Introduction

It is not too soon for policymakers to begin thinking about the reforms to the *financial architecture*, broadly conceived, that could prevent a similar crisis from developing in the future. [...] First, we must address the problem of financial institutions that are deemed too big – or perhaps too interconnected – to fail. Second, we must strengthen what I will call the financial infrastructure: the systems, rules, and conventions that govern trading, payment, clearing, and settlement in financial markets.
— Ben Bernanke, March 2009, *Financial Reform to Address Systemic Risk*

Financial institutions, market design are of first-order importance (see, for example, Allen, 2001; Thakor, 1996, and the references therein). The financial system architecture shapes the decisions of market participants and intermediaries. In neoclassical theory, a frictionless environment implies the irrelevance of market design (Merton and Bodie, 2005). However, in practice financial markets are hardly ever perfect: Frictions such as agency problems, asymmetric information, or contract incompleteness generate a wedge between neoclassical equilibria and observed market outcomes. Consequently, there is scope for financial system regulation (Degryse, 2007).

Financial market design is particularly challenging due to the complexity of the system. Innovation occurs at a very rapid pace: both through new products (e.g., collateralized debt obligations) and through new technologies (e.g., high frequency trading). At the same time, financial institutions are increasingly interconnected, usually through opaque, short-term contracts. One lesson from the recent global financial crisis is that market regulation needs to be both flexible and systemic in nature. A well-crafted financial system architecture reduces frictions and improves resource allocation (Boot and Thakor, 1997). The importance of market design is well recognized: The 2014 Nobel Memorial Prize in Economic Sciences was awarded to Jean Tirole, in part for his insights on optimal financial market regulation.

Modern financial intermediaries and capital markets are strongly intertwined. Market-based activities of banks are sometimes outside the perimeter of regulation (shadow banking, securitization). Consequently, capital market dynamics – such as the collapse of the U.S. housing price boom – can have major negative consequences on the solvency of intermediaries. Conversely, large financial institutions have key roles in trading and settlement as, for example, systemically important counterparties. It is natural to consider both capital markets and banks when re-approaching financial architecture design in the aftermath of the crisis (Song and Thakor, 2010). The G20 meeting in
Washington calls for a reassessment of the scope of financial regulation: “a special emphasis on institutions, instruments and markets that are currently unregulated, along with ensuring that all systemically-important institutions are appropriately regulated.” (International Monetary Fund, 2009)

In my thesis, I address Bernanke’s (2009) agenda on financial system architecture. Each chapter of this dissertation studies a specific post-crisis (proposed) reform in either the banking sector or capital markets. Policymakers in both Europe and the U.S. are rethinking the financial system. In 2016, a Single Resolution Mechanism will deal with banks in default across the Eurozone. The Dodd-Frank (in the United States, 2010) and EMIR (in Europe, 2012) acts mandate the central clearing of financial assets to reduce counterparty risk. The Flash Crash of May 6, 2010 sparked a debate in both academic and policy circles about the need to regulate high frequency trading.

I contribute to the academic and policy debate with three theoretical and empirical essays on recent financial architecture developments: How should we design better banking regulation? securities exchanges? post-trading infrastructures? The remainder of the introductory chapter is structured as a “road map” for the dissertation: I introduce, motivate, and briefly discuss the key insights of each of the three essays.

The design of banking regulation

The global financial crisis ignited the debate around a common European bank regulator — a banking union for the Eurozone. Ferry and Wolff (2012) look at the fiscal implications of a joint European regulator and advocate a common fiscal backstop mechanism. Colliard (2013) studies the optimal architecture of the single supervision mechanism (SSM), the supervisory pillar of the banking union, and argues there is a conflict of objectives between local and joint supervisors.

Chapter 2, based on Zoican and Górnicka (2014), shifts the debate to the Single Resolution Mechanism (SRM), the second important pillar of the banking union. Under the SRM the banking union, rather than the national authorities, decides whether to bail out an insolvent bank.

I argue that the SRM generates tension between increased regulatory efficiency in responding to bank defaults, on the one hand, and weaker commitment to liquidate failed systemic institutions, on the other hand. If the latter effect dominates, internationally-oriented banks take on more risk, augmenting existing moral hazard problems. If banks hold opaque assets, such as structured derivatives, the negative impact of the SRM is particularly large. A hybrid system in which the SRM coexists with the current national resolution mechanisms is arguably a superior arrangement, as it can both prevent insolvency contagion and keep risk taking incentives low.

Centralized bank default resolution has undeniable benefits. The International Monetary Fund (2013) argues the joint regulator reduces insolvency contagion through an intervention policy that minimizes costs for all member countries. If a multinational bank defaults, a centralized decision bypasses protracted negotiations between national authorities (see the Dexia, Fortis bailouts). Most importantly, the banking union is in the unique position to limit the spread of losses across European banks.
Cross-exposures of Eurozone banks are highly asymmetric: Banks in Germany, France, and the Netherlands have net loans of almost one trillion dollars to counterparts in Greece, Italy, Ireland, Portugal, or Spain. Cross-border links between banks create the scope for default contagion, as noted by Freixas, Parigi, and Rochet (2000). National authorities face a natural tension between domestic taxpayers and foreign creditors. If a highly indebted bank becomes insolvent, taxpayer money is needed to bail it out. A large part of the funds, however, is claimed by foreign institutions. Domestic authorities are understandably biased towards domestic taxpayers (Allen, Carletti, and Gimber, 2011). They are more likely to impose losses on international creditors through the liquidation of a heavily indebted bank. On the other hand, a banking union represents the interests of both domestic and foreign actors. To avoid contagion, it is more likely to be lenient with systemic international banks. Systemic contagion risks are contained and welfare improves.

Since banks in larger countries have more consistent international exposures, they benefit most from lower contagion risk. A banking union offers them a form of systemic risk insurance. Consequently, the contributions of Eurozone core countries to the common resolution fund should include a premium. Banks in Germany, France, or the Netherlands will be able to invest in growing European markets while bearing a smaller contagion risk than under the existing mechanism.

**Too-European-to-fail problem** While a lenient banking union minimizes contagion, it can also engender risky behavior at the bank level. Internationally indebted banks might become too-European-to-fail: their systemic status is re-emphasized. A lesson from the past crisis is that systemic banks are able to take on increasing amounts of risk while relying on implicit regulatory support. Is too-European-to-fail a major concern? I argue so: Modern banks hold large positions in opaque and complex assets, such as structured products and derivatives. For these asset classes, a marginal drop in risk management standards can have a very large negative impact on the returns and solvability of the bank.

**Alternative design solution** A hybrid regulatory architecture solves the banks incentive problem. The banking union and the national regulators coexist. The banking unions mandate, however, is limited to European-wide crises, e.g., triggered by an external shock such as the United States house market crash. National crises are dealt with by national authorities who bear the full costs of insolvent bank interventions. Contagion risks are contained when it matters most, but internationally systemic banks sometimes face a tougher stance from national regulators. It can prove to be the best of both worlds.

**The design of securities exchanges** Chapter 3, based on Menkveld and Zoican (2014) turns to the design of equity markets. Do exchanges get better when trading platforms reduce their latency? The landscape of equity markets changes at a fast pace. Exchanges invest in fast trading platforms to reduce order execution times for their members. However, speeding up the exchange does not necessarily improve market quality.
CHAPTER 1. INTRODUCTION

Few activities embraced the computer age so actively as trading. Loud and hectic pits have been progressively replaced by silent computer server rooms. Transactions are no less dynamic for it, however. A London-based trader can buy stocks in Frankfurt within 2.21 milliseconds. Light needs only 2.12 milliseconds to travel the same distance.

There is a very active ongoing debate around high-frequency trading. Supporters claim that in the few years since algorithmic trading took off, market liquidity and exchange competition improved. Critics point to aggressive high-frequency trading strategies that generate losses for human investors. Michael Lewis’ book “Flash Boys,” a champion of the latter view, is a very well-known example. How exactly does high-frequency trading affect markets? How can researchers and policymakers improve financial markets in the twenty-first century?

The benefits of algorithmic and fast trading  High-frequency traders (HFTs) have a comparative advantage in providing liquidity as market-makers. Market-makers are suppliers of quotes: a bid price, at which they are ready to buy an asset, and an ask price, at which they are ready to sell it. The difference between the two is referred to as the spread, and represents the profit of the market-maker.

Why are high-frequency traders better at making markets? Their computer algorithms monitor in real time all information relevant to the traded asset: news headlines, demand and supply changes, or related assets data. HFTs are able to incorporate this wealth of information into their price quotes faster than anyone else. Two advantages follow directly. First, price discovery improves: Price quotes accurately reflect all available information with minimal lag as documented by Riordan and Storkenmaier (2012). Second, a savvy trader could have exploited the delay between news and price updates to earn a profit at the market-makers expense. Fast trading minimizes this delay: The risk for an HFT market-maker is lower than for a human one. Consequently, HFTs are able to charge lower spreads. Trading costs are smaller for everybody (Hendershott and Riordan, 2011).

Algorithmic traders also promote competition between exchanges. In the past, assets were only traded at a single exchange. It made sense: having all potential buyers and sellers in one place increases the likelihood to find counterparties. The exchange had a natural monopoly and the power to set large fees. Algorithmic trading made it easier to automatically search for counterparties across multiple exchanges (Menkveld, 2013). Computer traders can take a position from a seller in one market and offload it to a buyer on a different one. There is no need for everybody to trade in the same place anymore. Under renewed competitive pressure, exchanges decrease trading fees.

The costs of algorithmic and fast trading  High-frequency traders are not always market makers, however. They can be, for instance, speculators. Foucault, Kadan, and Kandel (2013) argue that speculator-type HFTs trade on quickly-processed information to take advantage of other participants delay in updating price quotes. A natural reaction of HFT market-makers is to become even faster to close the gap. A socially costly arms race ensues, with all high-frequency traders aiming to marginally increase speed.
Are faster exchanges always good? A starting point to answer this question is to acknowledge the empirical evidence showing high-frequency traders behave both as market-makers and as speculators. Hagstromer and Norden (2013) document such order type specialization. A speed improvement of a few microseconds directly affects high-frequency traders, irrespective of their strategy. Since human reaction time is hundreds of milliseconds, it does not directly affect human traders.

High frequency traders act both as market makers and as short term speculators. In a low latency market, high frequency market makers are able to update their price quotes faster. At the same time however, high frequency speculators are also able to act faster on new information and profit by trading against quotes with an outdated information content. Additionally, the speed gap between high and low frequency traders widens.

As the exchange latency drops, high frequency market makers trade more often with high frequency speculators. The market makers face higher expected adverse selection costs, raising the bid-ask spreads as a response. At the same time, a lower latency offers market makers better incentives to monitor the asset value: as the adverse selection cost increase, reducing it through quote monitoring becomes optimal. To be able to quote the lowest possible spread, competitive market makers acquire more information in faster markets.

There are economies of scope from information acquisition. Market makers posting quotes on both sides of the market earn two times as much as market makers posting on only one side of the market. Both incur the same monitoring cost, however: monitoring a single quote is thus relatively more expensive. Market makers will monitor one sided quotes only in very fast markets: the adverse selection risk needs to be large enough to compensate for the relatively higher monitoring cost. Informed market makers can exploit the economies of scope and earn rents by being the first ones to submit a quote.

To empirically test the implications of the model, we use a natural experiment on the Nordic markets. On February 8, 2010, NASDAQ OMX introduces the INET Core Technology on the equity markets in Denmark, Finland, and Sweden. The latency drops from 2.5 milliseconds to 0.25 milliseconds, the lowest across all exchanges trading in Nordic equities. Following the technology upgrade, adverse selection costs for high frequency market makers increase by 2.50 basis points, the effect being stronger for assets with higher volatility. A standard deviation increase in volatility leads to a 1.95 bps increase in the average adverse selection costs. After the INET implementation, the effect is stronger by 1.55 basis points.

A market design challenge High-frequency trading has complex effects on market quality. Overall, evidence shows that the transition to the computer age improved market liquidity. However, after some point, increasingly faster trading has less to do with better liquidity provision and more with the interactions between high-frequency traders themselves. The social benefits of prices reflecting information one microsecond earlier are most likely not so important. Being one microsecond faster than a competitor just might tip the balance. How fast should trading be exactly? Can we design
different markets that maximize the benefits and eliminate the costs? For example, Budish, Cramton, and Shim (2013) advocate replacing existing limit order markets with a frequent auctions setup.

**The design of post-trading infrastructure**

A third recent financial architecture development refers to the post-trade infrastructure design. The design of clearing infrastructure is the focus of Chapter 4, based on Menkveld, Pagnotta, and Zoican (2013).

In the aftermath of the recent global financial crisis, a growing consensus has emerged on the fragility of bilateral clearing designs and on the benefits of a central clearing counterparty (CCP). Two important regulatory acts, the Dodd-Frank in the United States and the EMIR in Europe, mandate the central clearing of over-the-counter-derivatives. Although the systemic risk benefits of a CCP are well recognized in the literature, it is less clear how the mandatory introduction of a CCP may directly affect market quality. The analysis of voluntary central clearing reforms, as in Loon and Zhong (2014), may overestimate the positive effects of a CCP introduction.

**The debate** Ex ante, the relationship between clearing and market volatility is not trivial. One potential channel is counterparty risk: Bernstein, Hughson, and Weidenmier (2014) find that central clearing reduces the risk of default contagion. In turbulent periods, lower default contagion should limit fire sales and consequently price volatility.

As a CCP requires higher collateral (margins) than alternative regimes, a second channel emerges. However, the relationship between margins and asset volatility is complex. On the one hand, Pirrong (2011) argues that rigid collateral policies can exacerbate price movements (larger volatility). On the other hand, recent work by Rytchkov (2014) suggests that higher margins reduce volatility as they reduce the wedge between optimal portfolios of risk-averse and risk-tolerant traders.

**The experiment** An quasi-experimental design is used to quantify the effect of mandatory central clearing on asset volatility, as well as the interactions with market liquidity and price efficiency. Equities in Sweden, Denmark and Finland (henceforth, the Nordic markets) were traditionally cleared bilaterally through a given local Central Securities Depository (CSD). In October 2009, however, the large majority of stocks in the Nordic markets migrated to central clearing through the European clearing house EMCF.

The mandatory nature of the reform is crucial. Voluntary clearing was possible for six months before the reform. However, traders on Nordic markets did not move to centralized clearing before its mandatory implementation. Less than 1% of volume was cleared through EMCF in September 2009, one month before the reform. The finding can be partly explained by the higher collateral requirements: Voluntarily cleared volume was significantly lower in stocks with higher margin requirements.

To assess the impact of the mandatory reform, a quasi-experimental difference-in-differences approach is used. First, Nordic stocks are matched with European equities using propensity scores.
on a number of dimension: size, liquidity, systematic risk. Then, the time-series behaviour around the reform date for Nordic stocks relative to the matched controls captures the effect of central clearing.

**The results**  We find that mandatory central clearing reduces volatility. Following the reform, daily squared returns for Nordic stocks decreased by 21 basis points: a 9.21% decline. Lower volatility is not associated with better market efficiency, as in Foucault, Sraer, and Thesmar (2011). Also, turnover drops by 11.27% – effect which is primarily driven by volatility. As in Lo and Wang (2000) and Rytchkov (2014), lower volatility leads to lower portfolio re-balancing needs and hedging demand. Consequently, trading activity decreases.

A natural next step is to identify the exact channel using the cross-section of treated Nordic stocks. Does the effect stem from higher margins or lower counterparty risk? Evidence points to the former: The volatility decline is significantly stronger for stocks with higher margin requirements, in line with Rytchkov (2014).