THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Market Quality and Contagion in Fragmented Markets

Rohit Rahi

Jean-Pierre Zigrand

SRC Discussion Paper No 2

September 2013

Systemic Risk Centre

Discussion Paper Series

ISSN xxx-xxxx

Abstract
Financial market liquidity has become increasingly fragmented across multiple trading platforms. We propose an intuitive welfare-based market quality metric that can properly aggregate local market conditions across both securities and trading venues. Our analysis rests on a general equilibrium model with segmented markets. Arbitrageurs reap profits by effectively providing intermediation services (i.e. “liquidity"). Our market quality measure is equal to the additional consumption enjoyed by investors as a result of this intermediation, and can be represented by means of a number of observable proxies. The model is especially well-suited to study the contagion-like effects of liquidity shocks.

JEL classification: G10, G20, D52, D53.
Keywords: Fragmented markets, intermediation, arbitrage, liquidity, contagion.

This paper is published as part of the Systemic Risk Centre’s Discussion Paper Series. The support of the Economic and Social Research Council (ESRC) in funding the SRC is gratefully acknowledged [grant number ES/K002309/1].

Acknowledgements
This paper was circulated earlier under the title “A Theory of Strategic Intermediation and Endogenous Liquidity”. The authors thank the late Sudipto Bhattacharya, Douglas Gale, Joel Peress, Tano Santos, José Scheinkman as well as participants at workshops and seminars at several universities for helpful discussions.

Rohit Rahi is Reader in Finance at the Department of Finance and Financial Markets Group, and Research Associate of Systemic Risk Centre, London School of Economics and Political Science. Jean-Pierre Zigrand is Reader in Finance at the Department of Finance and Financial Markets Group, and Co-Director of Systemic Risk Centre, London School of Economics and Political Science.

Published by
Systemic Risk Centre
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

One of the most disruptive recent changes in the financial industry has been the widespread proliferation of trading venues following the Regulation National Market System (Reg NMS) in the US and the Markets in Financial Instruments Directive (MiFID) in Europe. The same stocks are traded not only on several exchanges but also on alternative trading systems such as Multilateral Trading Facilities (MTFs), Electronic Communication Networks (ECNs) and various Dark Pools.[1] The regulations, which were designed to enhance competition between trading venues, have in turn spawned a new breed of intermediary in the form of high-frequency traders (HFTs) or latency arbitrageurs[2] who trade simultaneously across multiple trading venues in order to exploit, and thus reduce or eliminate, price discrepancies. A very large percentage of trading volume has been attributed to such traders.[3] There is growing concern that competition in security markets in the US and Europe has led to trading liquidity becoming fragmented across too many venues. At the same time, the HFTs, who provide liquidity and help to align prices across venues, have been viewed with suspicion by the press, the traditional real-money investors and even by the regulators who to some extent created the need for this intermediation.

The Flash Crash of May 6th 2010, in which the Dow Jones index fell nearly 10% only to recover a few minutes later, has accelerated that discussion and has brought the topic of modern market making to the forefront. Attention has focused on the interconnectedness of trading venues and the implications for liquidity and welfare. For instance, a report by the CFTC and SEC (CFTC and SEC (2010)) points out that during the Flash Crash, "hot-potato volumes" spiked up as HFTs passed securities around in a musical chair-like fashion within and across trading venues, and shocks were transmitted across markets for stocks, options and futures in a complex fashion. When latency arbitrageurs withdrew from the markets and prices of identical securities diverged across trading venues, panic set in as market

[1]Examples at the time of writing are BATS (merged with Chi-X), Turquoise, Burgundy, ITG Posit, Equiduct, QuoteMTF, Liquidnet, UBS MTF, Sigma X MTF, Instinet Blockmatch, Nomura NX and Smartpool.

[2]Examples of HFTs include proprietary quantitative hedge funds and market makers at firms such as Citadel Group, D.E. Shaw Group, Renaissance Technologies, Getco, Optiver, Knight and Tradebot, as well as trading desks in some of the major investment banks.

[3]Various sources estimate that the fraction of equity trades involving HFT algorithms is 60–70% in the US, 30% in the UK, 40% in Europe, and 30% in Japan (see, for instance, Beddington et al. (2013)). The TABB group estimates that annual aggregate profits from latency arbitrage currently exceed $21bn, Donefer (2008) provides a range of $15-25bn, and Strasbourg (2011) estimates that HFT profits in the US were around $7.2bn in 2009. Other observers believe profits to be smaller. Even if these estimates are in the right ballpark, it is unfortunately not known what fraction of the profits are due to cross-trading-venue arbitrage as opposed to within-venue market making, although an indirect indication points to large profits: Kearns et al. (2010) have estimated that had HFTs had perfect foresight, they could have reaped about $21bn of within-venue market making profits in US markets. Since HFTs do not have perfect foresight, actual within-venue profits are bound to be much smaller, and given the estimates of overall HFT profits, across-venue profits are likely to be sizable.

participants no longer trusted the price discovery mechanism.[4] This suggests that conventional measures of intermediation and liquidity provision may not adequately reflect market conditions when trading and liquidity are fragmented.

And yet the bulk of academic research on financial markets still relies on a price formation mechanism on a single centralized market, leaving regulators with no modeling tools to rigorously understand the impact of new policy initiatives designed to influence diverse trading platforms. In Europe, for instance, policy makers have indicated that models that explicitly account for fragmented markets in a general equilibrium setting are desperately needed to think through the MiFID 2 process that is currently under review and to engage in economic impact and market quality analysis. They have also indicated that it is not sufficient to merely account for the fact that the same stocks are traded across multiple trading venues, but that the model ought to be flexible enough to allow for stocks, derivatives on stocks, exchange-traded funds and related securities traded on distinct venues,[5] while providing guidance on how market quality is affected by fragmentation and the resulting linkages created by latency arbitrageurs. This paper can be viewed as a first step in that direction.

We propose a simple model that explicitly allows for multiple assets traded in multiple markets or venues[6] that are linked by profit-motivated arbitrageurs or intermediaries. We specifically focus on cross-venue market making and abstract from within-venue market making. We model intermediaries as imperfectly competitive with entry into the intermediation business unrestricted but entailing a fixed cost (say in terms of human capital, software, or co-location of servers at the various mar-

[4]Consider for example the E-Mini index futures contract traded on CME Globex and the SPY exchange-traded fund traded on NYSE, both of which track the S&P500. During the Flash Crash, trading in the E-Mini was paused for 5 seconds while trading in SPY continued. Uncertainties about pricing accuracy, exacerbated by the uncoordinated introduction of circuit breakers, led many arbitrageurs to cease operating their cross-market strategies. For four minutes, very profitable arbitrage mispricings occurred (for details see CFTC and SEC (2010) and Hunsader (2010)). For a detailed analysis of financial stability in computer-based trading environments, the reader is referred to Chapter 4 of Beddington et al. (2013).

[5]As an example, consider the SPY exchange-traded fund that we alluded to in footnote 4. SPY enters into a no-arbitrage relationship with the portfolio of equities underlying the S&P500 index. In addition, there are over 2000 options on SPY. Each such option needs to satisfy no-arbitrage relationships not only with SPY, but also with all sorts of combinations of other options on SPY. Furthermore, SPY options are traded on six options exchanges simultaneously, adding another layer of law-of-one price relationships. Finally, options on SPY are closely related to options on the S&P500 itself as well as to options on the S&P500 futures contract.

[6]While the "venue" metaphor is a helpful one and fits some situations such as latency arbitrage in which the same or similar securities are traded simultaneously on multiple trading venues, it is equally natural to think of the segmentation as being functional rather than geographical, e.g. in terms of investors restricted to certain asset classes (on-the-run versus off-the-run bonds, stock index arbitrage, equities versus derivatives on those equities, investment grade versus junk bonds etc.). A trading venue can also be interpreted as an over-the-counter (OTC) market in which an intermediary trades with a clientele; the intermediary then tries to offload the exposure from this OTC trade either with offsetting OTC counterparties or in the organized markets.

ket centers). Just as in the real world where latency arbitrageurs hit limit order bids and asks by market orders (in the US by using intermarket sweep orders to bypass the Order Protection Rule), the intermediaries in our model use market orders to hit the net excess demand schedules left by the marginal investors on each trading venue. In equilibrium, gains from trade are intermediated and local valuations and liquidities are aggregated across trading venues through an arbitrage network.

This framework allows us to address common questions about the fragmented structure of modern financial markets. If cross-venue arbitraging is competitive, are equilibrium allocations and prices identical to those that would obtain in an economy with a single centralized and perfectly competitive venue? What is the effect of barriers to entry into the intermediation sector? What are the relationships between volumes, liquidity and welfare across trading venues? How is a liquidity shock affecting one venue transmitted through the entire network connected by such intermediaries? Whatever the reputation of cross-venue arbitrageurs, they would seem to provide a valuable service and given the general equilibrium setting of our model, we can answer welfare questions in a straightforward manner. Is the liquidity offered by latency arbitrageurs welfare-improving, and if yes, how can it be measured? How can the overall welfare be disaggregated into the contributions of individual securities? If intermediaries can design and trade securities to extract maximum profits, what is the effect on welfare?

We define market quality as the welfare gains achieved in equilibrium through the trading of securities via intermediaries. These gains are reflected in state prices across trading venues, before and after intermediation, and can be quantified as the additional consumption enjoyed by investors as a result of the intermediation. Market quality can thus be viewed as intermediated liquidity that channels gains from trade across multiple trading venues. Intermediaries provide liquidity in the very direct sense of being the counterparties to trades made possible by their diverse customer base that reaches across various clienteles or market centers.

Trading liquidity is often regarded as a salient feature of well-functioning security markets. Traditional liquidity metrics such as depth (the market impact of a trade), breadth (the size of bid-ask spreads), volume, transaction costs, as well as timeliness and ease of execution of trades can be viewed as *symptoms* or *attributes* of an appropriate provision of liquidity that exploits gains from trade. Unfortunately, such measures are rarely welfare-based, not least because they view assets one by one and ignore interdependencies across assets and markets. A particular asset may not be liquid in those metrics, but substitutes may be liquid enough to compensate for it in a way that the underlying payoff is liquid. Looking at the liquidity of one security or on one platform is therefore unlikely to reveal the whole story and a *global* metric is needed, such as the one proposed in this paper. We characterize some of the relationships between our metric and the conventional measures or attributes. While relying on the attributes themselves is no doubt useful for market practitioners, thinking of them as sufficient proxies for market quality or overall welfare can be misleading.

The main contributions of our paper are the following.

First, and relative to conventional measures of liquidity, our metric is not only *welfare-based*, but it also has the advantage that it can be *aggregated and disaggregated easily*, across securities as well as trading venues. For instance, it is not obvious how one can infer the overall quality of a market from the spreads or volumes of individual securities. Usually this involves picking a few assets that are deemed representative of the market as a whole. Furthermore, since identical assets, or more generally payoffs, may not exist on multiple trading venues, one would need to compare substitute assets. Both points raise a Pandora's box of judgmental issues which can be avoided entirely by using a metric built on state prices instead.

Second, we derive useful *proxies* of our market quality metric that can in principle be empirically estimated. One such proxy is equilibrium volume per unit of depth. Neither high volume nor depth are necessarily desirable attributes of a financial market, for if a market is deep and yet attracts little volume, it does not serve a useful role. Our measure of market quality can also be deduced directly from the costs of entry into the intermediation sector, for such costs determine the degree of competition between intermediaries and therefore the gains from trade realized.

Third, our model provides a coherent framework for understanding recent market phenomena. For instance, much has been made of the design of unwise complex securities. We evaluate the impact on market quality of equilibrium *asset innovation* by intermediaries designed to extract the maximum surplus from investors (as is the case for example with many categories of OTC derivatives). We find that such innovation enhances overall market quality, mainly because intermediaries find it optimal in equilibrium to offer what investors desire – and are willing to pay for – most, though market quality in some sectors of the economy may be adversely affected.

Our model also provides a framework within which one can understand the logic of *securitization*. The boom in collateralized debt obligations (CDOs) was made possible not only by the low interest rate environment, but also by the arbitrage profits reaped by CDO structurers due to the difference between the price paid for debt, and the monies raised by selling tranches of that debt tailored to the needs of individual clienteles. Our framework offers a rationale for the CDO mechanism. Quite naturally, it also illustrates the dangers inherent in such a mechanism: should the demand for one of the tranches wane, this local liquidity shock ripples through all the tranches.

Fourth, our model lends itself directly to the study of the *transmission of liquidity shocks* from one sector of the economy to other sectors through cross-market intermediation. Over and above the direct transmission through the network, which is a function of the tightness of integration and the degree of complementarity of trading needs across the various market segments, we find a feedback effect through which a detrimental liquidity shock lowers the number of intermediaries, which in turn lowers liquidity and so on. An example of such a "liquidity spiral" can be seen in the demise of Lehman Brothers. Triggered by a liquidity shock originating in the US

housing sector, the exit of Lehman in turn led to a further deterioration of liquidity, forcing other intermediaries to curtail their operations. We also illustrate contagion through a natural experiment that occurred on the London Stock Exchange when a server outage resulted in a suspension of trading, with consequent knock-on effects on alternative trading venues. Finally, we provide an example of macro contagion caused by the bursting of the Japanese bubble in the 1990s.

The paper is organized as follows. In the next section we introduce our definition of market quality and outline some of its general properties, including the relationship of our metric to standard depth and spread measures. In Section 3 we describe and characterize our notion of equilibrium. In Section 4 we elaborate on the role played by intermediaries in the provision of liquidity. In the next few sections we relate our measure to the market quality of individual assets, and to depth, volume, and welfare. In Section 8 we allow intermediaries to introduce new securities and analyze the impact on market quality. In Section 9 we show how our setup can be used to study contagion. Illustrations of contagion in equity and CDO markets follow in Section 10. Section 11 is devoted to a review of the literature. Section 12 concludes. Proofs are collected in the Appendix.

2 Market Quality as Intermediated Gains from Trade

We formalize the notion of market quality in a two-period economy in which assets are traded at date 0 and pay off at date 1.[7] Uncertainty is parametrized by a finite state space $S := \{1, \ldots, S\}$.[8] Assets are traded on several "venues," the set of venues being given by $K := \{1, \ldots, K\}$. There are J^k assets available to agents on venue k, with the random payoff of a typical asset j denoted by d^k_j. Asset payoffs on venue k can then be summarized by the random payoff vector $d^k := (d^k_1, \ldots, d^k_{J^k})$. Our framework can easily handle heterogeneity of agents both within and across venues, but in order to focus on cross-market arbitraging we assume that there is no within-venue heterogeneity. It is convenient then to think of a single (representative) agent on venue k, and refer to this agent as agent or investor or clientele k. Trading between venues is intermediated by arbitrageurs.

We will describe the characteristics of investors and arbitrageurs in the next section. At this stage we motivate our market quality metric and describe some of its general properties that do not depend on the particular way in which equilibrium prices are determined. Our measure of market quality involves a comparison of state-price deflators. Given a collection of J assets with random payoffs $d := (d_1, \ldots, d_J)$

[7] This need not be interpreted literally. In the case of latency arbitrage, for example, the aim of the HFTs is to start and end each day holding no risky positions and only limited capital. Their strategy does not involve holding inventories overnight with the explicit aim of hedging intertemporal investment opportunities. Hence a repeated one-shot game is a factually satisfactory approximation of their behavior.

[8] Following standard convention, we use the same symbol to denote a set and its cardinality.

and prices $q := (q_1, \ldots, q_J)$, a random variable p is called a state-price deflator if $q_j = E[d_j p]$ for every asset j, or more compactly, $q = E[dp]$. Let $\hat{q}^k$ be an equilibrium asset price vector on venue k, and $\hat{p}^k$ a corresponding state-price deflator. Similarly, let $\mathring{q}^k$ be the asset price vector on venue k in autarky (these are prices at which agent k chooses not to trade), and p^k an associated state-price deflator.

Consider first the benchmark case of complete markets in which all the Arrow securities are traded on venue k. The additional date 0 consumption available to investor k, as a result of trading the Arrow security corresponding to state s, is given by $\theta_s^k(\mathring{q}_s^k - \hat{q}_s^k)$, where θ_s^k is the amount of the security bought by k. In terms of state-price deflators $p_s^k := \mathring{q}_s^k/\pi_s$ and $\hat{p}_s^k := \hat{q}_s^k/\pi_s$, where π_s is the probability of state s, this measure of gains from trade can be written as $\pi_s\theta_s^k(p_s^k - \hat{p}_s^k)$. Assuming risk aversion, the marginal valuation of consumption in state s is decreasing in the amount of this consumption, so as a first-order approximation we can say that $\hat{p}_s^k = p_s^k - \beta^k\theta_s^k$, for some $\beta^k > 0$ (this linear relationship holds exactly in a CAPM economy, as assumed below). Solving for the equilibrium demand, we get $\theta_s^k = \frac{1}{\beta^k}(p_s^k - \hat{p}_s^k)$. Thus the realized gains from trade in Arrow security s are $\frac{1}{\beta^k}\pi_s(p_s^k - \hat{p}_s^k)^2$. Aggregating over all Arrow securities gives us a measure of gains from trade on venue k:

$$\tilde{\mathcal{Q}}^k := \frac{1}{\beta^k}E[(p^k - \hat{p}^k)^2].$$

This definition is unambiguous if markets are complete. If markets are incomplete, however, there are multiple state-price deflators consistent with the same asset prices and payoffs. Consider the set of marketable payoffs for the assets $d = (d_1, \ldots, d_J)$, given by $M := \{z : z = d \cdot \theta, \text{ for some portfolio } \theta \in \mathbb{R}^J\}$. For an arbitrary random variable z, let z_M denote the least-squares projection of z on M. If markets are incomplete, there are many state-price deflators p that price the payoffs in M identically, i.e. for which $E[zp]$ is the same for any given z in M. However, there is a unique state-price deflator that lies in M. This *traded* state-price deflator is p_M, the least-squares projection on M of any of the deflators p (see Proposition 2.1 below). The metric $\tilde{\mathcal{Q}}$ can therefore be extended to the incomplete-markets case as follows:

$$\mathcal{Q}^k := \frac{1}{\beta^k}E[(p_{M^k}^k - \hat{p}_{M^k}^k)^2],$$

where M^k is the marketed subspace for venue k. Aggregating over all venues gives us a measure of overall market quality:

$$\mathcal{Q} := \sum_{k \in K} \mathcal{Q}^k.$$

We defined market quality in the complete-markets case as the additional consumption enjoyed by investors as a result of intermediation. We will show later (in Section 5) that this pecuniary interpretation carries over to our general definition of market quality given by $\mathcal{Q}$.

The term $E[(p^k_{M^k} - \hat{p}^k_{M^k})^2]$ in the definition of market quality is the mean-square distance between agent k's (traded) valuation $p^k_{M^k}$ and the equilibrium (traded) valuation of venue k, $\hat{p}^k_{M^k}$. This has the interpretation of gains from trade reaped by agent k, constrained by the assets available to him. More generally, we can rely on the work of Chen and Knez (1995) on market integration to provide a characterization of mean-square distance between state-price deflators:

Proposition 2.1 *Given random variables p and p', and a marketed subspace M for some collection of assets, we have:*

1. *$p_M = p'_M$ if and only if $E[zp] = E[zp']$, for all payoffs $z \in M$. In particular, $E[zp] = E[zp_M]$, for all $z \in M$.*

2. $$E[(p_M - p'_M)^2] = \max_{z \in M : E[z^2]=1} \left[E(zp) - E(zp')\right]^2$$

   i.e. $E[(p_M - p'_M)^2]$ is the maximal squared pricing error induced by p and p' among marketed payoffs z with $E[z^2] = 1$.

3. $$E[(p_M - p'_M)^2] = \max_{z : E[z^2_M]=1} \left[E(zp_M) - E(zp'_M)\right]^2$$

   i.e. $E[(p_M - p'_M)^2]$ is the maximal squared pricing error induced by p_M and p'_M among payoffs z with $E[z^2_M] = 1$.

The first statement says that two random variables are valid state-price deflators for a given collection of assets if and only if their marketed components are the same; moreover, this common marketed component is itself a state-price deflator. Thus our market quality measure does not depend on which state-price representation is chosen (i.e. p^k could be any autarky state-price deflator, and $\hat{p}^k$ any equilibrium state-price deflator, for venue k). The last two statements characterize the mean-square distance between the traded state-price deflators p_M and p'_M as a bound on the difference in asset valuations implied by them. More precisely, it is the maximal squared pricing error using p and p' to price (normalized) payoffs in M, or alternatively it is the maximal squared pricing error using the traded state-price deflators themselves to price all (normalized) payoffs, whether marketed or not.

Our market quality metric can be thought of as a measure of intermediated liquidity. It has been usual in the literature on liquidity, especially in applied work, to focus on depth and spreads. While we will be more precise later on the relationship of depth in particular to our measure of market quality, a few general remarks are in order.

First, a small trading impact or a small spread means that trades that have not (yet) transpired would not be costly to execute, but it says nothing about the cost of trades that have already occurred in equilibrium. An additional marginal trade

may be illiquid while most infra-marginal trades may in fact have been executed at tight spreads and little market impact. Our measure amalgamates the liquidity benefiting all equilibrium trades, rather than the liquidity posted for the marginal trade. Second, with multiple assets, there are as many ways to impact markets as there are portfolios that can be perturbed. Not all perturbations are economically useful. For instance, a small additional trade in a security that leads to a change in the intertemporal marginal rate of substitution that is uncorrelated with the payoff of the security being perturbed will have zero market impact and reflect a very deep market, although nobody desires or trades that economically irrelevant security. Third, spreads have been analyzed by picking a few assets and then arguing that the spreads in these assets are representative of the economy as a whole. In our framework, on the other hand, price discrepancies are measured in terms of the distances between state-price deflators. The advantage of such a measure is that it considers willingness to pay directly, rather than indirectly through proxies computed from a limited number of securities. It follows from Proposition 2.1 that the mean-square distance between the traded state-price deflators on two venues on which the same assets are traded is equal to the bound on the squared pricing errors in using these state-price deflators to price any payoff. In other words, it represents exactly what one is looking for when computing price differentials, and has the virtue of using all available information.

It is easy to see that the level of mispricing, e.g. the size of bid-ask spreads of individual securities, need not have any relationship to our measure of market quality or indeed to any welfare-based notion of liquidity. Consider, for the sake of illustration, an asset with payoff z, $E[z] = 0$, that is traded on two venues, 1 and 2. These "venues" need not be distinct market centers; we can simply interpret the venue with the higher valuation of the asset as the "buy side," and the other venue as the "sell side." The mispricing or bid-ask spread of this asset, given by $|E[(\hat{p}^1 - \hat{p}^2)z]|$, may be very low. For instance, it is zero if the covariance between z and $\hat{p}^1 - \hat{p}^2$ is zero. Yet market quality may also be very low, for instance if there are no intermediaries or if the potential gains from trade are insignificant. And the same applies to the converse: market quality may be relatively high and yet the bid-ask spread for some asset may be large. In other words, the bid-ask spread for one particular asset may not provide a reliable indication of how well markets are performing their reallocative function. All information impounded into the pricing relationships and gathered from the equilibrium actions of all agents needs to be taken into consideration, as is the case when using state prices.

In summary, market quality or intermediated liquidity as we see it is a general snapshot spread, properly aggregated across all payoffs and all market segments. An apparent drawback of our definition is that it involves terms such as autarky state-price deflators, which are hard to estimate. In the next few sections, we provide several characterizations of our metric in terms of variables that are in principle observable, such as the number of intermediaries and the cost of intermediation.

3 Equilibrium

The definition of market quality proposed in this paper does not crucially depend on any particular choice of timing, agent characteristics or market structure, and is therefore of universal application. However, in order to derive closed-form solutions and to relate market quality to traditional liquidity metrics and to welfare, a modeling choice must be made.

A tractable framework is obtained by making assumptions that yield a local CAPM on each venue, as follows. Investor $k \in K$ has date 0 endowment ω_0^k, and date 1 (random) endowment ω^k. He has quadratic preferences:

$$U^k(x_0^k, x^k) = x_0^k + E\left[x^k - \frac{\beta^k}{2}(x^k)^2\right],$$

where β^k is a positive parameter, x_0^k is date 0 consumption, and x^k is date 1 consumption. In addition, there are N arbitrageurs (with the set of arbitrageurs also denoted by N) who possess the trading technology which allows them to trade across venues, or in other words, which allows them to act as intermediaries if they so wish. Arbitrageurs care only about date 0 consumption and are imperfectly competitive.

Investors behave competitively and can trade only on their own venue. Thus all trades between investors are intermediated by arbitrageurs. Arbitrageurs have no endowments, so they can be interpreted as pure intermediaries.

The interaction between price-taking investors and strategic arbitrageurs involves a Nash equilibrium concept with a Walrasian fringe. Let $y^{k,n}$ be the supply of assets on venue k by arbitrageur n, and $y^k := \sum_{n\in N} y^{k,n}$ the aggregate arbitrageur supply on venue k. For given y^k, $q^k(y^k)$ is the market-clearing asset price vector on venue k, with the asset demand of investor k denoted by $\theta^k(q^k)$.

Definition *Given an asset structure $\{d^k\}_{k\in K}$, a Cournot-Walras equilibrium (CWE) of the economy is an array of asset price functions, asset demand functions, and arbitrageur supplies, $\{q^k : \mathbb{R}^{J^k} \to \mathbb{R}^{J^k}, \theta^k : \mathbb{R}^{J^k} \to \mathbb{R}^{J^k}, y^{k,n} \in \mathbb{R}^{J^k}\}_{k\in K, n\in N}$, such that*

1. *Investor optimization: For given q^k, $\theta^k(q^k)$ solves*

$$\max_{\theta^k\in\mathbb{R}^{J^k}} \; x_0^k + E\Big[x^k - \frac{\beta^k}{2}(x^k)^2\Big]$$

subject to the budget constraints:

$$x_0^k = \omega_0^k - q^k \cdot \theta^k$$
$$x^k = \omega^k + d^k \cdot \theta^k.$$

2. *Arbitrageur optimization: For given $\{q^k(y^k), \{y^{k,n'}\}_{n'\neq n}\}_{k\in K}$, $y^{k,n}$ solves*

$$\max_{y^{k,n}\in\mathbb{R}^{J^k}} \sum_{k\in K} y^{k,n} \cdot q^k\Big(y^{k,n} + \sum_{n'\neq n} y^{k,n'}\Big)$$

subject to the no-default constraint:

$$\sum_{k\in K} d^k \cdot y^{k,n} \leq 0.$$

3. *Market clearing:* $\left\{q^k(y^k)\right\}_{k\in K}$ *solves*

$$\theta^k(q^k(y^k)) = y^k, \qquad \forall k \in K.$$

A complete characterization of the CWE can be found in Rahi and Zigrand (2009, 2013). In the remainder of this section, we provide a brief synopsis of the relevant results. We refer the reader to the original papers for more details, including proofs.

Let $p^k := 1 - \beta^k \omega^k$, which we assume to be non-negative. This is consistent with our usage of p^k in Section 2, as it can be shown that p^k is an autarky state-price deflator for venue k. Indeed, for given arbitrageur supply y^k,

$$q^k(y^k) = E\left[d^k[p^k - \beta^k(d^k \cdot y^k)]\right]. \tag{1}$$

Thus $p^k - \beta^k(d^k \cdot y^k)$ is a state-price deflator for venue k. The autarky state-price deflator p^k is obtained by setting $y^k = 0$. Asset prices in autarky are given by $\mathring{q}^k := q^k(0) = E[d^k p^k]$.

Proposition 3.1 (Cournot-Walras equilibrium: Rahi and Zigrand (2009)**)** *There is a unique CWE.*[9]

1. *Equilibrium arbitrageur supplies are given by*

$$d^k \cdot y^{k,n} = \frac{1}{(1+N)\beta^k}\left(p^k_{M^k} - p^A_{M^k}\right), \qquad k \in K, \tag{2}$$

   where $p^A \geq 0$ *is a state-price deflator for the arbitrageurs.*

2. *Equilibrium asset prices on venue* k *are given by* $\hat{q}^k = E[d^k \hat{p}^k]$*, where*

$$\hat{p}^k := \frac{1}{1+N}p^k + \frac{N}{1+N}p^A. \tag{3}$$

   Thus $\hat{p}^k$ *is an equilibrium state-price deflator for venue* k*.*

3. *Aggregate arbitrageur profits originating from venue* k *are given by*

$$\Phi^k := \hat{q}^k \cdot y^k = \frac{N}{(1+N)^2\beta^k} E[(p^k_{M^k} - p^A_{M^k})^2]. \tag{4}$$

[9]Unlike Rahi and Zigrand (2009), here we denote equilibrium asset prices on venue k by $\hat{q}^k$ instead of q^k.

4. The equilibrium demands of investors are given by

$$d^k \cdot \theta^k = \frac{1}{\beta^k}(p^k_{M^k} - \hat{p}^k_{M^k}), \qquad k \in K. \tag{5}$$

5. The equilibrium utilities of investors are given by

$$U^k = \bar{U}^k + \frac{1}{2}\beta^k E[(d^k \cdot \theta^k)^2], \qquad k \in K, \tag{6}$$

where $\bar{U}^k$ is a constant that does not depend on the asset structure or investor portfolios.

The random variable p^A is a state-price deflator for the arbitrageurs in the sense that $p^A(s)$ is the arbitrageurs' marginal shadow value of consumption in state s.[10] Note that p^A can be chosen so that it does not depend on N.

Given the centrality of the arbitrageur valuation p^A, it is important to provide an explicit characterization of it. To this end, we define a Walrasian equilibrium with restricted consumption as an equilibrium in which agents can trade any asset on a centralized venue, facing a common state-price deflator p^{RC}, but agent k can consume claims in M^k only.[11] There are no arbitrageurs.

Proposition 3.2 (Arbitrageur valuations: Rahi and Zigrand (2013) **)**
Arbitrageur valuations in the CWE coincide with valuations in the Walrasian equilibrium with restricted consumption, i.e. $p^A_{M^k} = p^{RC}_{M^k}$, for all k. Consequently $\lim_{N\to\infty} \hat{q}^k = E[d^k p^{RC}]$.

Thus asset prices in the arbitraged economy converge to asset prices in the restricted-consumption Walrasian equilibrium, as the number of arbitrageurs goes to infinity (note that this is an immediate consequence of (3), once it is established that $p^A_{M^k} = p^{RC}_{M^k}$).[12]

We obtain a sharper characterization of p^A under some restrictions on the asset structure $\{d^k\}_{k\in K}$. Let p^* denote the complete-markets Walrasian state-price deflator of the entire integrated economy with no participation constraints. It can be shown that

$$p^* = \sum_{k\in K} \lambda^k p^k,$$

[10] More concretely, the algorithms used by latency arbitrageurs are known to revolve around the concept of a "micro price" that corresponds to the "true price" as perceived by the latency arbitrageur, prompting the algorithm to buy if the actual price on a venue is below this value and to sell if it is above, as in Equation (2).

[11] In other words, each investor can arbitrage all markets, but must then purchase a final consumption stream in the span of his local assets. See Rahi and Zigrand (2013) for a formal definition, and also for a discussion of the subtle difference between this notion of equilibrium and Walrasian equilibrium with restricted participation. In the latter, agents face a common state-price deflator, but agent k can trade claims in M^k only.

[12] The equilibrium allocation (for investors) in the arbitraged economy also converges to the restricted-consumption Walrasian equilibrium allocation.

where

$$\lambda^k := \frac{\frac{1}{\beta^k}}{\sum_{j=1}^{K} \frac{1}{\beta^j}}, \qquad k \in K.$$

The state-price deflator p^* reflects the autarky valuation of each venue in proportion to its depth. Now consider the following spanning condition:

(S) Either (a) $M^k = M$, $k \in K$, or (b) $p^k - p^* \in M^k$, $k \in K$.

Under **S**(a) we have a standard incomplete-markets economy in which all investors trade the same payoffs, though on different venues. **S**(b) is the condition that characterizes an equilibrium security design (see Section 8). We have the following analogue of Proposition 3.2:

Proposition 3.3 (Arbitrageur valuations II: Rahi and Zigrand (2009) **)** *Suppose condition* **S** *holds. Then, arbitrageur valuations in the CWE coincide with valuations in the complete-markets Walrasian equilibrium, i.e.* $p^A_{M^k} = p^*_{M^k}$*, for all* k*. Consequently* $\lim_{N\to\infty} \hat{q}^k = E[d^k p^*]$.

4 Intermediation and Market Quality

Now that we are armed with a model and a closed-form solution of the unique equilibrium, we can explicitly characterize the properties of the market quality measure defined in Section 2.

So how does intermediation create liquidity? Intermediation does not affect the spans $\{M^k\}_{k\in K}$, as there is no asset with a new dimension of spanning that becomes available due to pure intermediation.[13] What is achieved through intermediation is that the existing assets can be used more fruitfully. Thanks to intermediation, investors can trade on better terms. Suppose for example there are two venues, 1 and 2, with the same asset structure. Suppose there is an asset with payoff z for which the autarky price on venue 1, $\mathring{q}^1 = E[zp^1]$, is lower than the autarky price on venue 2, $\mathring{q}^2 = E[zp^2]$. Investor 1 wants to short the asset and sell it to investor 2 who wants to go long. By Proposition 3.3, we can choose $p^A = p^*$, which is a convex combination of p^1 and p^2. Hence the arbitrageurs' valuation of this asset, $q^A := E[zp^A]$, lies between p^1 and p^2. In the intermediated equilibrium, q^1 is pushed up and q^2 is pulled down (due to (3), $\hat{p}^k$ is closer to p^A than is p^k, for both venues). Intermediaries allow investor 1 to sell on better terms, while investor 2 can buy on better terms, with the spread narrowing. The welfare of both investors increases even though intermediaries take home some profits.

Notice that the market quality metric for venue k is scaled by $1/\beta^k$. From (1), it is clear that β^k is the price impact of a unit of arbitrageur trading on venue k: the state s value of the state-price deflator $p^k - \beta^k(d^k \cdot y^k)$ falls by β^k for a unit

[13] The case where intermediaries can issue assets to optimally intermediate is studied in Section 8.

increase in arbitrageur supply of s-contingent consumption. Later we show that β^k also measures the impact on the price of any asset on venue k of an additional unit of the asset supplied to that venue (see equation (11) and the ensuing discussion). Thus $1/\beta^k$ is the depth of venue k.

The equilibrium arbitrageur supply, given by (2), is very intuitive. Assuming for the moment that markets are complete on all venues, an arbitrageur supplies state s consumption to those venues which value it more than he does ($p_s^k - p_s^A > 0$). How much he supplies to venue k depends on the size of the mispricing $|p_s^k - p_s^A|$, on the depth $1/\beta^k$, with more consumption supplied the deeper the venue, and finally on the degree of competition N. If markets are incomplete, however, the difference between state prices may not be marketable. The arbitrageur would then supply state-contingent consumption as close to $p^k - p^A$ as permissible by the available assets d^k. The closest such choice is the projection $(p^k - p^A)_{M^k} = p^k_{M^k} - p^A_{M^k}$. The greater the number of arbitrageurs competing for the given opportunities, the smaller is each arbitrageur's residual demand, and so the less each one supplies. In the limiting equilibrium, as N goes to infinity, arbitrageurs virtually disappear in that individual arbitrageur trades vanish (but not their aggregate trades), as does their aggregate consumption, $\sum_k \Phi^k$. Ultimately they perform the reallocative job of the Walrasian auctioneer at no cost to society (as formalized in Proposition 3.2).

Another way to see this is to compare realized and potential gains from trade. Since arbitrageur valuations are Walrasian (Proposition 3.2), we can define the potentially achievable or maximal gains from trade as

$$\bar{\mathcal{Q}} := \sum_{k \in K} \bar{\mathcal{Q}}^k, \tag{7}$$

where

$$\bar{\mathcal{Q}}^k := \frac{1}{\beta^k} E[(p^k_{M^k} - p^A_{M^k})^2]. \tag{8}$$

$\bar{\mathcal{Q}}$ measures the gains from trade that can be reaped if the economy moves from autarky to a perfectly intermediated Walrasian equilibrium, with the asset spans remaining unchanged. $\bar{\mathcal{Q}}^k$ measures the total gains from trade between k and the rest of the economy. These gains ultimately arise from differences in preferences and endowments. In this sense, one can interpret date 0 as the time when investors learn about their preferences and endowments, i.e. about their idiosyncratic "liquidity shocks."

Proposition 4.1 (Competition and market quality)

$$\mathcal{Q}^k = \left(\frac{N}{1+N}\right)^2 \bar{\mathcal{Q}}^k, \qquad k \in K. \tag{9}$$

In particular, local market quality $\mathcal{Q}^k$ is strictly increasing in N, $\mathcal{Q}^k = 0$ at $N = 0$, and $\lim_{N\to\infty} \mathcal{Q}^k = \bar{\mathcal{Q}}^k$. Consequently, overall market quality $\mathcal{Q}$ is increasing in N, $\mathcal{Q} = 0$ at $N = 0$, and $\lim_{N\to\infty} \mathcal{Q} = \bar{\mathcal{Q}}$.

This result follows from the fact that $p^k - \hat{p}^k = \frac{N}{1+N}(p^k - p^A)$, due to (3). The expression (9) shows how our market quality measure captures the general costs of trading due to the noncompetitive nature of the intermediation. More competition improves upon the extent of gains from trade realized in the markets. In the limit, as competition becomes perfect, the liquidity offered by intermediaries is sufficient for all potential gains from trade to be exploited.

One of the advantages of our setup is that it is straightforward to endogenize the number of intermediaries as a function of the cost of entry into the intermediation business. While there are a number of related concepts of entry, the following is simple and sensible. Suppose each arbitrageur must bear a fixed cost c in order to set up shop and intermediate across all markets. First we determine the number of arbitrageurs N', not necessarily a natural number, so that each one of the N' arbitrageurs makes a net profit of zero after having borne the fixed cost. Using (4), (7) and (8), N' solves

$$c = \frac{1}{N'} \sum_k \Phi^k(N') = \frac{\bar{\mathcal{Q}}}{(1+N')^2}. \tag{10}$$

Second, this number is rounded down to the nearest natural number:

Proposition 4.2 *The equilibrium level of intermediation is given by*

$$N(c) = \mathrm{rd}\left(\sqrt{c^{-1}\bar{\mathcal{Q}}} - 1\right), \qquad c \leq \frac{\bar{\mathcal{Q}}}{4}.$$

The operator "rd" rounds the real number in parenthesis down to the next natural number. In particular, arbitrageurs make profits in equilibrium, but not enough to attract one further arbitrageur. We must have $c \leq \bar{\mathcal{Q}}/4$ in order for intermediation to arise (this will be a standing assumption for the rest of the paper). N increases as c falls, with $\lim_{c\to 0} N(c) = \infty$.

The assumption of unrestricted but costly entry provides us with a simple proxy for market quality. Using (9) and (10), and ignoring integer constraints on N, we get:

Proposition 4.3 *Market quality is decreasing in c and is given by*

$$\mathcal{Q} = cN(c)^2.$$

With estimates of c and N, an estimate of market quality is then simply the cost of entry times the square of the number of intermediaries, or equivalently the total cost borne by the intermediation sector times the number of intermediaries. Notice that even though depth is a crucial ingredient of market quality, it appears only insofar as it affects the endogenous number of intermediaries N. An added bonus is that N is a variable which can in principle be observed directly rather than having to be estimated.

Finally, it follows from Propositions 4.2 and 4.3 (again ignoring integer constraints) that

$$\mathcal{Q} = \left(\sqrt{\bar{\mathcal{Q}}} - \sqrt{c}\right)^2.$$

Market quality is increasing in the maximal amount of gains from trade allowed by preferences and securities, $\bar{\mathcal{Q}}$, and decreasing in the entry cost c. Lower entry costs mean more competition amongst arbitrageurs, which leads to improved terms of trade and improved quantities offered to investors, and consequently higher market quality.

5 Market Quality of Individual Assets

We have defined market quality or intermediated liquidity as the overall ease with which gains from trade can be exploited. In this section we deduce asset-by-asset market quality measures from the aggregate measure, and establish a compelling feature of our measure, namely additivity.

The first step is to identify the common factors that contribute to the market quality of different assets. The empirical findings of Chordia et al. (2000) that liquidity can be correlated between certain assets is not surprising from a theoretical point of view. The assets supplied in large amounts by arbitrageurs all share the characteristic of being valuable to investors, and these assets will see high volumes and liquidity. Assets that do not contribute towards the realization of gains from trade will not see active trading. Quite naturally in our setting, the common factors that underlie the market quality of individual assets are the portfolios mimicking the gains from trade, i.e. the portfolios whose payoffs are $p^k - \hat{p}^k$, k in K.

Recall that $\mathring{q}^k = E[d^k p^k]$ is the autarky asset price vector on venue k, and $\hat{q}^k = E[d^k \hat{p}^k]$ is the equilibrium asset price vector on k. We can formally disaggregate market quality $\mathcal{Q}^k$ into the diverse contributions of the J^k assets on venue k as follows:

Proposition 5.1

$$\mathcal{Q}^k = \frac{1}{\beta^k}\, b^k \cdot (\mathring{q}^k - \hat{q}^k),$$

where $b^k := \{b^k_j\}_{j \in J^k}$ *is the regression coefficient of the multiple regression of* $p^k - \hat{p}^k$ *on* d^k.

The coefficient b^k_j is the portion of the variation of the trading gains $p^k - \hat{p}^k$ on venue k that is explained by asset j. Accordingly, we define the local market quality of this asset on venue k as

$$\mathcal{Q}^k_j := \frac{1}{\beta^k}\, b^k_j (\mathring{q}^k_j - \hat{q}^k_j),$$

so that indeed

$$\mathcal{Q}^k = \sum_{j=1}^{J^k} \mathcal{Q}^k_j.$$

The market quality of asset j on venue k is equal to the depth of venue k times the usefulness of asset j in generating overall gains from trade on venue k, b_j^k, times the gains from trade directly reaped from trading asset j on venue k, $\mathring{q}_j^k - \hat{q}_j^k$. The term $\frac{1}{\beta^k} b_j^k$ is in fact equal to θ_j^k, the equilibrium holding of asset j on venue k (see Rahi and Zigrand (2009)). The local market quality of asset j can therefore be characterized as follows:

Proposition 5.2 (Local asset market quality)

$$\mathcal{Q}_j^k = \theta_j^k (\mathring{q}_j^k - \hat{q}_j^k),$$

i.e. the market quality of asset j on venue k equals the amount of date 0 consumption gained by investor k due to the more favorable equilibrium asset prices induced by intermediation.

Thus market quality has a purely pecuniary interpretation as the additional amount of consumption investors can enjoy due to more efficient pricing. Note that $\mathcal{Q}_j^k$ is positive. The equilibrium holding θ_j^k is equal to the arbitrageur supply y_j^k. From (11), we can see that the own-price effect of arbitrageur supply is negative. For example, if $y_j^k > 0$, then $\hat{q}_j^k < \mathring{q}_j^k$.

Finally, consider the case in which the same assets (or, more generally, payoffs) trade in all locations. Let $\mathcal{Q}_j := \sum_{k \in K} \mathcal{Q}_j^k$ be the global, or economy-wide, market quality of asset j.

Proposition 5.3 (Global asset market quality) *Suppose $d^k = d$, for all $k \in K$. Then*

$$\mathcal{Q}_j = N \Phi_j,$$

where $\Phi_j := \sum_k y_j^k \hat{q}_j^k$ is the aggregate arbitrageur profit in asset j.

Thus the global market quality of asset j is equal to the number of arbitrageurs times the total profits reaped by them in intermediating this asset. One might think that large arbitrage profits are indicative of an inefficient economy. But for large N, large aggregate profits mean small individual profits, and together they imply an economy that has achieved large efficiency gains relative to autarky. For instance, the sizable aggregate profits from latency arbitrage can be thought of as the result of many trades that serve to improve allocative efficiency.[14]

6 Depth and Volume

Depth, $1/\beta^k$, enters directly into the market quality measure $\mathcal{Q}^k$, as one would expect. It is constant, and in particular independent of arbitrage trades. This is a

[14] See footnote 3 for estimates of latency arbitrage profits. In applying the logic of our static model to such high-frequency trading activities, each round of which generates only very small profits, it should be understood that we have a repeated version of our model in mind.

very convenient feature of our model, for it allows us to show the endogenous nature of liquidity, even though depth is constant.

While depth is constant, the supply of an asset on venue k has a differential impact on the prices of other assets on k depending on the payoff structure d^k. From (1),

$$\frac{\partial q_j^k(y^k)}{\partial y_{j'}^k} = -\beta^k E[d_j^k d_{j'}^k]. \tag{11}$$

The price impact of one unit of trade in asset j' on venue k is more pronounced for those assets on k that are close substitutes in the sense of having a higher noncentral comovement with j'. For normalized payoffs z, with $E[z^2] = 1$, β^k measures the own-price effect.

Since arbitrageur supply is scaled by depth, there is a natural connection between depth and volume of trade. We define the volume originating from venue k as

$$\mathcal{V}^k := E[(d^k \cdot y^k)^2].$$

This is the overall equilibrium volume of trade in state-contingent consumption implied by intermediated asset trades on venue k. From (2),

$$\mathcal{V}^k = \left[\frac{N}{(1+N)\beta^k}\right]^2 E[(p_{M^k}^k - p_{M^k}^A)^2].$$

Using (8) and (9), we obtain the following result:

Proposition 6.1 (Market quality and volume) *Market quality equals volume per unit of depth:* $\mathcal{Q}^k = \beta^k \mathcal{V}^k$.

As one would expect, a welfare-based notion of market quality is associated not with the volume of asset transactions, but with the volume of the induced net trade in the underlying state-contingent consumption.[15] It is the latter that empirical researchers should try to measure when looking for a volume-based proxy for liquidity. Implicit in these trades are the motivations that gave rise to them as well as the microstructure considerations of asset spans and the degree of competition in the intermediation sector.

The relationship between volume and market quality highlighted in Proposition 6.1 is quite intuitive. For a given volume, more gains from trade are realized the closer state prices move towards Walrasian ones. State prices do not move very

[15] If there is a single asset on venue k, so that d^k is a scalar random variable, and we normalize the payoff so that $E[(d^k)^2] = 1$, then $\mathcal{V}^k = (y^k)^2$. With multiple assets, it would obviously not be sensible to compute the overall volume on a trading venue by simply summing up the volumes across the various securities traded on that venue, nor would a value-weighted volume metric capture the idea of quantity traded. In the complete-markets case, however, there is a straightforward connection of $\mathcal{V}^k$ to the volume of asset transactions. If markets are complete on venue k, with S linearly independent assets, and y_s^k is the volume of trade in the portfolio that replicates the Arrow security corresponding to state s, then $\mathcal{V}^k = E[(y^k)^2]$.

much in deep markets. Therefore volume needs to be large relative to depth to exploit the gains from trade, which market quality measures. Of course, volume is itself increasing in depth, and the net effect of depth on market quality is positive, indicating that the volume effect of depth dominates the direct depth effect.

7 Welfare

The equilibrium welfare of investors is given by (6). We measure economy-wide welfare by $U := \sum_{k \in K} U^k$. Using (5) and (6),

$$U^k = \bar{U}^k + \frac{1}{2}\mathcal{Q}^k$$

and

$$U = \bar{U} + \frac{1}{2}\mathcal{Q},$$

where $\bar{U} := \sum_{k \in K} \bar{U}^k$. Similarly, from (4), (8) and (9), total arbitrageur profits originating from venue k are

$$\begin{aligned} \Phi^k &= \frac{N}{(1+N)^2}\bar{\mathcal{Q}}^k \\ &= \frac{1}{N}\mathcal{Q}^k, \end{aligned}$$

so that aggregate economy-wide profits are

$$\sum_{k \in K} \Phi^k = \frac{1}{N}\mathcal{Q}.$$

This leads us to the following result:

Proposition 7.1 (Market quality, volume and welfare) *The following measures, local as well as global, are monotonically related: market quality, volume, investor welfare, arbitrageur profits, and social welfare.*

As we argued in the introduction, we feel that any measure of market quality would have to be tightly related to welfare in order to be economically meaningful. The above proposition confirms that this is indeed the case in our model.

8 Security Design

In this section we allow intermediaries to innovate and add assets to the ones already available for trade. We shall see that the optimally innovated assets not only augment intermediary profits, but also allow a better exploitation of gains from trade, leading to higher market quality, volume and welfare.

One might guess that any innovation would be welfare-improving. The reasoning might be as follows: since intermediaries can always choose not to trade the new assets, volumes, and therefore market quality, cannot be lower than in the absence of innovation. The reality is more complicated though, since market quality as defined here captures the extent to which markets allow the economy to move closer to the ideal Walrasian equilibrium for the given asset structure. Since an asset innovation perturbs the Walrasian equilibrium also (in particular the deflator p^A), it is not necessarily true that pricing at the new equilibrium is closer to the new Walrasian equilibrium than the old pricing was to the old Walrasian equilibrium. It turns out, however, that the aforementioned logic is correct if the innovations are optimal for arbitrageurs.

We have already seen in Section 3 that there is a unique CWE for any given asset structure $\{d^k\}_{k\in K}$. We now allow each arbitrageur to add assets to each venue before any trading takes place. This determines a new asset structure $\{d^k_{innov}\}_{k\in K}$. The payoffs of the arbitrageurs in this security design game are the profits they earn in the ensuing CWE.[16] Which asset(s) would arbitrageurs introduce at a Nash equilibrium of this game? Rahi and Zigrand (2009) show that there is a unique asset added to each venue (if not already present):

Proposition 8.1 (Optimal innovation: Rahi and Zigrand (2009)**)**
For a given $\{d^k\}_{k\in K}$, the asset structure

$$\begin{array}{cccc} [d^k & (p^k - p^*)] & \text{if} & p^k - p^* \notin M^k; \\ & d^k & \text{if} & p^k - p^* \in M^k; \end{array}$$

is

1. *a minimal optimal asset structure for arbitrageurs; and*
2. *a minimal Nash equilibrium of the security design game.*

The reader is referred to Rahi and Zigrand (2009) for a proof and a detailed discussion of this result. The term "minimal" refers to the fact that there are other optimal (or equilibrium) configurations, but involving more assets – all of these configurations have the property that $p^k - p^* \in M^k$, all $k \in K$. If there is an innovation cost, howsoever small, the chosen structure would unambiguously be a minimal one.

Since arbitrageur profits are higher in the post-innovation economy (condition 1 of Proposition 8.1), so is market quality due to the monotonic relationship between profits and market quality (Proposition 7.1)):

Proposition 8.2 (Innovation and market quality) *Market quality $\mathcal{Q}$ increases when intermediaries can innovate assets.*

[16] Note that all arbitrageurs are able to trade the assets introduced by any one arbitrageur. Also, due to the symmetry of the CWE (Proposition 3.1), all arbitrageurs have the same equilibrium payoff.

A clear distinction needs to be made between local and global market quality, however. While overall market quality improves with optimal innovation, even though the intermediaries act strategically, it is shown in Rahi and Zigrand (2009) that profits on any particular venue may fall. Invoking the monotonic relationship between *local* profits and market quality (Proposition 7.1), this means that innovation may hurt market quality on some venues. The intuition goes as follows. If innovation leads to lower volume on venue k due to decreased usefulness of trade, then market quality falls on k. This occurs for instance if venue k had an initial asset structure that permitted intermediaries to execute some crucial trades, say to borrow some state-contingent resources. When intermediaries can innovate optimally, they build such trades into the assets they innovate, thereby reducing the need to execute the trades on venue k.

9 Transmission of Liquidity Shocks

We now turn to the study of how liquidity shocks are transmitted across the economy. Starting from an initial equilibrium, we perturb fundamentals on one of the venues and analyze the economy-wide repercussions of this local shock. For simplicity, this is not a shock that could have been anticipated. In this regard we follow most of the literature on contagion.

In order to simplify the analysis, we shall assume that the spanning condition **S** holds, i.e. either the security design is optimal (as described in Proposition 8.1), or the same set of payoffs are tradable on all venues. Then we can choose $p^A = p^* = \sum_k \lambda^k p^k$ by Proposition 3.3.

We consider a local shock on venue ℓ. There are a number of ways to model this shock. The following turns out to be analytically tractable. Suppose there are I^ℓ investors on venue ℓ with identical preference parameters and endowments, $\{\bar{\beta}^\ell, \bar{\omega}^\ell\}$. Then the representative agent on ℓ has preference parameter $\beta^\ell = \bar{\beta}^\ell / I^\ell$ and endowment $\omega^\ell = I^\ell \bar{\omega}^\ell$. Consider a shock to the investor population (or participation) I^ℓ, while preserving individual investor characteristics. A withdrawal of participants on venue ℓ lowers its depth $1/\beta^\ell$ while keeping its autarky state-price deflator, $p^\ell = 1 - \beta^\ell \omega^\ell = 1 - \bar{\beta}^\ell \bar{\omega}^\ell$, constant. Consequently p^ℓ plays a less prominent role in p^A, but without making the economy more risk averse as would have happened had we simply lowered the depth of venue ℓ.

Let

$$\vartheta^{k\ell} := \frac{E[(p^k_{M^k} - p^A_{M^k})(p^\ell_{M^k} - p^A_{M^k})]}{E[(p^k_{M^k} - p^A_{M^k})^2]}.$$

Thus $\vartheta^{k\ell}$ is the regression coefficient of the (projected) mispricing on venue ℓ, $p^\ell_{M^k} - p^A_{M^k}$, on the mispricing on venue k, $p^k_{M^k} - p^A_{M^k}$. This measure of covariation is a noncentral "beta" in the language of the CAPM. Ignoring integer constraints on N, we have the following result:[17]

[17] This result requires the assumption that **S** holds in a neighborhood of I^ℓ, so that we can set

Proposition 9.1 (Contagion) *Suppose the spanning condition* **S** *holds. Then the effect on venue k of a population shock on venue ℓ is given by*

$$\frac{d\log \mathcal{Q}^k}{d\log I^\ell} = \underbrace{\mathbf{1}_{k=\ell} - 2\lambda^\ell \vartheta^{k\ell}}_{\frac{d\log \mathcal{Q}^k}{d\log I^\ell}\Big|_N} + \frac{\mathcal{Q}^\ell}{N\mathcal{Q}},$$

which is strictly decreasing in N.

The indicator function $\mathbf{1}_{k=\ell}$ takes the value 1 if $k = \ell$, and is zero otherwise. Effects can be split into two categories: direct effects for a given N, captured by the term $\mathbf{1}_{k=\ell} - 2\lambda^\ell \vartheta^{k\ell}$, and indirect effects via entry or exit which are represented by the term $\mathcal{Q}^\ell/(N\mathcal{Q})$. Notice that the first term does not depend on the initial level of N, while the second term is decreasing in N (since $\mathcal{Q}^\ell/\mathcal{Q} = \bar{\mathcal{Q}}^\ell/\bar{\mathcal{Q}}$, by Proposition 4.1, and hence does not depend on N).

Consider first a venue $k \neq \ell$, and suppose N is fixed. The effect on venue k's market quality (or intermediated liquidity) is $-2\lambda^\ell \vartheta^{k\ell}$. If the parameter $\vartheta^{k\ell}$ is negative, venues k and ℓ are *complements* in the sense that arbitrageurs tend to buy on one when they are selling to the other, i.e. there is intermediated trade between the two venues. If venue ℓ experiences a reduction in its investor base, and a consequent deterioration of its depth, these intermediated trades become less valuable and less plentiful in equilibrium, thus reducing liquidity on k.

With endogenous N, this effect is exacerbated: fewer investors and lower depth on ℓ lead to less trade and to lower liquidity, which in turn leads to lower profits and thereby to fewer intermediaries, which in turn affects liquidity adversely and so forth. It is this cascade of deteriorating liquidities that has received significant attention in the contagion literature. The net effect of this feedback loop is represented by the term $\mathcal{Q}^\ell/(N\mathcal{Q})$. The effect is more pronounced the larger the role of venue ℓ in generating trades, as measured by its relative size $\mathcal{Q}^\ell/\mathcal{Q}$, and the smaller the initial N. A smaller initial N means that the feedback loop of liquidity on N and again of N on liquidity etc. is stronger as each arbitrageur is more powerful and holds a larger portfolio.

If, on the other hand, $\vartheta^{k\ell} > 0$, valuations on venues k and ℓ are similar in the sense of being on average on the same side as the economy-wide valuation p^*. The two venues therefore compete for trades, and can be said to be *substitutes*. In this case, a shallower ℓ induces intermediaries to migrate to k, thereby increasing liquidity on k, for given N. The contagion effect operating through a lower N is however the same as in the case of complementary venues.

If we measure the degree to which markets are integrated by N, we see that contagion (in the sense of an adverse spillover) is more pronounced the more fragmented

p^A equal to p^* both before and after the shock. This is clearly not an issue if the same payoffs are traded on all venues (condition **S**(a)). However, if we invoke **S**(b), the result should be interpreted as the long-run effect of a population shock, allowing for optimal adjustment of the security design. While it is difficult to obtain an analytical result if we fix the (initially optimal) security design, numerical examples can be worked out, as we do in Section 10.2.

markets are. More precisely, the derivative in Proposition 9.1 is strictly decreasing in N, and is minimized as N goes to infinity and perfect integration is achieved. If k and ℓ are substitutes, this minimized value is negative; in this case the spillover of a negative shock is actually benign.

Now consider the effect of a population shock on venue ℓ on its own liquidity. For fixed N, this effect is given by $(1 - 2\lambda^\ell)$. If λ^ℓ is small, this has the straightforward interpretation of the direct loss of liquidity due to the flight of investors. This is compounded by the consequent flight of intermediaries in the same way as for the rest of the economy. If λ^ℓ is non-negligible, however, there is a countervailing effect. Indeed, if $\lambda^\ell > 1/2$, $\mathcal{Q}^\ell$ actually increases when the population on ℓ falls, for given N. This might at first appear odd, but the effect stems from the endogenous nature of Walrasian prices. Fewer investors on venue ℓ lower the depth of venue ℓ, and everything else constant, liquidity is lower. But the smaller size of this clientele also means that it will now play a less prominent role in the determination of the economy-wide valuation p^*. The valuation p^* will become more dissimilar from p^ℓ, thereby increasing the potential gains from trade between ℓ and the rest of the economy, stimulating intermediated trades and increasing liquidity on ℓ. If $\lambda^\ell > 1/2$, this effect is strong enough to compensate for the loss of depth, before accounting for the knock-on effect on the number of intermediaries.

Evidently, in an economy with many venues, loss of liquidity is more likely to go hand in hand with a decline in the number of active investors. But there might be situations where a dominant venue optimally limits or rations participants. It may be that the arrival of more (identical) investors can hurt local liquidity. The converse implication is that liquidity can suffer on a venue that experiences a rise in its investor population while substitute venues at the same time benefit from higher liquidity. These examples show that there is a clear externality in our economy that can go in either direction.

For $k \neq \ell$, assuming that $\lambda^\ell < 1/2$, it is easy to verify that

$$\frac{d \log \mathcal{Q}^k}{d \log \mathcal{Q}^\ell} > 1 \qquad \text{iff} \qquad 2\lambda^\ell(1 - \vartheta^{k\ell}) > 1 \tag{12}$$

for the population-type shocks considered above. Thus, if ℓ is large in terms of relative depth, and k is sufficiently complementary with respect to ℓ, a liquidity shock on ℓ has an even bigger impact on k than on ℓ itself. This is an illustration of the dictum that "when Russia sneezes, Brazil catches a cold."

What is the effect on asset prices of a liquidity shock? It is instructive to consider the case where the same assets trade on all venues so that price comparisons are straightforward. Accordingly, we assume that $d^k = d$, all k. Then $q^* := E[dp^*]$ is the asset price vector implied by the hypothetical complete-markets state-price deflator for the entire integrated economy.

Proposition 9.2 *Suppose $d^k = d$, for all $k \in K$. Then*

$$\frac{\partial \hat{q}^k}{\partial I^\ell} = \frac{N}{1+N} \frac{\lambda^\ell}{I^\ell} (\mathring{q}^\ell - q^*), \qquad k \in K.$$

Thus, if venue ℓ in isolation values assets more highly than the economy as a whole ($\mathring{q}^\ell > q^*$), an adverse participation shock on ℓ depresses asset prices worldwide. This is because the tendency of venue ℓ to pull up asset prices, via intermediated trades, is reduced when its weight in the world economy is lower. Quite naturally, the effect is more pronounced the greater the degree of intermediation.

10 Examples of Contagious Illiquidity

In this section we illustrate how our framework can be used to understand the diffusion of liquidity shocks in a number of recent market events.

10.1 Trading Halt on the LSE

The UK FTSE stock market basically consists of the London Stock Exchange (LSE) as the main venue with around 60% of trading volume for FTSE-100 stocks, with BATS, Chi-X and Turquoise as the main MTFs.[18] Since these venues trade a large common set of securities, one could reasonably view them as being competing venues, or substitutes. On Thursday 26th of November 2009, the LSE halted trading at 10:33 due to a server error, placing all order books into auction mode until trading resumed at 14:00. If these venues were strong substitutes, our model would predict that a negative liquidity shock on the LSE would lead to higher liquidity on the MTFs. But the opposite happened. Liquidity dried up immediately on all the MTFs and recovered only on the dot at 14:00 (see Intelligent Financial Systems (2009)).

Our model suggests that these markets should instead be understood as complements, with arbitrageurs typically buying on one and selling on the other. By Proposition 9.1, an adverse shock to I^{LSE} has a negative impact on the liquidity of an MTF if and only if $2\lambda^{LSE}\vartheta^{LSE,MTF} < \frac{\mathcal{Q}^{LSE}}{N\mathcal{Q}}$. So all trading venues that are either weak enough substitutes or complements of the LSE would have their liquidity negatively affected by a liquidity shock to the LSE. In fact, (12) tells us that the impact on an MTF would be more pronounced than on the LSE itself if $2\lambda^{LSE}(1-\vartheta^{LSE,MTF}) > 1$ (we can safely assume that $\lambda^{LSE} < 1/2$). This condition is more likely to be satisfied the larger the relative weight of the LSE in pricing the true value of stocks, and the greater the degree of complementarity. It would be an interesting empirical exercise to estimate these numbers.

10.2 CDO Boom and Bust

Consider the CDO mechanism. The profit to intermediaries from structuring and marketing CDOs ultimately stems from the fact that the tranched cash flows can be sold for more than the procurement cost of the cash flows from credit, such as loans and mortgages.

[18] BATS acquired Chi-X in 2011. They were separate entities at the time of the trading halt on the LSE.

For simplicity, in the following example there are four clienteles. Venue $k = 4$ represents the clientele from which the credit originates, modeled as a single security with payoff d^4. Suppose there are three states of the world, and the promised cash flows from credit are 3. Due to default, however, the effective cash flows are $d^4 = (3, 2, 1)$, where we write the random variable d^4 as a vector of state-contingent payoffs. In other words, in state $s = 1$ all loans are repaid, in state $s = 2$ two-thirds are repaid, and in state $s = 3$ only one-third are repaid. Intermediaries slice these cash flows into three tranches. The supersenior tranche is sold off to the highest bidders, here represented by investors of type $k = 1$. We assume that the supersenior tranche always pays off,[19] with $d^1 = (1, 1, 1)$. The mezzanine tranche, paying off $d^2 = (1, 1, 0)$, is sold to the highest bidding clientele, $k = 2$. Notice that the mezzanine tranche suffers a loss in state 3. Finally, the highest bidders for the junior tranche are investors on venue $k = 3$. The junior tranche only pays off in state $s = 1$ as it is the first to absorb any losses: $d^3 = (1, 0, 0)$. To summarize, the asset structure is:

$$d^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad d^2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad d^3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad d^4 = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}. \tag{13}$$

We construct an economy in which the equilibrium strategies of the arbitrageurs consist of buying the debt on venue 4, tranching it, and selling each tranche off to the clientele that values it most. We are interested in the transmission of liquidity shocks across this economy. In particular, based on current accounts of the subprime crisis, the relevant question is what the repercussions on overall liquidity are of a diminished clientele for the supersenior tranche.

To simplify our calculations, we assume that the three states are equally probable, and all investors have the same preference parameter $\beta^k = 1/4$. Furthermore, we assume that venues 2, 3 and 4 have the same population, which we normalize to one (i.e. $I^2 = I^3 = I^4 = 1$). We denote the population on venue 1 by I (i.e. $I^1 = I$). We shall reduce I to reflect investor flight from the supersenior CDO tranche. Date 1 endowments are as follows:

$$\omega^1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \omega^2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \omega^3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \omega^4 = \begin{bmatrix} 4 \\ 3 \\ 2 \end{bmatrix}.$$

The corresponding autarky state-price deflators, given by $p^k = 1 - \beta^k \omega^k$, are:

$$p^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad p^2 = \begin{bmatrix} 1 \\ 1 \\ \frac{3}{4} \end{bmatrix}, \quad p^3 = \begin{bmatrix} 1 \\ \frac{3}{4} \\ \frac{3}{4} \end{bmatrix}, \quad p^4 = \begin{bmatrix} 0 \\ \frac{1}{4} \\ \frac{1}{2} \end{bmatrix}.$$

Thus clientele 1 has the highest willingness to purchase the supersenior payoff d^1. Likewise, clienteles 2 and 3 are the highest bidders for the mezzanine and junior tranches, d^2 and d^3, respectively.

[19] This is irrelevant for our results. With more states, superseniors can default as well.

To understand the rationale for the CDO structure, consider first the benchmark case in which $I = 1$. Then the complete-markets Walrasian state-price deflator for the integrated economy, p^*, is 3/4 in all three states. It is easy to check that the asset structure (13) is the optimal security design, i.e. tranching is optimal for arbitrageurs. For every unit of d^4 that arbitrageurs buy, they sell one unit each of the tranches d^1, d^2 and d^3. The arbitrageurs' valuation p^A is equal to p^*.

Compare this, for instance, to the case in which a pass-through security is sold to all investors. Then the asset structure is $(3, 2, 1)$ on all venues. The arbitrageurs' valuation is the same as above and equal to p^*. For every unit that arbitrageurs buy on venue 4, they sell 6/14, 5/14 and 3/14 units on venues 1, 2 and 3, respectively. Maximal market quality, $\bar{\mathcal{Q}}^k$, is unchanged for venue 4 but is lower for the other venues. The equilibrium level of intermediation is lower as well, leading to lower market quality, liquidity and welfare on all four venues.

While the CDO structure is optimal for $I = 1$, it is not so for other values of I. In particular, we are interested in what happens if appetite for the supersenior tranche diminishes, given this CDO structure. For $I \neq 1$, the spanning property **S** fails, which means that we cannot use the convenient condition $p^A = p^*$. The following can be verified to be a Lagrange multiplier for the arbitrageurs' first-order conditions, and therefore a valid state-price deflator:

$$p^A = \frac{3}{17I+3}\begin{bmatrix} 4I+1 \\ 4I+1 \\ 9I-4 \end{bmatrix},$$

provided $I \geq 4/9$, which we will henceforth assume.[20] Equilibrium arbitrageur supplies are:

$$y^{1,n} = y^{2,n} = y^{3,n} = -y^{4,n} = \frac{1}{1+N} \cdot \frac{20I}{17I+3}.$$

Thus the pattern of trade is the same as in the benchmark case of $I = 1$. These trades are simply scaled down as I falls. Notice that arbitrageur trades are exactly offsetting, so that $\sum_k y^{k,n} d^k = 0$. Equilibrium asset prices are given by:

$$\begin{array}{ll} \hat{q}^1 = 1 - \frac{N}{1+N}\frac{5}{17I+3}, & \hat{q}^2 = \frac{2}{3} - \frac{N}{1+N}\frac{10I}{3(17I+3)}, \\ \hat{q}^3 = \frac{1}{3} - \frac{N}{1+N}\frac{5I}{3(17I+3)}, & \hat{q}^4 = \frac{1}{3} + \frac{N}{1+N}\frac{70I}{3(17I+3)}. \end{array}$$

Maximal economy-wide market quality is

$$\bar{\mathcal{Q}} = \frac{100I}{3(17I+3)}.$$

As I falls, so does $\bar{\mathcal{Q}}$. This means that, even for fixed N, overall market quality or liquidity $\mathcal{Q}$, which is given by $(\frac{N}{1+N})^2\bar{\mathcal{Q}}$, falls. In fact, it can be verified that the

[20] The results are less clear-cut when I falls below 4/9. This is because there are not enough investors to absorb consumption in state 3, so it ends up in the hands of the arbitrageurs. Then our assumption that arbitrageurs only care about consumption at date 0, which is fairly innocuous as long as the asset structure does not deviate too far from one that satisfies **S**, starts to matter.

same is true for the liquidity of tranches 2 and 3, and the liquidity of the underlying debt. Moreover, as I falls, intermediaries start going out of business, with N given by $\mathrm{rd}(\sqrt{c^{-1}\overline{\mathcal{L}}} - 1)$. This exacerbates the drying up of liquidity.

Figures 1 and 2 illustrate the effects on liquidity and intermediation of a change in I, both above and below 1, for $c = .001$. That there is contagion is evident: as

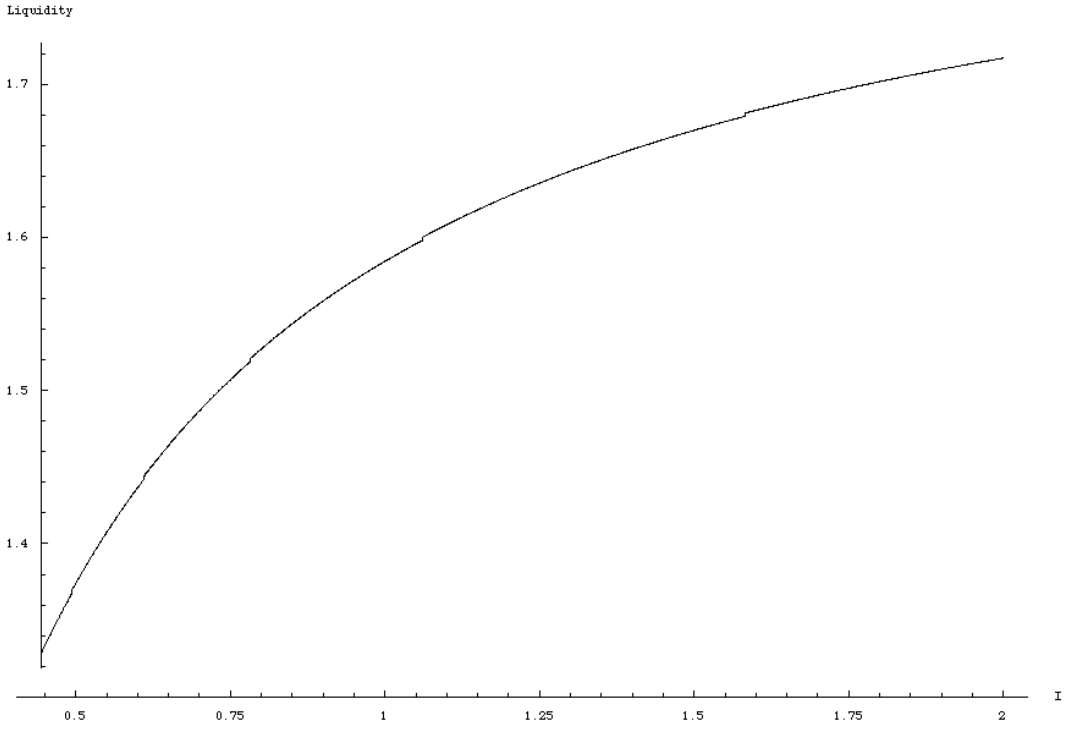

Figure 1: Overall liquidity, $\mathcal{Q}$, as a function of I

the natural clientele for the supersenior tranche is eroded, the entire CDO market seizes up. A 50% decline in the size of this clientele (starting from $I = 1$) causes overall liquidity to decline by more than 13%. This effect aggregates the impact of a change in I on relative depths, on shadow prices p^A, as well as on N. The plots for the liquidity of tranches 2 and 3, and for the liquidity of the securitized debt, are similar to that for overall liquidity.

During the boom phase, before doubts about the creditworthiness of CDOs and related products became prevalent, demand for tranches was in part fueled by the quest for yield in a low interest rate environment. In our model, the CDO mechanism leads to lower prices of the various tranches than would have obtained in its absence (i.e. $\hat{q}^k < \mathring{q}^k, k = 1, 2, 3$). In other words, the CDOs allow the credit and money markets to deliver higher yields. Likewise, the CDO mechanism allows debtors to borrow at a more attractive rate ($\hat{q}^4 > \mathring{q}^4$).

Everything else constant, higher demand for the supersenior tranche leads to

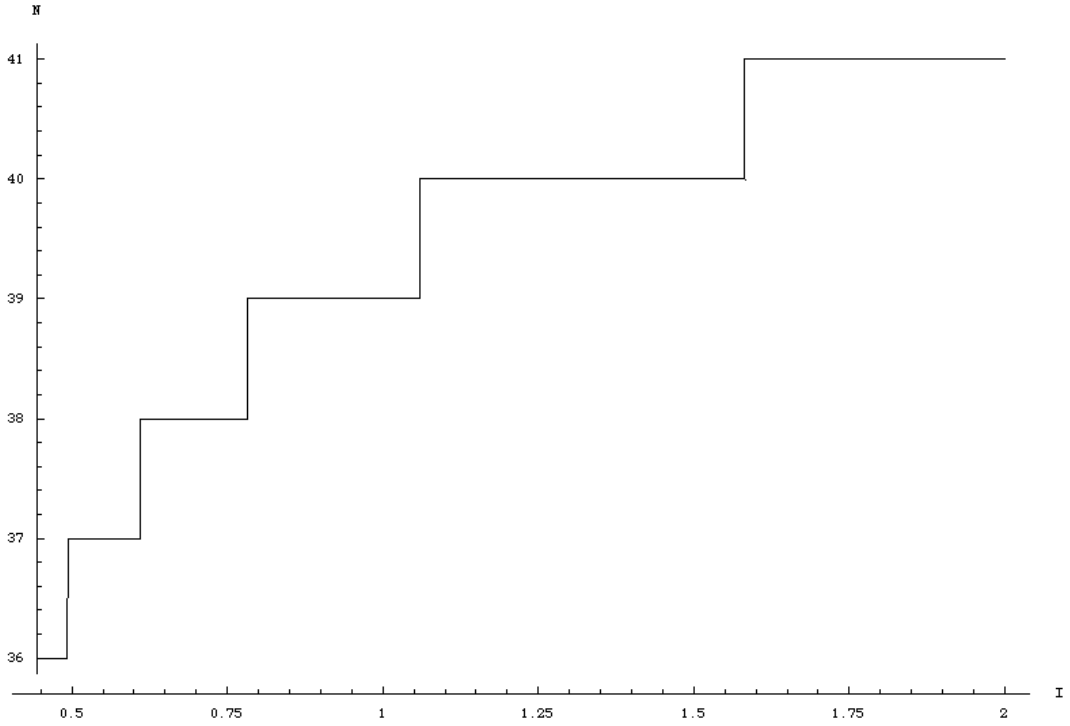

Figure 2: Equilibrium Number of Arbitrageurs, N, as a function of I

higher supersenior prices,[21] as well as higher prices for the underlying securitized debt. Concurrently, prices for the other tranches fall – and yields rise – since these investors find more counterparties for their trades. And if on the contrary demand for the supersenior tranche wanes, these effects are reversed: prices for tranches 2 and 3 rise and the corresponding yields fall as arbitrageurs are forced to reduce their shorts and buy back those tranches.

The crisis events unfolding in the credit markets from Summer 2007 onwards cannot be fully captured by this simple version of our model. Contrary to our assumptions here, banks in the real world did have their own capital and used it to keep the supersenior tranches when they found no buyers for them. They went on structuring CDOs and selling the remaining lower graded tranches off, pocketing the "arbitrage" profits (they were arbitrage trades for the structuring desks, who sold the supersenior tranches to the treasury department of the same organization, but not for the intermediary as a whole). This overextension into CDOs then became plain when an "unexpected" state was realized wherein the supersenior tranches were no longer perceived to pay back their face value. More elaborate versions of our model can be constructed to allow for arbitrageur capital and for default, but this is beyond the scope of this paper.

[21] One can check that this is true in spite of the countervailing effect of higher N.

10.3 Japan-US in the Early 1990s

As a further illustration of contagion, this time of the macro type, consider the liquidity shock emanating from Japan at the end of the 1980s and beginning of the 1990s, as documented for example by Peek and Rosengren (1997). We can interpret this shock as a drop in the Japanese local investor base. While Japan was a major financial power, it is safe to assume that it did not account for more than half of the world's financial depth. Given that the flow of capital was from Japan to the US, Japan and the US were complements, and on average asset prices were higher in Japan than in the rest of the world. The adverse shock to Japanese liquidity depressed stock prices in Japan. The authors documented that the result of this liquidity shock was a sharp decline in Japanese investment in the US, which in turn adversely affected liquidity in the US, an instance of contagion along the lines suggested by our model (in particular, Proposition 9.2).

11 Relationship to the Literature

Despite the recent interest in liquidity fragmentation, the increasing complexity of structured products exploiting segmentation, the growth of latency arbitrage by HFTs, and the need for a rigorous analysis of the forces of market integration for regulatory purposes, academic research on these subjects is still in its infancy. While there is a vast literature studying market liquidity directly or indirectly, market quality in general remains a bit nebulous. We are not aware of any papers that define market quality or liquidity via an explicit metric that itself has a clear welfare meaning, or that relate this definition to the different attributes of liquidity, such as depth, bid-ask spreads or volume.

Traditionally, liquidity has been studied empirically in single-asset models (see, for example, the papers cited in Chordia et al. (2000)), with little attention given to multi-asset liquidity, common factors, liquidity substitutes and so forth. Recently, however, a few papers have started to address this omission, among them Chordia et al. (2000), Hasbrouck and Seppi (2001) and Korajczyk and Sadka (2008). Similarly, the effect of multiple trading venues on liquidity has not been studied extensively. While many papers compare liquidity metrics such as bid-ask spreads on an ECN with those on an exchange, with both being analyzed in isolation, fewer study the effects on liquidity of the interaction between ECNs and exchanges. In the latter camp are Hendershott and Mendelson (2000), Weston (2002), Foucault and Menkveld (2008) and Biais et al. (2010), who find that in general the growth of electronic competitors has had a positive impact on bid-ask spreads in the underlying markets. None of these papers explicitly studies the effects of cross-venue trades, however, with the exception of Karolyi et al. (2012) who argue that during periods of market stress, commonalities appear that are due to the crisis-induced trades by cross-market arbitrageurs, a theme that we also explore in this paper.

There is a growing empirical literature in support of segmentation and clientele

pricing in asset markets; see Rahi and Zigrand (2009) for a discussion of this literature. That there are arbitrage opportunities across markets, often fleeting, is well-known. Most such opportunities occur between less well-researched or tailored securities, between related derivatives, or between derivatives and replicating strategies of varying degrees of sophistication. But even bids and asks for liquid Nasdaq stocks cross occasionally across trading venues (i.e. the bid on one venue is higher than the ask on another).[22] Garvey and Murphy (2006) for instance show that in their sample, consisting of the 10 most traded stocks on Nasdaq in addition to the 10 stocks with the largest market caps, crosses occur about .5% of the time.

A key ingredient of our theoretical model is the close connection of liquidity to arbitrage activity. This connection has been documented empirically by Hu et al. (2012) in the market for US Treasuries. They find that during market crises, shortage of arbitrage capital allows yields to move more freely relative to the curve, resulting in more "noise" in prices and hence in "less liquid" markets.

In terms of theoretical work, not much has been done on multi-asset liquidity. Gromb and Vayanos (2009) study the provision of liquidity by arbitrageurs in a segmented markets setting, with arbitrage opportunities arising across pairs of assets traded on different market segments, and a continuum of competitive arbitrageurs who face a separate margin constraint in each asset. Their main concern is the effect on liquidity (defined as depth) of this financial constraint, which is the source of the limits to arbitrage in their model. In our paper, on the other hand, arbitrage is limited due to a cost of entering the arbitraging business, and imperfect competition among arbitrageurs. Our framework also allows for an arbitrary asset structure across trading venues. Brunnermeier and Pedersen (2009) model liquidity needs as arising from the asynchronous arrival of investors. Like Gromb and Vayanos (2009), they study the link between the margin constraints faced by speculators and the liquidity they provide. Fernando (2003) models "liquidity shocks" as additive shocks that affect investors' marginal valuations of risky assets. His main interests are the price effects of idiosyncratic versus systematic liquidity shocks and the impact of liquidity shocks to one asset on prices of other assets. Cespa and Foucault (2012) model liquidity spillovers across assets as arising through an informational channel whereby a less liquid market for one security carries less information, which affects the information set of dealers in other markets and hence the liquidity provided by these dealers.

Our analysis builds on our earlier work in Rahi and Zigrand (2009, 2013). We use some results from these papers, in particular on the characterization of equilibrium.

[22] In 2005, 42% of trades in Nasdaq stocks occurred on Nasdaq with the remaining 58% occurring in non-linked market centers. Note that the observed crosses are the outcomes of segmentation left unarbitraged by intermediaries: absent arbitrageurs many more crosses would be observed.

12 Conclusion

In this paper we study financial market quality in a world in which trading is fragmented across multiple venues or platforms. We define an intuitive metric of market quality that captures the true economic meaning of the liquidity provided by intermediaries, namely the extent to which gains from trade are realized through intermediation. This metric can be expressed in terms of the real resources gained by investors as a result of this intermediation. It can be applied to one security traded on a single platform, but its strength lies in being able to aggregate liquidity across multiple securities and trading venues.

Standard liquidity metrics rely on a particular choice of both the metric, that is seldom explicitly welfare-grounded (such as bid-ask spreads or depth), and of a small number of securities, thereby ignoring substitutes as well as the overall market equilibrium. Our market quality metric does not suffer from such shortcomings and "sees through" the structure of markets to provide a measure that captures the magnitude of welfare-improving transactions realized in equilibrium. Our metric also has a clear relationship to volume, depth, intermediation costs and the like.

We explicitly model the role of profit-maximizing, liquidity-providing intermediaries who link markets together. The number of intermediaries is part of the equilibrium and is both influenced by market quality, and influences it in turn. The intermediation can be functional or geographic, and can encompass a variety of trading activities, from arbitrage between derivatives and the underlying markets to the huge industry of latency arbitrage across multiple lit and dark trading venues that has been in the regulatory limelight recently. We also allow intermediaries to introduce new securities, and we show that security design improves overall market quality, though local market quality may decline on some venues. The intermediaries form endogenous links across markets, and the strength of these linkages determines the transmission of liquidity shocks through the system.

We illustrate how in a number of recent market events the diffusion of liquidity shocks can be understood through the lens of interconnected markets provided by the setup in this paper. For instance, the advent of HFTs offering market making services across exchanges and MTFs suggests that some of these MTFs are complements rather than substitutes of the main venues in the sense that liquidity on one relies positively on liquidity on the others. The outage on the London Stock Exchange on the 26th of November 2009 rippled through the network of interconnected markets and brought liquidity across all related MTFs crashing down, even though the alternative venues were supposed to pick up the liquidity lost on the LSE. Our model suggests that most of the trades on the MTFs were arbitrage trades by HFTs between these venues and the LSE.

We show that the impact of a local shock on the size of the intermediation sector has a feedback multiplier effect on market quality – for instance, a negative liquidity shock forces some intermediaries to exit, thus reducing liquidity, inducing more intermediaries to exit, and so forth. We illustrate this with an example of conta-

gion in CDO markets, wherein a demand shock to one tranche reverberates through the entire system, impacting the liquidity of all the other tranches. By interpreting trading venues as countries, our setup can also shed light on cross-border investment flows following a shock in one country, as in the case of the bursting of the Japanese bubble and its effect on the US stock market.

Appendix

In the Appendix we adopt matrix notation in order to simplify the proofs. We represent asset payoffs d^k by the $S \times J^k$ matrix R^k whose j'th column lists the state-by-state payoffs of the j'th asset. The set of traded payoffs M^k is then the column space of R^k.

Let Π be the diagonal matrix whose diagonal elements are the probabilities of the states, $\pi_1, \ldots, \pi_S$. A state-price deflator for (q, R) is a vector $p \in \mathbb{R}^S$ such that $q = R^\top \Pi p$.[23] In other words, state-price deflators can be viewed as vectors instead of random variables. Similarly, the expectation $E[xy]$ can be written as $x^\top \Pi y$, where the random variables x and y are viewed as vectors in $\mathbb{R}^S$. In our finite-dimensional setting, the inner product space L^2 is the space $\mathbb{R}^S$ endowed with the inner product $\langle x, y \rangle_2 := x^\top \Pi y$. Then $x_{M^k} = P^k x$, where P^k is the orthogonal projection operator in L^2 onto M^k, given by the idempotent matrix $P^k := R^k({R^k}^\top \Pi R^k)^{-1}{R^k}^\top \Pi$. An explicit derivation of P^k can be found in Rahi and Zigrand (2009). P^k depends on R^k only through the span M^k. The L^2-norm of $x \in \mathbb{R}^S$ is $\|x\|_2 := (x^\top \Pi x)^{\frac{1}{2}}$.

In this notation, market quality on venue k is

$$\mathcal{Q}^k = \frac{1}{\beta^k} \|P^k(p^k - \hat{p}^k)\|_2^2.$$

Proof of Proposition 2.1 The proof is adapted from arguments in Chen and Knez (1995). Let the asset payoff matrix be R with marketed subspace M and corresponding projection matrix P.

The first statement says that $P(p - p') = 0$ if and only if $R^\top \Pi(p - p') = 0$. If $P(p-p') = 0$, then $R^\top \Pi P(p-p') = 0$. But $R^\top \Pi P = R^\top \Pi$, so that $R^\top \Pi(p-p') = 0$. Conversely, $R^\top \Pi(p - p') = 0$ implies that $(p - p') \in M^\perp$. Hence $P(p - p') = 0$.

Next we prove the third statement. Consider a payoff z, not necessarily in M. The mispricing of z using Pp versus Pp' is $m(z) := z^\top \Pi P(p-p')$. Since $\Pi P = P^\top \Pi P$, by the Cauchy-Schwartz inequality we have $m(z) \leq \|Pz\|_2 \|P(p-p')\|_2$; equality occurs if and only if Pz and $P(p - p')$ are collinear. It follows that

$$\|P(p - p')\|_2 = \max_{z:\|Pz\|_2=1} z^\top \Pi P(p - p'). \tag{14}$$

[23] The symbol $\top$ denotes "transpose." We adopt the convention of taking all vectors to be column vectors by default, unless transposed.

For the second statement, consider a payoff $z \in M$. Then $z = R\theta$ for some $\theta \in \mathbb{R}^J$. Using again the fact that $R^\top \Pi P = R^\top \Pi$, we see that $m(z) = z^\top \Pi(p - p')$. Hence, (14) can be written as

$$\|P(p - p')\|_2 = \max_{z \in M : \|z\|_2 = 1} z^\top \Pi(p - p').$$

■

Proof of Proposition 5.1

$$\begin{aligned}
\mathcal{Q}^k &= \frac{1}{\beta^k} \|P^k(p^k - \hat{p}^k)\|_2^2 \\
&= \frac{1}{\beta^k} (p^k - \hat{p}^k)^\top {P^k}^\top \Pi P^k (p^k - \hat{p}^k) \\
&= \frac{1}{\beta^k} \underbrace{(p^k - \hat{p}^k)^\top \Pi R^k ({R^k}^\top \Pi R^k)^{-1}}_{{b^k}^\top} \underbrace{{R^k}^\top \Pi (p^k - \hat{p}^k)}_{\mathring{q}^k - \hat{q}^k}.
\end{aligned}$$

■

Proof of Proposition 5.3 Using (3), we have

$$\hat{q}^k = \frac{1}{1+N} \mathring{q}^k + \frac{N}{1+N} q^*,$$

where $q^* = E[dp^*]$. It follows that $\mathring{q}^k - \hat{q}^k = N(\hat{q}^k - q^*)$. From Proposition 5.2, $\mathcal{Q}_j^k = N\theta_j^k(\hat{q}_j^k - q_j^*) = N y_j^k (\hat{q}_j^k - q_j^*)$. Since the same assets trade on all venues, arbitrageur holdings of any asset, aggregated across all venues, must be zero, i.e. $\sum_k y^k = 0$. Hence, $\mathcal{Q}_j = N \sum_k y_j^k \hat{q}_j^k$. ■

Proof of Proposition 9.1 In order to calculate the effect of a proportional change in I^ℓ, $d \log I^\ell$, it is convenient to write the population of venue ℓ as αI^ℓ, with corresponding depth α/β^ℓ, and compute derivatives with respect to α evaluated at $\alpha = 1$. Using (9), we can write $\mathcal{Q}^k$ as a function of α, p^A and N:

$$\begin{aligned}
\mathcal{Q}^k(\alpha, p^A(\alpha), N(\alpha)) &= \frac{\alpha^k}{\beta^k} \left(\frac{N}{1+N} \right)^2 \|P^k(p^k - p^A)\|_2^2 \\
&= \frac{\alpha^k}{\beta^k} \left(\frac{N}{1+N} \right)^2 (p^k - p^A)^\top \Pi P^k (p^k - p^A),
\end{aligned}$$

where $\alpha^k = \alpha$ for $k = \ell$, and $\alpha^k = 1$ for $k \neq \ell$. The total derivative of $\mathcal{Q}^k$ with respect to α is

$$\frac{d\mathcal{Q}^k}{d\alpha} = \underbrace{\frac{\partial \mathcal{Q}^k}{\partial \alpha} + \frac{\partial \mathcal{Q}^k}{\partial p^A} \cdot {p^A}'(\alpha)}_{\frac{d\mathcal{Q}^k}{d\alpha}\big|_N} + \frac{\partial \mathcal{Q}^k}{\partial N} N'(\alpha). \tag{15}$$

Noting that

$$p^A(\alpha) = \frac{\frac{\alpha}{\beta^\ell}p^\ell + \sum_{k\neq\ell}\frac{1}{\beta^k}p^k}{\frac{\alpha}{\beta^\ell} + \sum_{k\neq\ell}\frac{1}{\beta^k}},$$

we have

$$\begin{aligned}
\frac{\partial \mathcal{Q}^k}{\partial \alpha} &= \mathcal{Q}^\ell \mathbf{1}_{k=\ell} \\
\frac{\partial \mathcal{Q}^k}{\partial p^A} &= -\frac{2}{\beta^k}\left(\frac{N}{1+N}\right)^2 \Pi P^k(p^k - p^A) \\
p^{A\prime}(\alpha) &= \lambda^\ell(p^\ell - p^A) = \frac{\beta^k\lambda^k}{\beta^\ell}(p^\ell - p^A),
\end{aligned} \tag{16}$$

where all the derivatives are evaluated at $\alpha = 1$. Hence the effect on $\mathcal{Q}^k$ for given N is

$$\left.\frac{d\mathcal{Q}^k}{d\alpha}\right|_N = \mathcal{Q}^\ell \mathbf{1}_{k=\ell} - \frac{2\lambda^k}{\beta^\ell}\left(\frac{N}{1+N}\right)^2 \varphi^{k\ell}, \tag{17}$$

where

$$\varphi^{k\ell} := (p^k - p^A)^\top \Pi P^k (p^\ell - p^A).$$

We now solve for $N'(\alpha)$. With free entry, $\mathcal{Q} = cN^2$ (Proposition 4.3). Therefore, the function $N(\alpha)$ is defined by the identity

$$\sum_k \mathcal{Q}^k(\alpha, p^A(\alpha), N(\alpha)) - c[N(\alpha)]^2 \equiv 0.$$

Implicit differentiation gives us

$$N'(\alpha) = \frac{\sum_k \left.\frac{d\mathcal{Q}^k}{d\alpha}\right|_N}{2cN - \sum_k \frac{\partial \mathcal{Q}^k}{\partial N}}.$$

Now

$$\frac{\partial \mathcal{Q}^k}{\partial N} = \frac{2}{N(1+N)}\mathcal{Q}^k, \tag{18}$$

so that

$$\sum_k \frac{\partial \mathcal{Q}^k}{\partial N} = \frac{2}{N(1+N)}\mathcal{Q} = \frac{2cN}{1+N},$$

where we have once again used the result that $\mathcal{Q} = cN^2$.

Under the spanning condition **S**, either $P^k = P$ or $P^k(p^k - p^A) = p^k - p^A$. In both cases $\sum_k \lambda^k \varphi^{k\ell} = 0$, since $p^A = p^*$. Therefore, from (17),

$$\sum_k \left.\frac{d\mathcal{Q}^k}{dI^\ell}\right|_N = \mathcal{Q}^\ell.$$

Altogether, this yields

$$N'(\alpha) = \frac{(1+N)\mathcal{Q}^\ell}{2cN^2} = \frac{(1+N)\mathcal{Q}^\ell}{2\mathcal{Q}}. \tag{19}$$

Substituting (17), (18) and (19) into (15) gives us

$$\frac{d\mathcal{Q}^k}{d\alpha} = \mathcal{Q}^\ell \mathbf{1}_{k=\ell} - \frac{2\lambda^k}{\beta^\ell}\left(\frac{N}{1+N}\right)^2 \varphi^{k\ell} + \frac{\mathcal{Q}^k\mathcal{Q}^\ell}{N\mathcal{Q}}.$$

Dividing through by $\mathcal{Q}^k$ we get the desired result (note that $\vartheta^{k\ell} = \frac{\varphi^{k\ell}}{\varphi^{k,k}}$). ∎

Proof of Proposition 9.2 Let $R^k = R$, all k. Then, using (3),

$$\begin{aligned}\hat{q}^k &= R^\top \Pi \hat{p}^k \\ &= R^\top \Pi \left(\frac{1}{1+N}p^k + \frac{N}{1+N}p^A\right).\end{aligned}$$

Therefore, from (16),

$$\frac{\partial \hat{q}^k}{\partial d \log I^\ell} = \frac{N}{1+N}\lambda^\ell R^\top \Pi (p^\ell - p^A).$$

Moreover, since condition **S**(a) is satisfied, we have $p^A = p^*$. This yields the desired result. ∎

References

Beddington, J., Bond, P., Cliff, D., Houstoun, K., Linton, O., Goodhart, C., and Zigrand, J.-P. (2013). *The Future of Computer Trading in Financial Markets: An International Perspective*. Foresight.

Biais, B., Bisière, C., and Spatt, C. (2010). Imperfect competition in financial markets: An empirical study of Island and Nasdaq. *Management Science*, 56(12):2237–2250.

Brunnermeier, M. and Pedersen, L. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22(6):2201–2199.

Cespa, G. and Foucault, T. (2012). Illiquidity contagion and liquidity crashes. Working Paper.

CFTC and SEC (2010). Findings regarding the market events of May 6, 2010. Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues.

Chen, Z. and Knez, P. J. (1995). Measurement of market integration and arbitrage. *Review of Financial Studies*, 8(2):287–325.

Chordia, T., Roll, R., and Subrahmanyam, A. (2000). Commonality in liquidity. *Journal of Financial Economics*, 56(1):3–28.

Donefer, B. S. (2008). Risk management and electronic trading. Mimeo, FIX Protocol Conference.

Fernando, C. S. (2003). Commonality in liquidity: Transmission of liquidity shocks across investors and securities. *Journal of Financial Intermediation*, 12:233–254.

Foucault, T. and Menkveld, A. J. (2008). Competition for order flow and smart order routing systems. *Journal of Finance*, 63:119–158.

Garvey, R. and Murphy, A. (2006). Crossed markets: Arbitrage opportunities in Nasdaq stocks. *The Journal of Alternative Investments*, 9(2):46–58.

Gromb, D. and Vayanos, D. (2009). Financially constrained arbitrage and cross-market contagion. Mimeo, London School of Economics.

Hasbrouck, J. and Seppi, D. (2001). Common factors in prices, order flows and liquidity. *Journal of Financial Economics*, 59(3):383–411.

Hendershott, T. and Mendelson, H. (2000). Crossing networks and dealer markets: Competition and efficiency. *Journal of Finance*, 55:1–40.

Hu, X., Pan, J., and Wang, J. (2012). Noise as information for illiquidity. Working Paper.

Hunsader, E. (2010). Nanex flash crash summary report. Technical report, Nanex. `http://www.nanex.net/FlashCrashFinal/FlashCrashSummary_III.html`.

Intelligent Financial Systems (2009). What was the impact of the LSE outage on Thurs 26th Nov 2009? LiquidMetrix Short Article, `http://www.if5.com/LiquidMetrix/Articles/LM001`.

Karolyi, G. A., Lee, K.-H., and van Dijk, M. A. (2012). Understanding commonality in liquidity around the world. *Journal of Financial Economics*, 105(1):82–112.

Kearns, M., Kulesza, A., and Nevmyvaka, Y. (2010). Empirical limitations on high-frequency trading profitability. *Journal of Trading*, 5(4):50–62.

Korajczyk, R. and Sadka, R. (2008). Pricing the commonality across alternative measures of liquidity. *Journal of Financial Economics*, 87(1):45–72.

Peek, J. and Rosengren, E. S. (1997). The international transmission of financial shocks: The case of Japan. *American Economic Review*, 87(4):495–505.

Rahi, R. and Zigrand, J.-P. (2009). Strategic financial innovation in segmented markets. *Review of Financial Studies*, 22(8):2941–2971.

Rahi, R. and Zigrand, J.-P. (2013). Walrasian foundations for equilibria in segmented markets. Mimeo, London School of Economics.

Strasbourg, J. (2011). A wild ride to profits. *Wall Street Journal*. http://online.wsj.com/article/SB10001424053111904253204576510371408072058.html, accessed September 6, 2012.

Weston, J. (2002). Electronic communication networks and liquidity on the Nasdaq. *Journal of Financial Services Research*, 22(1&2):125–139.