Working Paper Series

Demand Externalities and Price Cap Regulation: Learning from the U.S. Debit Card Market

WP 13-06R

Zhu Wang
Federal Reserve Bank of Richmond

THE FEDERAL RESERVE BANK OF RICHMOND
RICHMOND ▪ BALTIMORE ▪ CHARLOTTE

Demand Externalities and Price Cap Regulation: Learning from the U.S. Debit Card Market

Zhu Wang*

July 2014
Working Paper No. 13-06R

Abstract

This paper studies unintended consequences of price cap regulation in the presence of demand externalities in the context of payment cards. The recent U.S. debit card regulation was intended to lower merchant card acceptance costs by capping the maximum interchange fee. However, small-ticket merchants found their fees instead higher after the regulation. To address this puzzle, I construct a two-sided market model and show that card demand externalities across merchant sectors rationalize card networks' pricing response. Based on the model, I study socially optimal card fees and an alternative cap regulation that may avoid the unintended consequence on small-ticket merchants.

Keywords: Price cap regulation; Demand externalities; Two-sided market
JEL Classification: D4; L5; G2

---

*Research Department, Federal Reserve Bank of Richmond. Email: zhu.wang@rich.frb.org. I thank Wilko Bolt, Huberto Ennis, Darren Filson, Boyan Jovanovic, Grace Bin Li, and participants at the Economics of Payments VI Conference hosted by the Bank of Canada, 2013 International Industrial Organization Conference and various seminars for helpful comments. The views expressed herein are solely those of the author and do not necessarily reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

1 Introduction

Credit and debit cards have become an important part of our payments system and they affect a large number of consumers and merchants. Recent Federal Reserve studies show that 80 percent of U.S. consumers have debit cards and 78 percent have credit cards. In a typical month, 31 percent of consumer payments are made with debit cards, and 21 percent with credit cards.[1]

However, the pricing in the payment card markets has been controversial. As Rochet and Tirole (2006) pointed out, payment cards are so-called "two-sided markets," in which card networks serve two distinct end-user groups, namely, cardholders and merchants.[2] In practice, card networks and their issuers typically charge high interchange fees to merchants for card acceptance, but provide rewards to consumers for card usage. Many industry observers and policymakers have become concerned that this highly skewed pricing structure may distort payments efficiency by inflating merchants' costs of accepting cards. Meanwhile, more than 20 countries have regulated or started investigating interchange fees.

In the U.S., the Durbin Amendment to the Dodd-Frank Act has recently required the Federal Reserve to regulate debit card interchange fees. Under the regulation, the maximum permissible debit interchange fee for covered issuers is capped at a half of its pre-regulation industry average level. As a direct impact, card issuers lost multibillion-dollar annual interchange revenues to merchants. However, the regulation has also generated unintended consequences on certain merchant groups. Particularly, prior to the regulation, merchants were charged differentiated interchange fees based on their sectors. Post regulation, however, card networks set a uniform interchange fee at the maximum cap amount. As a result, small-ticket merchants who used to pay lower interchange fees found their rates instead increased. In essence, the price cap has become a price floor.

The unintended consequence on small-ticket merchants made headlines and resulted in a lawsuit filed by merchant groups against the Federal Reserve's debit interchange

[1] Kevin Foster et al. (2010).

[2] The research on two-sided markets recently has gained wide attention (Rysman 2009). Other examples include HMOs (patients and doctors), operating systems (computer users and software developers), video game consoles (gamers and game developers), and newspapers (advertisers and readers).

regulation.[3] This presents a puzzle: Why would card networks raise fees on small-ticket merchants in response to a fee cap? If each merchant sector is independent in terms of card acceptance and usage and networks find that they maximize profits by charging lower fees to small-ticket merchants, it is not obvious why they would abandon this strategy in the face of a cap that is higher than the fees they were charging.

This puzzle is not readily explained by the existing two-sided payment card market models (e.g. Rochet and Tirole 2002, 2011, Wright 2003, 2011). Those studies find that privately determined interchange fees tend to exceed the socially efficient level because of the wrong incentives at the point of sale, i.e. consumers pay the same retail price regardless of the payment instrument they use. However, those models typically treat merchant sectors independent from one another in terms of card acceptance and usage, so they do not predict or explain why some merchants would be adversely affected by an interchange cap that is not binding for them.

In this paper, I address this puzzle by introducing card demand externalities into a two-sided market framework. In the model, merchant sectors are charged differentiated interchange fees due to their (observable) heterogenous benefits of card acceptance and usage. In addition, consumers' benefits of using cards in a merchant sector are positively affected by their card usage in other sectors, which I call "ubiquity externalities."[4] This type of demand externalities is shown to drive card networks' response to the cap regulation: Before the regulation, card networks were willing to offer subsidized interchange fees to small-ticket merchants because their card acceptance boosts consumers' card usage for large-ticket purchases from which card issuers can collect higher interchange fees. Once a cap on interchange fees was imposed, however, card issuers profit less from this kind of externalities so they discontinued the subsidy.

[3] E.g. see "Debit-Fee Cap Has Nasty Side Effect," *Wall Street Journal*, December 8, 2011.

[4] Ubiquity has always been a top selling point for brand cards. This is clearly shown in card networks' campaign slogans, such as Visa's "It is *everywhere* you want to be," and MasterCard's "There are some things money can't buy. For *everything* else, there's MasterCard." Ubiquity externalities may arise from various sources. First, in the presence of a fixed adoption cost, consumers are more likely to adopt payment cards if the card is accepted by more merchants. Second, for consumers who have adopted cards, universal card acceptance may allow them to carry less cash and as a result rely more on cards for making payments. Third, universal card usage may allow card networks and issuers to collect more complete information on consumer shopping patterns, so that they can design better services to encourage further card usage (e.g. by offering more targeted card reward programs). All these ubiquity externalities, regardless of their sources, are consistent with our following analysis.

Based on the model, I then study socially optimal card fees and alternative regulations. The analysis shows that the social optimum generally would require lower interchange fees than those chosen by the private market, but nevertheless it may maintain the differentiated fee structure by charging high (respectively, low) interchange rates to large-ticket (respectively, small-ticket) merchants. This is because both the social and the private optima seek to internalize the positive externalities of card usage across merchant sectors by subsidizing small-ticket transactions. In the presence of card demand externalities, I further show that capping the weighted average interchange fee, instead of the maximum interchange fee, may help restore the social optimum.

The contribution of the paper is threefold. First, I address a puzzle of the debit card interchange regulation by showing a "waterbed effect" may be at work, where regulating down the price of one sector may reduce the cross-subsidies that this sector provides to another one. This provides a rational explanation for the unintended consequences following the regulation. Second, I embed the analysis in an extended two-sided market model. In contrast to the existing payment card literature, the new model considers endogenous issuer competition, heterogenous merchant sectors, and card demand externalities. Exploring these features yields a better understanding of both the structure and the levels of socially optimal interchange fees. I show that the socially optimal fee structure may allow price discrimination, and the fee levels are determined by multiple factors, including merchant-and-consumer net benefits of card usage (which are subject to ubiquity externalities), the competitiveness of issuers, and the acquirers' cost. These new results suggest that the popular interchange regulations adopted in various countries, solely based on either issuer costs or merchant benefits, may have inadequate theoretical foundation.[5] Finally, I propose an alternative regulation that caps the weighted average interchange fee. The alternative regulation is shown to provide incentives for card net-

[5] Two types of interchange fee regulations are currently in practice. One is based on issuers' costs, first adopted by the Reserve Bank of Australia in early 2000. The Durbin regulation in the U.S. is a recent example. The issuer-cost based regulation has been criticized for ignoring the two-sided nature of payment card markets. Instead, Rochet and Tirole (2011) proposed regulating the interchange fee based on merchant transaction benefit of card acceptance, which was adopted by the European Commission. The merchant-benefit based regulation addresses the two-sided market concerns, but relies on a strong assumption that issuers set a constant markup. Moreover, neither type of the regulations has considered card demand externalities across merchant sectors.

works and issuers to internalize card demand externalities and hence avoid unintended consequences on small-ticket merchants.

The paper is organized as follows. Section 2 provides the background of the payment card industry and the debit interchange fee regulation. Section 3 lays out a two-sided payment card market model with heterogenous merchant sectors and differentiated interchange fees. The model allows for card demand externalities across merchant sectors. Section 4 characterizes the model equilibria with and without the interchange cap regulation. Section 5 discusses socially optimal interchange fees and an alternative cap regulation. Section 6 provides concluding remarks.

2 Industry background

Credit and debit cards have become an increasingly important part of the U.S. payments system. Recent data show that the share of their transactions in personal consumption expenditures rose to 48 percent in 2011. Among those, credit cards were used in 26 billion transactions for a total value of $2.1 trillion, and debit cards were used in 49 billion transactions for a value of $1.8 trillion.[6]

Credit cards typically provide float or credit to cardholders, while debit cards directly draw from the cardholder's bank account right after each transaction. In practice, debit card payments are authorized either by the cardholder's signature or with a PIN number. The former accounts for 60 percent of debit transactions and the latter accounts for 40 percent.

Visa and MasterCard are the two major card networks in the United States. They provide card services through member financial institutions (issuers and acquirers) and account for 85 percent of the U.S. consumer credit card market.[7] Visa and MasterCard are also the primary providers of debit card services. The two networks split the signature debit market, with Visa holding 75 percent of the market share and MasterCard holding

[6]Source: *Nilson Report*, December 2011. Prepaid cards are another type of general-purpose cards but with much smaller volumes. They accounted for 2% of U.S. personal consumption expenditures in 2011.

[7]American Express and Discover are the other two credit card networks holding the remaining market share. They handle most card issuing and merchant acquiring by themselves and are called "three-party" systems. For a "three-party" system, interchange fees are internal transfers.

25 percent.[8] In contrast, PIN debit transactions are routed over PIN debit networks. Interlink, Star, Pulse and NYCE are the top four networks, together holding 90 percent of the PIN debit market share. The largest PIN network, Interlink, is operated by Visa.

2.1 Interchange controversy

Along with the development of payment card markets, there has been a long-running controversy about interchange fees. Merchants are critical of the fees that they pay to accept cards. These fees are referred to as the "merchant discounts," which are composed mainly of interchange fees paid to card issuers (i.e., banks issuing cards and make payments on behalf of cardholders) through merchant acquirers (i.e., banks collecting payments on behalf of merchants). Merchants believe that the card networks and issuers have wielded their market power to set excessively high interchange fees. The card networks and issuers counter that these interchange fees are necessary for covering issuers' costs as well as providing rewards to cardholders, which may also benefit merchants by making consumers more willing to use the cards.

In recent years, merchant groups launched a series of litigation against what they claim is anticompetitive behavior by the card networks and their issuers. Some of the lawsuits have been aimed directly at interchange fees of credit and debit cards. For example, a group of class-action suits filed by merchants against Visa and MasterCard alleged that the networks violated antitrust laws by engaging in price-fixing. As a result, Visa, MasterCard and their major issuers reached a $5.7 billion settlement agreement with U.S. retailers in December 2013, which is the largest antitrust settlement in U.S. history.

The heated debate on interchange fees has also attracted attention from researchers and regulatory authorities. On the research side, a sizeable body of literature, called "two-sided market theory," has been developed to evaluate payment card market competition and pricing issues.[9] On the regulatory side, three bills restricting interchange fees were

[8] Discover has recently entered the signature debit market, but its market share is small.

[9] For example: Baxter (1983), Carlton and Frankel (1995), Katz (2001), Schmalensee (2002), Rochet and Tirole (2002, 2006, 2011), Gans and King (2003), Wright (2003, 2004, 2010, 2012), Cabral (2005), Armstrong (2006), Schwartz and Vincent (2006), Rysman (2007, 2009), Bolt and Chakravorti (2008), Robin Prager et al. (2009), Rochet and Wright (2010), Wang (2010), Weyl (2010), Shy and Wang (2011), McAndrews and Wang (2012), and Bedre-Defolie and Calvano (2013).

introduced in Congress shortly before the Durbin Amendment was passed.[10] Similar trends are also taking place in many other countries.[11]

2.2 Durbin regulation

In 2010, an amendment sponsored by Sen. Dick Durbin was added to the Dodd-Frank bill, which was passed and signed into law in July 2010. The Durbin Amendment directs the Federal Reserve Board of Governors to ensure that debit card interchange fees are "reasonable and proportional to the cost incurred by the issuer with respect to the transaction." The Federal Reserve Board thereafter issued Regulation II (Debit Card Interchange Fees and Routing), which went into effect on October 1, 2011.

The new regulation establishes a cap on the debit interchange fees that banks with more than $10 billion in assets can collect from merchants through merchant acquirers. The permissible fees were set based on the Fed's evaluation of issuers' costs associated with debit card processing, clearance and settlement. The resulting interchange cap is composed of the following: A base fee of 21 cents per transaction to cover the issuer's processing costs, a five basis point adjustment to cover potential fraud losses, and an additional 1 cent per transaction to cover fraud prevention costs if the issuer is eligible. This cap applies to both Signature and PIN debit transactions.

The regulation has a major impact on card issuers' interchange revenues. According to a recent Federal Reserve study, the average debit card transaction in 2009 was approximately $40. Based on the regulation, the interchange fee applicable to a typical debit card transaction would be capped at 24 cents (21 cents + ($40 × .05%) + 1 cent), which is about half of its pre-regulation industry average level. As a result, card issuers were expected to lose an estimated $8.5 billion annual interchange revenues.[12]

In response to the reduced interchange revenues, many card issuing banks have cut back their debit reward programs and free checking services. A recent Pulse debit issuer

[10] The three bills are a House version of the Credit Card Fair Fee Act of 2009, a Senate version of the same act, and the Credit Card Interchange Fees Act of 2009.

[11] Recent examples of interchange fee regulation include Argentina, Australia, Austria, Brazil, Canada, Chile, China, Colombia, Denmark, European Union, France, Hungary, Israel, Mexico, New Zealand, Norway, Panama, Poland, Portugal, South Africa, South Korea, Spain, Switzerland, Turkey, and United Kingdom.

[12] Wang (2012) provides some estimates of issuers' lost interchange revenues using Call Report data.

study shows that 50 percent of regulated debit card issuers with a reward program ended their programs in 2011, and another 18 percent planned to do so in 2012.[13] Meanwhile, the Bankrate's 2012 Checking Survey shows that the average monthly fee of noninterest checking accounts rose by 25 percent compared with the year before, and the minimum balance for free-checking services rose by 23 percent.[14] Several major banks including Bank of America, Wells Fargo, and Chase attempted to charge a monthly debit card fee to their customers, though they eventually backed out due to customer outrage.[15]

2.3 Small-ticket effect

Merchants as a whole may have greatly benefited from the reduced debit interchange fees under the regulation.[16] However, the distribution of the benefits is quite uneven. Particularly, an unintended consequence quickly surfaced: Small-ticket merchants find their interchange fees higher after the regulation.

Prior to the regulation, Visa, MasterCard, and most PIN networks offered discounted debit interchange fees for small-ticket transactions as a way to encourage card acceptance by merchants specializing in those transactions.[17] For instance, Visa and MasterCard set the small-ticket signature debit interchange rates at 1.55 percent of the transaction value plus 4 cents for sales of $15 and below. As a result, a debit card would only charge a 7 cents interchange fee for a $2 sale or 11 cents for a $5 sale. However, in response to the regulation, most card networks eliminated the small-ticket discounts, and all transactions (except those on cards issued by exempt issuers) have to pay the maximum cap rate set by the Durbin regulation.[18] For merchants selling small-ticket items, this means that the

[13] The 2012 Debit Issuer Study, commissioned by Pulse, is based on research with 57 banks and credit unions that collectively represent approximately 87 million debit cards and 47,000 ATMs.

[14] Bankrate surveyed banks in the top 25 U.S. cities to compare the average fees associated with checking accounts in their annual Checking Account Survey.

[15] See "Banks Adding Debit Card Fees," *The New York Times*, September 29, 2011.

[16] Depending on merchant comptition, some of the benefits may be passed along to consumers through lower retail prices.

[17] Visa and MasterCard introduced small-ticket discounted interchange fees in the early 2000s. The rates were applied to merchant sectors specializing in small-ticket transactions, including Local Commuter Transport, Taxicabs and Limousine, Fast Food Restaurants, Coffee Shops, Parking Lots and Garages, Motion Picture Theaters, Video Rental Stores, Cashless Vending Machines and Kiosks, Bus Lines, Tolls and Bridge Fees, News Dealers, Laundries, Dry Cleaners, Quick Copy, Car Wash and Service Stations, etc. In October 2010, Visa expanded the program to include more merchant sectors.

[18] Hayashi (2013) compares the increases of interchange fees for small-ticket transactions for Visa,

cost of accepting the same debit card doubled or even tripled after the regulation.

The increase of small-ticket interchange fees could affect a large number of transactions. According to the 2010 Federal Reserve Payments Study, debit cards were used for 4.9 billion transactions below $5, and 10.8 billion transactions between $5-$15 in 2009. The former accounts for 8.3 percent of all payment card transactions (including credit, debit, and prepaid cards), and the latter accounts for 18.3 percent. Depending on their compositions of transaction sizes, merchants in different sectors could be affected differently by the post-regulation debit interchange fees.[19] However, merchants who specialize in small-ticket transactions would be most adversely affected.

In response, many small-ticket merchants have tried to find ways to offset their higher interchange rates. Some raised prices, or chose to restrict or reject the use of debit cards.[20] Some others offer customers incentives to consolidate transactions using prepaid cards or online wallets.[21] In the meantime, a lawsuit was filed in November 2011 by a group of trade associations and retail companies against the Federal Reserve's debit interchange regulation. The lawsuit alleges that the Fed has set the interchange cap too high by including costs that were barred by the law, and "forcing small businesses to pay three times as much to the big banks on small purchases was clearly not the intent of the law and is further evidence that the Fed got it wrong."[22]

The unintended regulatory impact on small-ticket merchants calls for a further examination of the payment card market. According to the Federal Reserve's evaluation, debit card issuers incur a per-transaction cost around 21 cents, which exceeds the interchange

MasterCard, and most PIN debit networks.

[19]E.g. Shy (2012) used the data from a diary study of consumer payment choices to identify the types of merchants who are likely to pay higher or lower interchange fees under the debit regulation.

[20]Notable examples in the press include: the DVD-rental company Redbox raised rental prices from $1 to $1.20 to cover increased debit fees; USA Technologies and Apriva, two large payment facilitators in the vending industry, stopped accepting MasterCard debit cards; the fast food restaurant chain Dairy Queen asked customers to pay with cash for purchases under $10. See "Debit-Fee Cap Has Nasty Side Effect," *Wall Street Journal*, December 8, 2011.

[21]Merchants are charged one transaction fee when a customer loads the prepaid card or online wallet rather than multiple times each instance a user pays with a debit card. Notable examples in the press include coffeehouse chain Starbucks promoting in-store prepaid cards and Washington, D.C. parking operator Parkmobile offering discounts for customers who pay with an online wallet. E.g. see "Small-Ticket Retailers Squeezed By High Transaction Fees," *U.S.News & World Report*, October 26, 2012.

[22]See "Merchants' Lawsuit Says Fed Failed to Follow Law on Swipe Fee Reform," *Business Wire*, November 22, 2011.

fees that they charged for small-ticket transactions prior to the regulation. Considering that issuers typically do not recover costs from the cardholder side (cardholders often receive a reward rather than pay a fee for each card transaction), card issuers appeared to have subsidized small-ticket transactions. This is also the reason that card networks claim why they eliminate the small-ticket discounts under the regulation.[23] While the existing two-sided market theories have shed great light on the functioning of interchange fees, they do not explain the subsidies on small-ticket transactions prior to the Durbin regulation, nor do they explain why these subsidies were discontinued afterwards. In this paper, I try to address this puzzle and draw some new implications.

3 Model environment

I consider a payment card system composed of five types of players: consumers, merchants, acquirers, issuers, and the card network, as illustrated in Figure 1. The setup extends the standard two-sided market model, such as Rochet and Tirole (2002, 2011), to allow for card demand externalities across merchant sectors.

■ **Consumers:** There is a continuum of measure one of consumers, who purchase goods from two distinct merchant sectors h and l. In this setting, h (respectively, l) refers to the large-ticket (respectively, small-ticket) sector where merchants and consumers enjoy *high* (respectively, *low*) transaction benefits of card acceptance and usage.[24]

Consumers have inelastic demand and buy one good per sector. Within each sector, consumers need to decide which store to patronize. They know the stores' price and card acceptance policy before making the choice. Once in a store they then select a payment method (a card or an alternative payment method such as cash), provided that the retailer indeed offers a choice among payment means. I assume price coherence such that retailers charge the same price for purchases made by card and by cash.[25] Whenever a transaction

[23] According to MasterCard, "the company decided that it couldn't sustain the [small-ticket] discounts under the new rate model because the old rates had essentially subsidized the small-ticket discounts." See "Debit-Fee Cap Has Nasty Side Effect," *Wall Street Journal*, December 8, 2011.

[24] For both merchants and consumers, replacing cash with cards may reduce their handling, safekeeping and fraud expenses on payments, and the benefits typically increase with transaction values. Therefore, it is natural to assume that merchants and consumers benefit more from card usage in large-ticket transactions than in small-ticket transactions.

[25] Price coherence is the key feature that defines a two-sided market. Rochet and Tirole (2006) show

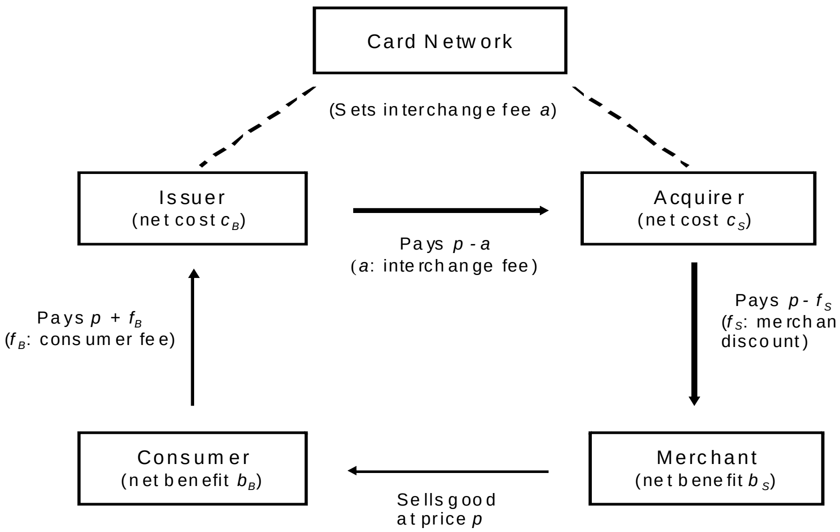

Figure 1: A Payment Card System

between a consumer (buyer) and a retailer (seller) is settled by card, the buyer pays a fee f_B^i to her card issuing bank (issuer) and the seller pays a merchant discount f_S^i to her merchant acquiring bank (acquirer). These fees, f_B^i and f_S^i, depend on the merchant sector $i \in \{h, l\}$, and f_B^i is allowed to be negative, in which case the cardholder receives a reward. There are no annual fees and all consumers have a card.[26]

A consumer's transaction benefit of purchasing good i with a card instead of cash is a random variable b_B^i drawn from a cumulative distribution function H_i on the support $[\underline{b}_B^i, \overline{b}_B^i]$. It is natural to think that the mean and variance of consumer card usage benefits positively relate to the transaction value. Denoting μ^i as the mean of b_B^i, this implies that $\mu^h > \mu^l$ and $var(b_B^h) > var(b_B^l)$. For simplicity, I thereafter assume H_h to be a uniform distribution on the support $[\mu^h(y^l) - \gamma, \mu^h(y^l) + \gamma]$, while H_l is a degenerate distribution

that the two-sided market pricing structure (e.g. interchange fees) would become irrelevant without the price coherence condition. In reality, price coherence may result either from network rules or state regulation, or from high transaction costs for merchants to price discriminate based on payment means. In the U.S., while merchants are allowed to offer their customers discounts for paying with cash or checks, few merchants choose to do so. On the other hand, card network rules and some state laws explicitly prohibit surcharging on payment cards.

[26]This model assumes a representative consumer framework developed by Wright (2004) and used in the subsequent literature. Alternatively, the model could use the framework developed by Rochet and Tirole (2002) and assume heterogenous consumers who differ systematically in their transaction benefits of using cards. As Rochet and Tirole (2011) show, these two alternative frameworks deliver convergent results, so the analysis and findings can be interpreted using either framework.

taking a single value μ^l. The latter is an innocuous assumption given that the variance of H_l is sufficiently small.[27] Moreover, I assume that μ^h is positively affected by the consumer's card usage in the small-ticket sector y^l, i.e. $d\mu^h/dy^l > 0$. (I define $y^l = q^l\chi_l$, where χ_l indicates whether l-sector merchants accept cards and q^l is a consumer's frequency of card usage in the l sector conditional on cards being accepted).[28] This assumption captures the idea that ubiquity externalities shift up consumers' valuation of paying with cards in the h sector.[29]

Cardholders are assumed to observe the realization of b^i_B once in the store. This is a standard assumption introduced by Wright (2004) and used in the subsequent literature. Because the net benefit of paying by card is equal to the difference $b^i_B - f^i_B$, a card payment is optimal for the consumer whenever $f^i_B \leq b^i_B$. Hence, whenever $f^h_B \leq \mu^h(y^l) + \gamma$, the proportion of card payments at an h-sector (i.e. large-ticket) store that accepts cards is

$$q^h(f^h_B) = \Pr(b^h_B \geq f^h_B) = \frac{\mu^h(y^l) + \gamma - f^h_B}{2\gamma}, \tag{1}$$

and the average net consumer benefit of paying with a card is

$$v^h(f^h_B) = E[b^h_B - f^h_B | b^h_B \geq f^h_B] = \frac{\mu^h(y^l) + \gamma - f^h_B}{2}. \tag{2}$$

Note that $q^h(f^h_B) = v^h(f^h_B) = 0$ if $f^h_B > \mu^h(y^l) + \gamma$.

Similarly, whenever $f^l_B \leq \mu^l$, the proportion of card payments at an l-sector (i.e. small-ticket) store that accepts cards is

$$q^l(f^l_B) = \Pr(\mu^l \geq f^l_B) = 1, \tag{3}$$

[27] Intuitively, we can think that a consumer's transaction benefit of paying by card relative to using cash is a random variable $b^i_B = bp^i$, where b is a random factor and p^i is the price of good i largely determined by the non-payment cost of the good. This implies that $E(b^h_B) > E(b^l_B)$ and $var(b^h_B) > var(b^l_B)$. Moreover, given that p^l is small, both $E(b^l_B)$ and $var(b^l_B)$ could be close to zero.

[28] Under the assumption that H_l is degenerate, consumer card usage in the l sector becomes a simple binary outcome, i.e. $q^l \in \{0, 1\}$. This makes it easier to model card usage externalities between the l and h sectors. Note that if H_l is a non-degenerate distribution, we then need to specify how card demand externalities vary by each of the multiple levels of card usage in the l sector, which significantly complicates the problem but does not provide greater intuition.

[29] For ease of exposition, I assume that consumers' transaction benefit of using cards in the l sector is fixed, unaffected by card usage in the h sector. However, relaxing this assumption would not change the qualitative findings.

and the average net consumer benefit of paying with a card is

$$v^l(f_B^l) = \mu^l - f_B^l. \tag{4}$$

Note that $q^l(f_B^l) = v^l(f_B^l) = 0$ if $f_B^l > \mu^l$.

■ **Merchants:** Merchants belong to one of the two sectors, h and l. A merchant in a given sector $i \in \{h, l\}$ derives the transaction benefit b_S^i of accepting payment cards (relative to handling cash), and naturally $b_S^h > b_S^l$. Moreover, the heterogeneity between sectors is observable to the card network so that the card network can perfectly price discriminate by charging differentiated interchange fee a^i to the merchant sector i.

By accepting cards, under the price coherence assumption, a merchant is able to offer each of its card-holding customers an additional expected surplus of $q^i(f_B^i)v^i(f_B^i)$, but faces an additional expected net cost of $q^i(f_B^i)(f_S^i - b_S^i)$ per cardholder from doing so. Here, f_S^i is the sector-specific merchant discount paid to the acquirer. Therefore, a merchant accepts cards if and only if $f_S^i \leq b_S^i + v^i(f_B^i)$. Rochet and Tirole (2011) show this condition holds for a variety of merchant competition setups, including monopoly, perfect competition and Hotelling-Lerner-Salop differentiated products competition with any number of retailers. Wright (2010) shows the same condition holds for Cournot competition.

I denote χ_i as an indicator function whether merchants in sector i accept cards or not. Accordingly,

$$\chi_i = \begin{cases} 1 & \text{if } f_S^i \leq b_S^i + v^i(f_B^i) \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

Note that merchants in the h and l sectors do not directly coordinate to internalize card usage externalities. This is a realistic assumption given that there could be a large number of merchants in each sector, which makes the coordination too costly. Moreover, due to antitrust restrictions, merchants in reality can not engage in group bargaining regarding interchange fees, so they typically face "take-it-or-leave-it" offers from card networks.

■ **Acquirers:** I assume acquirers incur a per-transaction cost c_S and are perfectly competitive. Thus, given the interchange fee a^i, they charge a sector-specific merchant discount f_S^i such that

$$f_S^i = a^i + c_S. \tag{6}$$

Because acquirers are competitive, they play no role in the analysis except passing through the interchange charge to merchants.

■ **Issuers:** There are $n \geq 1$ issuers who have market power.[30] Issuers incur a per-transaction cost c_B and receive an interchange payment of a^i in a card transaction. I consider a symmetric equilibrium at which all issuers charge the same consumer fee f_B^i, which can be negative if cardholders receive a reward.

As pointed out in Rochet and Tirole (2002, 2011), there are various ways to model issuer competition. To be concrete, I assume an explicit setting: Issuers coordinate on their pricing in the l sector where they make a loss (so that they internalize card demand externalities between the l and h sectors), but engage in a Cournot competition in the h sector where they make a profit. The former assumption simplifies the setting of small-ticket card fees in order to focus on the card demand externalities, while the latter assumption allows for endogenizing issuers' markup for large-ticket transactions. Note when $n = 1$, the model reduces to a special case where there is a monopoly issuer.[31]

For small-ticket transactions, issuers take the interchange fee a^l as given and set the consumer fee f_B^l to maximize their total profit conditional on merchants accepting cards (i.e. $\chi_l = 1$):

$$\hat{\Pi}^l = \max_{f_B^l} (f_B^l + a^l - c_B) q^l, \tag{7}$$

$$s.t.\ q^l = \begin{cases} 1 & \text{if } f_B^l \leq \mu^l \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

[30] This is a standard assumption in the literature. As pointed out in Rochet and Tirole (2002), the issuer market power may be due to marketing strategies, search costs, issuer reputation or the nature of the card. Note that were the issuing side perfectly competitive, issuers and card networks would have no preference over the interchange fee, and so the latter would be indeterminate.

[31] The assumption that payment card issuers engage in Cournot competition is consistent with Rochet and Tirole (2002). Alternatively, we could assume n symmetric monopoly issuers, each making their own pricing decisions in h and l sectors to internalize card demand externalities. The assumption of monopoly issuers is likely to be true for the case of debit cards because a debit card holder typically has long-term banking relationship with her card issuer. Our following analysis can equally apply to this alternative setup (by simply setting $n = 1$), in which case we show the welfare-maximizing interchange fees coincide with the market-determined ones, but the user-surplus-maximizing interchange fees are lower.

where (8) follows (3). Whenever $q^l = 1$, the highest possible consumer fee that issuers choose is

$$f_B^l = \mu^l, \tag{9}$$

and the corresponding total issuers' profit in the l sector is

$$\hat{\Pi}^l = \mu^l + a^l - c_B. \tag{10}$$

For large-ticket transactions, issuers engage in a Cournot competition if merchants accept cards (i.e. $\chi_h = 1$). Each issuer j sets the output level q_j^h taking the output by competing issuers, $q_{-j}^h = q^h - q_j^h$, as given and maximizes profit:

$$\hat{\pi}_j^h = \max_{q_j^h} (f_B^h + a^h - c_B) q_j^h, \tag{11}$$

$$s.t.\ f_B^h = \mu^h(y^l) + \gamma - 2\gamma(q_j^h + q_{-j}^h) \tag{12}$$

where (12) follows Eq (1). In a symmetric equilibrium, the total card usage q^h and the consumer fee f_B^h are pinned down as follows:

$$q^h = nq_j^h = \frac{n}{2\gamma(n+1)}[\mu^h(y^l) + \gamma + a^h - c_B], \tag{13}$$

$$f_B^h = \frac{1}{n+1}[\mu^h(y^l) + \gamma + n(c_B - a^h)], \tag{14}$$

and the total issuers' profit in the h sector is

$$\hat{\Pi}^h = \frac{n[\mu^h(y^l) + \gamma + a^h - c_B]^2}{2\gamma(n+1)^2}. \tag{15}$$

■ **Network:** I consider a monopoly network that sets sector-specific interchange fees a^h and a^l to maximize the total issuers' profit, namely

$$\Pi = \max_{a^h,\, a^l} (\hat{\Pi}^l \chi_l + \hat{\Pi}^h \chi_h), \tag{16}$$

where $\hat{\Pi}^l$ and $\hat{\Pi}^h$ are given by Eqs (10) and (15) above.

Because the network maximizes the issuers' profit, it makes a decision consistent with issuers on whether to provide card services to the l sector. Therefore, $y^l = \chi_l$ always holds at equilibrium, so we can simply replace $\mu^h(y^l)$ with $\mu^h(\chi_l)$ in the following analysis.

In the welfare and policy analysis (Section 5), I will also consider an alternative regime where the network is run by a social planner who maximizes social welfare or total user surplus.

■ **Timing**: I solve for a subgame perfect Nash equilibrium of the model. The timing of the game can be summarized in the following four stages.

1. The card network sets sector-specific interchange fees a^i.
2. Issuers and acquirers set fees f^i_B and f^i_S.
3. Depending on their value of b^i_S, merchants decide whether to accept cards and set retail prices.
4. Observing which merchants accept cards and their prices, consumers decide which merchants to purchase from. Once in the store, consumers receive their draw of b^i_B and decide how to pay.

4 Model characterization

I first consider a monopoly network, which sets sector-specific interchange fees a^i to maximize the total issuers' profit. In the absence of regulation, the network solves the following problem:

$$\Pi = \max_{a^h,\, a^l} \frac{n[\mu^h(\chi_l) + \gamma + a^h - c_B]^2}{2\gamma(n+1)^2}\chi_h + (\mu^l + a^l - c_B)\chi_l \tag{17}$$

$$s.t. \quad \chi_h = \begin{cases} 1 & \text{if } a^h \leq \frac{n[\mu^h(\chi_l)+\gamma+b^h_S - c_S - c_B]}{n+2} + b^h_S - c_S \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

$$\chi_l = \begin{cases} 1 & \text{if } a^l \leq b^l_S - c_S \\ 0 & \text{otherwise} \end{cases}. \tag{19}$$

The condition (18) is derived from (2), (5), (6) and (14), while the condition (19) is derived from (4), (5), (6) and (9).

Once an interchange fee cap $\overline{a}$ is introduced by regulation, the network then solves a similar problem as above but with an additional constraint:

$$a^i \leq \overline{a} \quad \text{for} \quad i \in \{h, l\}. \tag{20}$$

To help characterize the model equilibrium, I make three basic assumptions on parameter values.

Assumption A1.

$$Z^h(\chi_l) = b_S^h + \mu^h(\chi_l) + \gamma - c_B - c_S > 0 \quad \text{for} \quad \chi_l \in \{0, 1\}.$$

The first assumption states that the maximum merchant-and-consumer joint transaction benefit of using cards in the h sector net of costs is always positive. As will be shown, this ensures that issuers earn a positive profit for serving card transactions in the h sector.

Assumption A2.

$$Z^l = b_S^l + \mu^l - c_B - c_S < 0.$$

The second assumption states that the merchant-and-consumer joint transaction benefit of using cards in the l sector net of costs is negative. As will be shown, this implies that card issuers make a loss for serving card transactions in the l sector *per se*.

Assumption A3.

$$\mu^h(1) - \mu^h(0) = Z^h(1) - Z^h(0) > \frac{\gamma(n+2)^2(-Z^l)}{2n[Z^h(0) + Z^h(1)]}.$$

The third assumption states that card demand externalities are sufficiently large between the l and h sectors. As will be shown, this ensures that in the absence of regulation, the card network would charge differentiated interchange fees to serve card transactions in both the h and l sectors.

Under the above assumptions, I first characterize the model equilibrium in the absence of regulation. The findings are shown by the following proposition.

Proposition 1 *Under Assumptions A1-A3, an unregulated card network which maximizes total issuers' profit sets differentiated interchange fees such that cards are used in both the h and l sectors.*

Proof. Consider three options for the card network. First, when only the h sector is served with card services (i.e. $\chi_h = 1, \chi_l = 0$), the card network maximizes the total issuers' profit (17) by setting the h-sector interchange fee such that the constraint (18) is binding

$$a^h(\chi_l = 0) = b_S^h - c_S + \frac{n}{n+2} Z^h(0). \tag{21}$$

As a result, the total number of card transactions is

$$q^h = \frac{n}{\gamma(n+2)} Z^h(0), \tag{22}$$

and the total issuers' profit is

$$\Pi^h = \frac{2n}{\gamma(n+2)^2} [Z^h(0)]^2. \tag{23}$$

Under Assumption A1, this implies that $q^h > 0$ and $\Pi^h > 0$.

Second, when only the l sector is served with the card services (i.e. $\chi_h = 0, \chi_l = 1$), the card network maximizes the issuers' profit (17) by setting the l-sector interchange fee

$$a^l = b_S^l - c_S. \tag{24}$$

Under Assumption A2, the total issuers' profit is

$$\Pi^l = Z^l = b_S^l + \mu^l - c_S - c_B < 0. \tag{25}$$

Finally, when both the h and l sectors are served with card services (i.e. $\chi_h = \chi_l = 1$), the card network maximizes the issuers' profit (17) by charging differentiated interchange fees to the two sectors:

$$a^h(\chi_l = 1) = b_S^h - c_S + \frac{n}{n+2} Z^h(1), \tag{26}$$

$$a^l = b_S^l - c_S. \tag{27}$$

The resulting total issuers' profit is

$$\Pi^{h+l} = \frac{2n}{\gamma(n+2)^2}[Z^h(1)]^2 + Z^l. \tag{28}$$

Comparing Eqs (21), (26) and (27), it is found that the interchange fee is always higher in the h sector than the l sector, i.e.

$$a^h(\chi_l = 1) > a^h(\chi_l = 0) > a^l \tag{29}$$

given that $b_S^h > b_S^l$ and $Z^h(1) > Z^h(0) > 0$. Comparing (23) and (28), it is also verified that $\Pi^{h+l} > \Pi^h$ iff Assumption A3 holds. Therefore, under Assumptions A1-A3, the card network charges differentiated interchange fees given by (26) and (27) and serves card transactions in both the h and l sectors. ■

In comparison, I now characterize the model equilibrium under the interchange cap regulation. Under the regulation, the card network needs to solve the problem (17) subject to the cap constraint (20) in addition to (18)-(19). The goal here is to derive conditions that rationalize the card network's pricing response to the cap regulation as seen in the market. Namely, under the regulation, the card network charges a single interchange fee exactly at the cap level $\overline{a}$. As a result, merchants in the h sector continue to accept card, but merchants in the l sector do not.

Recall Eq (29) that $a^h(\chi_l = 1) > a^h(\chi_l = 0) > a^l$. For the purpose stated, I consider a cap level $\overline{a}$ that satisfies $a^h(\chi_l = 0) \geq \overline{a} > a^l$. This ensures that the cap is binding for the h sector regardless of whether or not the l sector is served with card services.[32] I now

[32]Note that if the cap value $\overline{a}$ is set at a level such that $a^h(\chi_l = 1) > \overline{a} > a^h(\chi_l = 0)$, the cap would not be binding for the h sector if the l sector is dropped out of the card services. The case could be a theoretical possibility, but is less relevant for explaining the market reality.

establish the following proposition.

Proposition 2 *Given any interchange cap $\overline{a}$ that satisfies $a^h(\chi_l = 0) \geq \overline{a} > a^l$, the card network sets a single interchange fee at $\overline{a}$ such that cards are used only in the h sector if the following condition holds*

$$\mu^h(1) - \mu^h(0) < \frac{2\gamma(n+1)^2(-Z^l)}{n[Z^h(1) + \frac{3n+2}{n+2}Z^h(0)]}. \tag{A4}$$

Proof. Given that $a^h(\chi_l = 0) \geq \overline{a} > a^l$, the cap $\overline{a}$ is binding for the h sector regardless of whether or not the l sector is served with card services. Therefore, Eqs (13) and (14) imply that

$$q^h = \frac{n}{2\gamma(n+1)}[\mu^h(\chi_l) + \gamma + \overline{a} - c_B],$$

$$f_B^h = \frac{1}{n+1}[\mu^h(\chi_l) + \gamma + n(c_B - \overline{a})].$$

If both the h and l sectors are served (i.e. $\chi_h = \chi_l = 1$), the total issuers' profit is

$$\Pi^{h+l} = (f_B^h + \overline{a} - c_B)q^h + Z^l = \frac{n}{2\gamma(n+1)^2}[\mu^h(1) + \gamma + \overline{a} - c_B]^2 + Z^l.$$

In contrast, if only the h sector is served (i.e. $\chi_h = 1$, $\chi_l = 0$), the total issuers' profit is

$$\Pi^h = (f_B^h + \overline{a} - c_B)q^h = \frac{n}{2\gamma(n+1)^2}[\mu^h(0) + \gamma + \overline{a} - c_B]^2.$$

Therefore, $\Pi^{h+l} < \Pi^h$ iff

$$\mu^h(1) - \mu^h(0) < \frac{2\gamma(n+1)^2(-Z^l)}{n[\mu^h(1) + \mu^h(0) + 2\gamma + 2\overline{a} - 2c_B]}. \tag{30}$$

Because $\overline{a} \leq a^h(\chi_l = 0)$, a sufficient condition for (30) to hold is that

$$\mu^h(1) - \mu^h(0) < \frac{2\gamma(n+1)^2(-Z^l)}{n[\mu^h(1) + \mu^h(0) + 2\gamma + 2a^h(\chi_l = 0) - 2c_B]}. \tag{31}$$

Inserting the expression of $a^h(\chi_l = 0)$ from Eq (21), the condition (31) can then be rewritten as (A4).

Under Assumptions A1-A2, it is straightforward to verify that

$$\frac{2\gamma(n+1)^2(-Z^l)}{n[Z^h(1)+\frac{3n+2}{n+2}Z^h(0)]} > \frac{\gamma(n+2)^2(-Z^l)}{2n[Z^h(0)+Z^h(1)]}.$$

Therefore, there exists a non-empty set of values that satisfy Assumption A3 and Condition A4. Hence, for any value of $\mu(1)-\mu(0)$ within that set, the card network sets differentiated interchange fees to serve both the h and l sectors in the absence of regulation, and sets a single interchange fee at $\overline{a}$ such that only the h sector is served with the card services under the cap regulation. ■

5 Welfare and policy analysis

I have provided a model that rationalizes card networks' interchange pricing before and after the cap regulation introduced by the Durbin Amendment. The analysis suggests that card demand externalities between the small-ticket and large-ticket sectors could play an important role in explaining card networks' response to the regulation. Based on the model framework, I now take a step further to conduct welfare and policy analysis.

5.1 Welfare maximization

I first consider an alternative regime where the network is run by a social planner who maximizes social welfare.[33] Social welfare is generated whenever consumers use cards for payment at retailers provided consumer-and-merchant combined transaction benefits exceed the combined costs (i.e., $b_S^i + b_B^i > c_B + c_S$), which is shown as

$$\sum_{i\in\{h,l\}} \left(\chi_i \int_{f_B^i}^{\overline{b}_B^i} [b_S^i + b_B^i - c_B - c_S] dH_i(b_B^i) \right). \tag{32}$$

To be comparable with the analysis in the previous section, I assume that the social planner can collectively set card fees (a^l, f_B^l) for small-ticket transactions to internalize

[33] In the welfare analysis, I abstract from the concern that social costs of alternative payment means may deviate from private costs (e.g. the cash and check services are partially sponsored by the government, so social costs of providing those services may diverge from private costs). Those are interesting but separate issues, which are beyond the scope of this paper.

card demand externalities (I will show later that this outcome can indeed be implemented by an alternative interchange cap regulation). Therefore, under the model's distributional assumptions of b_B^h and b_B^l, the social planner sets card fees a^h, a^l, f_B^l to maximize social welfare as follows,

$$W = \max_{a^h, a^l, f_B^l} \frac{\chi_h}{2\gamma}\left([b_S^h - c_B - c_S][\mu^h(\chi_l) + \gamma - f_B^h] + \frac{[(\mu^h(\chi_l)+\gamma]^2 - f_B^{h2}}{2}\right) \tag{33}$$

$$+ [b_S^l + \mu^l - c_B - c_S]\chi_l$$

$$s.t. \ (9), (14), (18), (19).$$

The following proposition characterizes the solution to the welfare maximization problem (33). The results show that under Assumptions A1-A3, the social planner would also set differentiated interchange fees to serve both the h and l sectors, but the fee level in the h sector tends to be lower than that set by the private network.

Proposition 3 *The social planner who maximizes social welfare sets differentiated interchange fees to serve card transactions in both the h and l sectors. In addition, (i) when issuer competition is high (i.e. $n > 2$), the h-sector interchange fee set by the social planner is lower than that set by the private network; (ii) when issuer competition is low (i.e. $n \leq 2$), the h-sector interchange fee set by the social planner coincides with that set by the private network.*

Proof. Consider that the card network is run by a social planner who maximizes social welfare. In the case where issuer competition is high (i.e. $n > 2$), the constraint (18) does not bind. The first order condition with regard to f_B^h yields that

$$\tilde{f}_B^h = c_B + c_S - b_S^h. \tag{34}$$

Eqs (14) and (34) then determine the interchange fee in the h sector

$$\tilde{a}^h = b_S^h - c_S + \frac{Z^h(\chi_l)}{n}. \tag{35}$$

The social planner can also set

$$\tilde{a}^l = b_S^l - c_S \quad \text{and} \quad \tilde{f}_B^l = \mu^l \tag{36}$$

to serve the l sector. Therefore, if the social planner sets a single interchange fee and only serves the h sector, the maximum welfare is determined by (33) as

$$W^h = \frac{[Z^h(0)]^2}{4\gamma}. \tag{37}$$

In contrast, if the social planner sets differentiated interchange fees and serves both the h and l sectors, the maximum welfare is

$$W^{h+l} = \frac{[Z^h(1)]^2}{4\gamma} + Z^l. \tag{38}$$

Under Assumption A3, $W^h < W^{h+l}$, so the social planner prefers the latter.[34] Comparing Eqs (35) and (26) for $\chi_l = 1$ shows that the h-sector interchange fee set by the social planner is lower than that set by the private network.

In the case where issuer competition is low (i.e. $n \leq 2$), the constraint (18) is binding. Hence,

$$\tilde{a}^h = b_S^h - c_S + \frac{n}{n+2} Z^h(\chi_l). \tag{39}$$

Eqs (39) and (14) then determine the consumer fee in the h sector

$$\tilde{f}_B^h = \frac{2n(c_S + c_B - b_S^h) - (n-2)[\mu^h(\chi_l) + \gamma]}{n+2}. \tag{40}$$

Again, the social planner can also set card fees

$$\tilde{a}^l = b_S^l - c_S \quad \text{and} \quad \tilde{f}_B^l = \mu^l \tag{41}$$

to serve the l sector. Therefore, if the social planner sets a single interchange fee and only

[34] For issuers to participate in the card network, they need to make a non-negative profit. This can be satisfied under plausible parameter values, i.e. $[Z^h(1)]^2 > 2\gamma n(-Z^l)$, or the social planner is allowed to conduct lump-sum transfers.

serves the h sector, the maximum welfare is determined by (33) as

$$W^h = \frac{2n[Z^h(0)]^2}{\gamma(n+2)^2}. \tag{42}$$

In contrast, if the social planner sets differentiated interchange fees and serves both the h and l sectors, the maximum welfare is

$$W^{h+l} = \frac{2n[Z^h(1)]^2}{\gamma(n+2)^2} + Z^l. \tag{43}$$

Under Assumption A3, $W^h < W^{h+l}$, so the social planner prefers the latter as well. Note that in this case, the social planner's decision is indeed equivalent to the private network's decision analyzed in Proposition 1. Therefore, the welfare-maximizing interchange fees coincide with those set by the private network. ■

The welfare findings can be intuitively explained as follows. There are two counteracting distortions in the card payment system that we consider (particularly in the h sector where consumer transaction benefit of card usage follows a non-degenerate distribution).[35] On the one hand, price coherence allows consumers to pay the same retail price regardless of the payment method they use. As a result, merchants internalize consumers' inframarginal card usage benefits when they decide whether to accept cards. This raises the interchange fee that merchants are willing to accept.[36] On the other hand, issuers impose a markup when setting consumer fees, which drives down the inframarginal card usage benefits and lowers the interchange fee that merchants are willing to accept. In the case where the issuers' market power is small (i.e. $n > 2$), the distortion due to price coherence dominates, so the privately determined interchange fee in the h sector exceeds the socially optimal level. In the case where the issuers' market power is large (i.e. $n \leq 2$), the distortion due to issuer markup dominates. However, because the so-

[35] Because of the assumption that consumers' transaction benefit of card usage in the l sector follows a degenerate distribution, we abstract from distortion in the l sector *per se* (which is supposed to be sufficiently small anyway). However, the regulator and the private network still have different objectives for internalizing cross-sector card demand externalities.

[36] As mentioned before, the analysis in this paper can be carried over to the framework of Rochet and Tirole (2002), where heterogenous consumers differ systematically in their transaction benefits of using cards. In that framework, price coherence implies that cash-paying consumers are subsidizing those who use cards.

cially optimal interchange fee is limited by the merchant card acceptance constraint, the privately determined interchange fee coincides with the social optimum.[37]

In spite of the result that the privately determined interchange fee in the h sector may exceed the socially optimal level, we find that under the same set of assumptions, the social planner behaves similar to the private network by setting differentiated interchange fees to serve card transactions in both the h and l sectors. Essentially, both the social planner and the private network treat the transactions in the l sector as a loss leader. In doing so, they subsidize the l-sector card transactions in order to internalize the positive externalities of card usage between the l and h sectors.

A similar analysis can be done if we assume that the social planner maximizes total user surplus instead of social welfare. Total user surplus is the sum of consumer surplus and merchants' profit (but not issuers' profit).[38]Focusing on total user surplus is legitimate when card issuers' profit is dismissed by competition authorities. In this case, the results turn out to be even stronger. Under plausible parameter values, I again find that the social planner who maximizes total user surplus would set differentiated interchange fees to serve card transactions in both the h and l sectors. Moreover, the resulting interchange fees in both the h and l sectors are lower than those maximizing total issuers' profit or social welfare. The proof of the results can be found in the Appendix.

5.2 Alternative regulation

The analysis above shows that privately determined interchange fees tend to be too high (based on the criterion of social welfare maximization or total user surplus maximization), a finding consistent with previous studies. This implies that payment cards could be overused at equilibrium. Therefore, lowering interchange fees may potentially improve payments efficiency, which provides some justification for regulating interchange fees.

However, Eq (35) points out that the socially optimal h-sector interchange fee is determined by multiple factors including merchant-and-consumer net benefits $Z^h(\chi_l)$, merchant

[37]In reality, a card network typically has a large number of issuers. Therefore, it is likely that the privately determined interchange fee exceeds the socially optimal level.

[38]Maximizing total user surplus is the criterion Rochet and Tirole (2011) used to derive the optimal interchange fee regulation based on the "merchant avoided-cost test."

transaction benefit b_S^h, issuer competition n, and the acquirers' cost c_S. The finding suggests that the Durbin regulation that requests interchange fees to be capped by issuers' marginal cost c_B lacks theoretical foundation.

Our analysis also suggests that in the presence of card demand externalities, capping the maximum interchange fee may not restore the social optimum because it could adversely affect small-ticket transactions, as shown by Proposition 2. On the other hand, policymakers may not want to directly regulate card fees in all merchant sectors (including interchange fees and consumer fees) because that would risk being too heavy-handed. Therefore, an interesting question is whether we could design an alternative interchange cap regulation restoring the social optimum. The following discussion illustrate how this can be done conceptually.

Proposition 4 *In the presence of card demand externalities across merchant sectors, capping the weighted average interchange fee, instead of the maximum interchange fee, may restore the social optimum.*

Proof. Consider a regulator who maximizes social welfare (a similar analysis can be done using the criterion of total user surplus). When $n > 2$, the privately determined interchange fee $a^h(\chi_l = 1)$ given by (26) exceeds the welfare-maximizing level $\tilde{a}^h(\chi_l = 1)$ given by (35). Assume that Assumptions A1-A3 and Condition A4 hold, so both the social planner and the private network would want to serve card transactions in the h and l sectors. Following the analysis in Section 2, I focus on the scenario where $\tilde{a}^h(\chi_l = 1) < a^h(\chi_l = 0)$.[39] In this case, as suggested by Proposition 2, capping the maximum interchange fee at $\tilde{a}^h(\chi_l = 1)$ would not restore the social optimum because the card network and issuers would stop subsidizing the l sector.

Instead, the regulator could consider a weighted average cap (WAC) on interchange fee as follows:

$$\lambda y^l \tilde{a}^l + \max(1 - \lambda, 1 - y^l) a^h \leq \bar{a}, \tag{WAC}$$

where $\bar{a}$ is the cap, $0 \leq \lambda \leq 1$ is the weight chosen by the regulator, and $\tilde{a}^l = b_S^l - c_S$ is the socially optimal l-sector interchange fee. The WAC rule essentially requires $a^h \leq \bar{a}$

[39] Eqs (21) and (35) imply that $\tilde{a}^h(\chi_l = 1) < a^h(\chi_l = 0)$ whenever $Z^h(1) < \frac{n^2}{n+2} Z^h(0)$.

when $y^l = q^l \chi_l = 0$ but allows $\lambda \tilde{a}^l + (1-\lambda) a^h \leq \bar{a}$ when $y^l = 1$, so it imposes a penalty on the maximum permissible level of a^h in case card networks and issuers raise fees to shut off small-ticket transactions.

Recall that the welfare-maximizing interchange fees are given by (35) and (36) that

$$\tilde{a}^h = \frac{Z^h(1)}{n} + b_S^h - c_S, \quad (44)$$

$$\tilde{a}^l = b_S^l - c_S. \quad (45)$$

Note that $\tilde{a}^h > \tilde{a}^l$ because $b_S^h > b_S^l$ and $Z^h(1) > 0$. The corresponding total issuers' profit is determined by (16), (44) and (45) as

$$\Pi^{h+l}(\tilde{a}^h, \tilde{a}^l) = \frac{n[\mu^h(1) + \gamma + \tilde{a}^h - c_B]^2}{2\gamma(n+1)^2} + Z^l. \quad (46)$$

Denote $\Pi^h(\tilde{a}^l)$ as the total issuer's profit for only serving the h sector at the interchange fee level $\tilde{a}^l = b_S^l - c_S$, so

$$\Pi^h(\tilde{a}^l) = \frac{n[\mu^h(0) + \gamma + \tilde{a}^l - c_B]^2}{2\gamma(n+1)^2}.$$

Under plausible parameter values, we have that[40]

$$\Pi^{h+l}(\tilde{a}^h, \tilde{a}^l) > \Pi^h(\tilde{a}^l) > 0. \quad (47)$$

Given that $\tilde{a}^h > \tilde{a}^l$, (47) implies that there exists a non-empty set of values that $\bar{a}$ can take, which satisfy $\tilde{a}^h > \bar{a} > \tilde{a}^l$ and

$$\Pi^{h+l}(\tilde{a}^h, \tilde{a}^l) > \Pi^h(\bar{a}) > \Pi^h(\tilde{a}^l) > 0. \quad (48)$$

The regulator can then choose any cap value $\bar{a}$ from this set and determine the corresponding weight λ by solving

$$\lambda \tilde{a}^l + (1-\lambda)\tilde{a}^h = \bar{a}. \quad (49)$$

[40]Note that the first inequality in (47) holds if $\left[\frac{n}{n+1} Z^h(1)\right]^2 - \left[Z^h(0) - (b_S^h - b_S^l)\right]^2 > \frac{n(-Z^l)}{2\gamma(n+1)^2}$, and the second inequality holds if $Z^h(0) > b_S^h - b_S^l$.

Given $\bar{a}$ and λ determined by (48)-(49), if the card network raises the l-sector interchange fee above $\tilde{a}^l$ (or if issuers raise the l-sector consumer fee so that $f_B^l > \mu^l$) and hence only serves the h sector, the highest total issuers' profit that can be achieved is $\Pi^h(\bar{a})$ under the WAC rule. In contrast, if the card network serves the l-sector with the interchange fee $\tilde{a}^l$, it is allowed to set the h-sector interchange fee as high as $\tilde{a}^h$, so the maximal profit is $\Pi^{h+l}(\tilde{a}^h, \tilde{a}^l)$. As (48) shows, the latter is more profitable. On the other hand, the card network has no incentive to set the l-sector interchange fee below $\tilde{a}^l$ under the WAC rule because this will result in further losses in the l sector but not help increase the h-sector interchange fee beyond the level of $\tilde{a}^h$. ■

Proposition 4 shows that a regulation capping the weighted average interchange fee may restore the social optimum. The analysis has intuitive implications for implementation. In principle, policymakers could design the interchange cap in such a way that penalizes card networks and issuers in case they stop subsidizing small-ticket transactions. Under the WAC rule that we consider, card networks and issuers would have to face a lower interchange cap for large-ticket transactions if they make fee adjustments that adversely affect card usage for small-ticket transactions. As long as the penalty is sufficiently large (cf conditions (48) and (49)), this will provide incentives for card networks and issuers to keep the small-ticket discounts and therefore internalize card demand externalities across merchant sectors.

6 Conclusion

The recent U.S. debit card regulation introduced by the Durbin Amendment to the Dodd-Frank Act has generated some unintended consequences. While the regulation was intended to lower merchant card acceptance costs by capping the maximum interchange fee, small-ticket merchants find their fees instead higher.

In this paper, I address this puzzle by introducing card demand externalities into a two-sided market model. The findings rationalize the card networks' response to the cap regulation: Before the regulation, card networks were willing to offer discounted interchange fees to small-ticket merchants because their card acceptance boosts consumers' card usage for large-ticket purchases from which card issuers can collect higher interchange fees. After the regulation, however, card issuers profit less from this kind of externality, so they discontinued the discounts. Based on the model, I then study socially optimal interchange fees and alternative interchange regulation. The analysis suggests that the social optimum generally would require lower interchange fees than those chosen by the private market, but nevertheless it may maintain the differentiated fee structure in order to internalize the positive externalities of card usage across merchant sectors. In this case, simply capping the maximum interchange fee would not restore the social optimum because of the side effect on small-ticket transactions. As an alternative, I propose a cap regulation based on the weighted average interchange fee.

Overall, the takeaway of the paper is that interchange fees encompass more than just the costs of processing payment card transactions. In the two-sided market, they also serve to balance demand between consumers and merchants, as well coordinate acceptance and usage among different merchant sectors. The model shows that privately determined interchange fees tend to be too high to maximize social welfare, so regulating down interchange fees may help to improve the market outcome. But regulation that only considers one-sided market logic (setting fees equal to issuers' costs, for example) or one sector of the market (ignoring card demand externalities between large- and small-ticket merchants, for example) may result in unintended consequences.

There are several avenues for future research. First, it would be useful to quantify the card demand externalities across merchant sectors. This can be an important step for assessing the empirical impact of current as well as alternative interchange regulations. Second, it would be useful to consider policy options other than price regulation. For instance, in theory, if merchants can set different retail prices conditioning on payment means, interchange fees may become less of an issue.[41] Finally and more broadly, our analysis can be extended beyond payment cards or two-sided markets. Policymakers may always want to be alert to cross-sector externalities when designing regulatory policies so that unintended consequences can be reduced or avoided.

[41] However, those policy options may also have their own limitations, so some cautions need to be taken. For example, in countries where card surcharging is allowed, few merchants choose to do so. Moreover, for some merchants who are indeed surcharging, they are found surcharging excessively or in nontransparent ways. See Hayashi (2012).

Appendix: Total user surplus maximization

Total user surplus is generated whenever consumers use cards for payment at retailers provided consumer-and-merchant joint transaction benefits exceed the joint fees that they pay, namely $b_S^i + b_B^i > f_B^i + f_S^i$. In other words, total user surplus is the sum of consumer surplus and merchants' profit (but not issuers' profit). The expression of total user surplus can be derived from (17) and (33), as shown below. Again, I assume that the social planner can collectively set card fees (a^l, f_B^l) for small-ticket transactions to internalize card demand externalities. Therefore, the social planner sets card fees a^h, a^l, f_B^l to maximize total user surplus as follows:

$$TUS = \max_{a^h, a^l, f_B^l} \frac{\chi_h}{2\gamma}\left([b_S^h - c_B - c_S][\mu^h(\chi_l) + \gamma - f_B^h] + \frac{[(\mu^h(\chi_l) + \gamma]^2 - f_B^{h2}}{2}\right) \quad (50)$$

$$- \frac{\chi_h}{2\gamma}\left(\frac{n[\mu^h(\chi_l) + \gamma + a^h - c_B]^2}{(n+1)^2}\right) + (b_S^l - c_S - a^l)\chi_l$$

$$s.t. \quad \frac{n[\mu^h(\chi_l) + \gamma + a^h - c_B]^2}{2\gamma(n+1)^2}\chi_h + (f_B^l + a^l - c_B)\chi_l \geq 0, \quad (51)$$

$$\text{and} \quad (9), (14), (18), (19).$$

The constraint (51) requires that card issuers make a non-negative profit.

The following proposition characterizes the solution to the problem. The results show that under plausible parameter values, the social planner who maximizes total user surplus would also set differentiated interchange fees to serve both the h and l sectors, but the interchange fees in both sectors are lower than those maximizing total issuers' profit or social welfare.

Proposition 5 *The social planner who maximizes the total user surplus sets differentiated interchange fees to serve card transactions in both the h and l sectors, and the interchange fees in both sectors are lower than those maximizing total issuers' profit or social welfare if*

$$[Z^h(1)]^2 > \frac{2\gamma(n+2)^2(-Z^l)}{n}. \quad (A5)$$

Proof. Consider that the card network is run by a social planner who maximizes total user surplus. The constraint (18) never binds and the first order condition with regard to f_B^h yields

$$\tilde{f}_B^h = c_B + c_S - b_S^h + \frac{2}{n+2} Z^h(\chi_l). \tag{52}$$

Eqs (14) and (52) then determine the interchange fee

$$\tilde{a}^h = b_S^h - c_S - \frac{Z^h(\chi_l)}{n+2}. \tag{53}$$

Therefore, if the social planner only provides card services to the h sector but not the l sector, it is optimal to set

$$\tilde{a}^h(\chi_l = 0) = b_S^h - c_S - \frac{Z^h(0)}{n+2}. \tag{54}$$

The total user surplus is

$$TUS^h = \frac{n}{4\gamma(n+2)} [Z^h(0)]^2, \tag{55}$$

and issuers make the total profit

$$\Pi^h = \frac{n}{2\gamma(n+2)^2} [Z^h(0)]^2. \tag{56}$$

Alternatively, if the social planner also provides card services to the l sector, it is optimal to set

$$\tilde{a}^h(\chi_l = 1) = b_S^h - c_S - \frac{Z^h(1)}{n+2}, \tag{57}$$

and set the lowest interchange fee $\tilde{a}^l$ that satisfies the constraint (51):

$$\tilde{a}^l = b_S^l - c_S - \left(\frac{n}{2\gamma(n+2)^2} [Z^h(1)]^2 + Z^l \right). \tag{58}$$

The total user surplus is

$$TUS^{h+l} = \frac{n^2 + 4n}{4\gamma(n+2)^2} [Z^h(1)]^2 + Z^l, \tag{59}$$

and the total issuers' profit is

$$\Pi^{h+l} = 0. \tag{60}$$

Under Condition A5, $TUS^{h+l} > TUS^{h}$, so the social planner achieves higher total user surplus by setting differentiated interchange fees to serve card transactions in both the h and l sectors. Moreover, under Condition A5, Eqs (57) and (58) confirm that the interchange fees in both the h and l sectors are lower than those maximizing total issuers' profit (given by (26) and (27)) or those maximizing social welfare (given by (35) and (36) when $n > 2$ or (39) and (41) when $n \leq 2$). ■

References

[1] Armstrong, Mark (2006), "Competition in Two-Sided Markets," *RAND Journal of Economics*, 37(3): 668–681.

[2] Baxter, William (1983), "Bank Interchange of Transactional Paper: Legal Perspectives," *Journal of Law and Economics*, 26: 541-588.

[3] Bedre-Defolie, Özlem, and Emilio Calvano (2013), "Pricing Payment Cards," *American Economic Journal: Microeconomics*, 5(3): 206-31.

[4] Bolt, Wilko and Sujit Chakravorti (2008), "Consumer Choice and Merchant Acceptance of Payment Media," DNB Working Papers 197, Netherlands Central Bank.

[5] Cabral, Luis (2005), "Market Power and Efficiency in Card Payment Systems: A Comment," *Review of Network Economics*, 5: 15-25.

[6] Carlton, Dennis and Alan Frankel (1995), "The Antitrust Economics of Credit Card Networks," *Antitrust Law Journal*, 63: 903-915.

[7] Foster, Kevin, Erik Meijer, Scott Schuh and Michael Zabek (2010), "The 2008 Survey of Consumer Payment Choice," Public Policy Discussion Paper, No. 09-10, Federal Reserve Bank of Boston.

[8] Gans, Joshua and Stephen King (2003), "The Neutrality of Interchange Fees in Payment Systems," *Topics in Economic Analysis & Policy,* 3 (1): Article 1.

[9] Hayashi, Fumiko (2012), "Discounts and Surcharges: Implications for Consumer Payment Choice," *Payments System Research Briefings,* June, Federal Reserve Bank of Kansas City.

[10] Hayashi, Fumiko (2013), "The New Debit Card Regulations: Effects on Merchants, Consumers, and Payments System Efficiency," *Federal Reserve Bank of Kansas City Economic Review*, Q1: 89-118.

[11] Katz, Michael (2001), "Reform of Credit Card Schemes in Australia II: Commissioned Report" RBA Public Document. August.

[12] McAndrews, James and Zhu Wang (2012), "The Economics of Two-Sided Payment Card Markets: Pricing, Adoption and Usage." Research Working Paper 12-08, Federal Reserve Bank of Richmond.

[13] Prager, Robin, Mark Manuszak, Elizabeth Kizer, and Ron Borzekowski (2009), "Interchange Fees and Payment Card Networks: Economics, Industry Developments, and Policy Issues," *Finance and Economics Discussion Series* 2009-23. Washington: Board of Governors of the Federal Reserve System.

[14] Rochet, Jean-Charles and Jean Tirole (2002), "Cooperation among Competitors: Some Economics of Payment Card Associations," *RAND Journal of Economics*, 33: 549-570.

[15] Rochet, Jean-Charles and Jean Tirole (2006), "Two-Sided Markets: A Progress Report," *RAND Journal of Economics*, 35: 645-667.

[16] Rochet, Jean-Charles and Jean Tirole (2011), "Must-Take Cards: Merchant Discounts and Avoided Costs," *Journal of the European Economic Association,* 9 (3): 462–495.

[17] Rochet, Jean-Charles and Julian Wright (2010), "Credit Card Interchange Fees," *Journal of Banking and Finance*, 34: 1788-1797.

[18] Rysman, Marc (2007), "An Empirical Analysis of Payment Card Usage," *Journal of Industrial Economics*, 60: 1-36.

[19] Rysman, Marc (2009), "The Economics of Two-Sided Markets," *Journal of Economic*

Perspectives, 23: 125-144.

[20] Schmalensee, Richard (2002), "Payment Systems and Interchange Fees," *Journal of Industrial Economics*, 50: 103-122.

[21] Shy, Oz (2012), "Who Gains and Who Loses from the 2011 Debit Card Interchange Fee Reform?" *Public Policy Discussion Paper,* No. 12-6, Federal Reserve Bank of Boston.

[22] Schwartz, Marius and Daniel Vincent (2006), "The No Surcharge Rule and Card User Rebates: Vertical Control by a Payment Network," *Review of Network Economics,* 5: 72-102.

[23] Shy, Oz and Zhu Wang (2011), "Why Do Payment Card Networks Charge Proportional Fees?" *American Economic Review*, 101 (4): 1575-1590.

[24] Wang, Zhu (2010), "Market Structure and Payment Card Pricing: What Drives the Interchange?" *International Journal of Industrial Organization*, 28: 86-98.

[25] Wang, Zhu (2012), "Debit Card Interchange Fee Regulation: Some Assessments and Considerations," *Federal Reserve Bank of Richmond Economic Quarterly*, 98 (3): 159-183.

[26] Weyl, Glen (2010), "A Price Theory of Multi-Sided Platforms," *American Economic Review*, 100 (4): 1642-1672.

[27] Wright, Julian (2003), "Optimal Card Payment Systems," *European Economic Review*, 47: 587-612.

[28] Wright, Julian (2004), "Determinants of Optimal Interchange Fees in Payment Systems," *Journal of Industrial Economics*, 52: 1-26.

[29] Wright, Julian (2010), "Why Do Merchants Accept Payment Cards?" *Review of Network Economics*, 9, Article 1.

[30] Wright, Julian (2012), "Why Payment Card Fees Are Biased Against Retailers," *RAND Journal of Economics*, 43: 761-80.