more cautious to be spidered.

Further research might apply our framework empirically to real estate markets or extend it to other online markets where spiders might be active. Our model might also prove helpful in current legal discussions, such as the Google books case.

Appendix

The Belgian Case

In this section, we discuss the relevant issues at stake related to the data collection used in section 5.1.

Platform For the choice of the platforms for online classified ads, we looked up the website with the largest content (immoweb.be) and ran a search on Google (related:immoweb.be). We selected the platforms from the list.

Ads Ads are the number of houses or apartments offered for sale or rent on the platform. We collected these data from the respective websites themselves, so there is a potential measurement problem if they don't report their ads in the same way. We only counted the houses offered in Belgium and we excluded holiday cottages. In some cases, such as immo4free.com it was not possible to disentangle the Belgian and the outside ads, therefore these figures might be an overestimation.

Users Users are the average number of unique visitors of the last week on the platform website, as measured by CIM Metriweb. The Center for Information and Media (CIM) is a non-profit organization which collects and verifies data from websites and media companies. It is widely trusted for its accurateness, but provides only data for companies who joined CIM. For those websites that are not member of CIM, we used the website Statbrain.com. Statbrain is less accurate than Metriweb, but it has the advantage that

it has results for nearly every website. Statbrain is based on the number of links a site receives and their Alexa ranking. Note here also that the number of visitors is for the whole website, so if the website contains other pages than real estate ads, these are also included. Similar to the CIM numbers, usage figures might be biased upwards for general platforms, since these websites are not exclusively reserved for real estate.

Start For start, we used the registration date of the website. These data are retrieved from dns.be, the non-profit organization responsible for managing the .be top level domain. We controlled this data with the data from the website archive.org. This website is run by the American non-profit organization Internet Archive and collects snapshots of websites. Though the first date of archiving might be different of the start date of the website, it gives an indication of the start period.

The Dutch Case

In this section, we discuss the relevant issues at stake related to the data collection used in section 5.2.

Platform For the choice of the platforms for online classified ads, we looked up the website with the largest familiarity (funda.nl) and ran a search on Google (related:funda.nl). We selected the platforms from the list.

Ads Ads are the number of houses or apartments offered for sale or rent on the platform. We collected these data from the respective websites themselves, so there is a potential measurement problem if they don't report their ads in the same way. We only counted the houses offered in the Netherlands and we excluded holiday cottages.

Users Users are the estimated average number of unique visitors of the last week on the platform website, as measured by Statbrain. This estimate is based on the number of links a site receives and their Alexa ranking. Note here also that the number of visitors is for the whole website, so if the website contains other pages than real estate ads, these are also included. Usage figures might be biased upwards for general platforms, since these websites are not exclusively reserved for real estate.

Start For start, we used the registration date of the website. These data are retrieved from sidn.nl, the non-profit organization responsible for managing the .nl top level domain. We controlled this data with the data from the website archive.org. This website is run by the American non-profit organization Internet Archive and collects snapshots of websites. Though the first date of archiving might be different of the start date of the website, it gives an indication of the start period.

References

Armstrong, M. (2006), 'Competition in two-sided markets', *RAND Journal of Economics* **37**(3), 668–91.

Bonanno, G. & Vickers, J. (1988), 'Vertical separation', *Journal of Industrial Economics* **36**(3), 257–65.

Caillaud, B. & Jullien, B. (2003), 'Chicken and egg: Competition among intermediation service providers', *RAND Journal of Economics* **34**(2), 309–28.

Coles, M. G. & Muthoo, A. (1998), 'Strategic bargaining and competitive bidding in a dynamic market equilibrium', *Review of Economic Studies* **65**(2), 235–60.

De Smet, D. (2008), Entry deterrence in two-sided markets, Master's thesis, Katholieke Universiteit Leuven.

Fischer, S. (2001), 'When aninals attack: Spiders and internet trespass', *Minnesota Intellectual Property Review* **2**(2), 139–182.

Fudenberg, D. & Tirole, J. (1984), 'The fat-cat effect, the puppy-dog ploy, and the lean and hungry look', *American Economic Review* **74**(2), 361–66.

Hemenway, K. & Calishain, T. (2003), *Spidering Hacks: 100 Industrial-Strength Tips & Tools*, 1 edn, O'Reilly. 432 pages.

Hendel, I., Nevo, A. & Ortalo-Magné, F. (2009), 'The relative performance of real estate marketing platforms: Mls versus fsbomadison.com', *American Economic Review* **99**(5), 1878–98.

Jennings, F. & Yates, J. (2009), 'Scrapping over data: are the data scrapers' days numbered?', *Journal of Intellectual Property Law & Practice* **4**, 120–129.

Katz, M. L. & Shapiro, C. (1985), 'Network externalities, competition, and compatibility', *American Economic Review* **75**(3), 424–40.

Peitz, M. & Waelbroeck, P. (2006), 'Why the music industry may gain from free downloading – the role of sampling', *International Journal of Industrial Organization* **24**(5), 907–913.

Rochet, J.-C. & Tirole, J. (2003), 'Platform competition in two-sided markets', *Journal of the European Economic Association* **1**(4), 990–1029.

Rosenfeld, J. (2002), 'Spiders and crawlers and bots, oh my: The economic efficiency and public policy of online contracts that restrict data collection', *Stanford Technology Law Review* **3**.

Short, J. (2004), 'An economic analysis of the law surrounding data aggregation in cyberspace', *Maine Law Review* **56**(1), 61–100.

Stahl, D. (1989), 'Oligopolistic pricing with sequential consumer search', *American Economic Review* **79**(4), 700–712.

Travis, H. (2006), 'Google book search and fair use: itunes for authors, or napster for books?', *University of Miami Law Review* **61**, 601–681.

Yost, J. (2008), *The Internet and American Business*, The MIT Press, chapter Internet Challenges for Nonmedia Industries, Firms and Workers, pp. 315–349.