ALGORITHMIC FINANCE

The topology of macro financial flows: An application of stochastic flow diagrams

Neil J. Calkin; Marcos López de Prado

Algorithmic Finance (2014), 3:1-2, 43-85
DOI: 10.3233/AF-140033

Abstract, HTML, and PDF:
http://algorithmicfinance.org/3-1-2/pp43-85

Subscription, submission, and other info:
www.iospress.nl/journal/algorithmic-finance

Algorithmic Finance 3 (2014) 43–85
DOI 10.3233/AF-140033
IOS Press

The topology of macro financial flows: An application of stochastic flow diagrams

Neil J. Calkin[a] and Marcos López de Prado[b,c,*]
[a]*Department of Mathematical Sciences, Clemson University, Clemson, SC, USA*
[b]*Senior Managing Director, Guggenheim Partners, New York, NY, USA*
[c]*Research Affiliate, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*

Abstract. A large portion of Macroeconomic and Financial research is built upon classical applications of Linear Algebra (such as regression analysis) and Stochastic Calculus (such as valuation models). As a result, most Macroeconomic and Financial research has inherited a focus on *geometric locations* rather than *logical relations*. Ideally, Econometric models could be complemented with Topological and Graph-Theoretical tools that recognize the hierarchy and relationships between system constituents.

Stochastic Flow Diagrams (SFDs) are topological representations of complex dynamic systems. We construct a network of financial instruments and show how SFDs allow researchers to monitor the flow of capital across the financial system. Because our approach is dynamic, it models how and for how long a financial shock propagates through the system. Practical applications include stress-testing of investment portfolios under user-defined scenarios, and the discovery of Macro trading opportunities. SFDs add Topology to the Econometric toolkit used by Macroeconomists, and may enlighten perennial controversies, such as the one involving Keynesians and Austrian-school economists. Our findings have important implications for regulators, market designers and Macro investors.

Keywords: Time series, graph theory, topology, financial flows, macro trading

AMS Classification: 90C35, 90B15, 05C21, 62P05, 37M10

> "*It is the aim of science to establish general rules which determine the reciprocal connection of objects and events in time and space. For these rules, or laws of nature, absolutely general validity is required –not proven.*"
>
> *Albert Einstein (1941)*

1. Introduction

Financial markets are price-discovery mechanisms, where investors can place their bets on future asset prices or hedge against price fluctuations. In a globalized economy, the same investors operate in multiple markets, allocating capital to equities, fixed income, commodities, currencies, etc. When investors alter their capital allocations, they generate financial flows, which disturb the pre-existing balance of supply and demand. For example, investors could reduce their exposure to bonds in order to increase their exposure to equities, generating a net flow towards the latter. In doing so, the valuations of bonds and stocks will

*Corresponding author: Marcos López de Prado, Guggenheim Partners, 330 Madison Ave., New York, NY 10017, USA. E-mail: marcos.lopezdeprado@guggenheimpartners.com; www.QuantResearch.info.

be impacted in ways that specialists in those markets will not fully understand. Specialists' myopic analysis of their markets gives an advantage to Macro fund managers, who study markets as a complex system.

A financial system can be defined as a set of interdependent markets forming an integrated whole. In a financial system, prices and quantities emanated from constituent markets cannot be understood individually. Prices are mutually interrelated, and a shock on one displaces the rest. Because of rigidities, adjustments may not be instantaneous, giving rise to lead-lag effects. When the financial system works properly, Central Banks can regulate the level of economic activity in the short run by setting official interest rates. This action propagates through the entire system, as a result of capital flows impacting each market. However, when capital does not flow through the system, such as during the capital freeze of 2008, Central Banks are forced to intervene in individual markets. Understanding the structure of capital flows is essential for Central Bankers to identify the tools that are most effective and least interventionist. The optimal tools will vary depending on the targeted asset class or sector of the economy.

One reason why Economics has been plagued with internal disputes, some of them almost a century old, may be the lack of tools to study Economic and Financial systems properly. Virtually all Macroeconomic and Financial research rests upon classical applications of Linear Algebra (such as regression analysis) and Stochastic Calculus (such as valuation models). Sala-i-Martin (1997) humorously criticized this over-reliance in regression analysis in his paper "*I Just Ran Two Million Regressions.*" Prof. Gilbert Strang sarcastically described those researchers relying solely on regression analysis as "least-squares-happy" statisticians (Strang, 2010):

> "*And that inverse is called the pseudo-inverse, and it's very, very, useful in application. Statisticians discovered, oh boy, this is the thing that we needed all our lives, and here it finally showed up, the pseudo-inverse is the right thing. Why do statisticians need it? [. . .] Because statisticians are like least-squares-happy. I mean they're always doing least squares.*"

This is a problem, because Linear Algebra was not devised to study complex dynamic systems. The first part of the Fundamental Theorem of Linear Algebra shows that a *mxn* matrix *A* produces a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, i.e. moving locations from the row space into the column space. The second part of that theorem tells us that if $Ax = b$ is not in the column space, since Ax cannot leave the column space, we can still choose the closest point to b in that subspace (Strang, 1993), an operation called projection or regression. Linear Algebra focuses on *locations* in Euclidian geometry, thus the "linear" term. Stochastic Calculus shares that focus, by concentrating on the measurement of changes of geometric location associated with a random variable, the speed of that change, etc. Macroeconomic and Financial research has inherited that focus on geometric locations. However, location is only one aspect relevant to the study of complex systems. While Linear Algebra can answer questions such as how closely together a system moves, it was not devised to recognize what is a system's shortest path of propagation, what nodes can shut down a network, or what connections are most critical to ensure network flow. In the year 1735, Euler proposed the "Seven Bridges of Königsberg" problem to evidence that Geometry could not answer basic questions regarding a system's organization. Euler showed that *logical relations* are more relevant to answer such questions, hence founding the subjects of Topology and Graph Theory.

> *Following Euler's argument, because Econometric and Statistical models rely on geometric arguments, they will not suffice to study Economic and Financial systems. The implication is, in order to better understand Economic and Financial systems, we must complement our Econometric models with Topological and Graph-Theoretical applications. Only the latter recognize the hierarchy and relationships between system constituents.*

A Topological study of the financial system provides a relationships map to how markets are arranged and interconnected. A Graph Theory study of the financial system tells us about the role that various nodes play. This is useful to market participants, because it explains how information propagates through the system, allowing them to monitor capital flows and anticipate price trends. It is useful to Central Bankers, because it helps them decide what monetary policy tools will be more effective, and simulate the secondary effects of their actions. Regulators and market designers can use the financial system's Topology to study

crashes, develop early warning systems and trigger ex-ante circuit breakers. For example, evidence suggests that the Flash Crash of May 6th 2010, which affected U.S. stock prices, was preceded by massive capital flows in the foreign exchange and rates markets, see Easley et al. (2011, 2012). Easley et al. (2013) present evidence of contagion effects across markets.

The main goal of this paper is to propose a framework that applies some of the analytical tools introduced in Calkin and López de Prado (2014). We believe that this framework can assist policy makers and investors in the systematic identification and analysis of financial flows. We are not suggesting that Statistical and Econometric applications are not useful, however we believe that Topology should be added to the Statistical and Econometric toolkit used to study systems. Econometricians need to add Euler's perspective to the tools that Gauss and Legendre gave us. In doing so, we will be following Laplace's advice: "*Lisez Euler, lisez Euler, c'est notre maître á tous.*" (Dunham, 1999)

The rest of the paper is organized as follows: Section 2 reviews the existing literature. Section 3 details our contributions. Section 4 discusses the estimation of SVAR specifications. Section 5 describes our framework. Section 6 presents a possible architecture for financial flows. This is not necessarily the only possible architecture, and our main interest is to illustrate the use of topological techniques rather than favoring a particular system architecture. Section 7 simulates the response of the system to a variety of shocks. Section 8 simulates the response of the system to structural changes. Section 9 summarizes our conclusions. The mathematical appendix explains in greater detail the arguments employed by the paper, and the Python code contains the implementation of some key aspects of our methodology.[1]

2. Review of existing literature

There is a growing literature of advanced papers presenting Topological and Graph Theory applications to computer science and operations research (for example, Bollobás (2013); Bondy and Murty (1976); Durrett (2007), etc.). However, as it relates to financial applications, the number of studies is still surprisingly scarce:

- Easley and Kleinberg (2010) and Jackson (2010) are two textbooks that explain how concepts from Graph Theory and Game Theory can be applied to answer important economic and financial questions. The first one in particular is an excellent source for getting introduced into this subject.
- Rebonato and Denev (2014) introduce a novel Bayesian network approach to coherent asset allocation. In the context of portfolio management, these authors model observations according to causation links, as opposed to mere association. This introduces a structure that allows the user to understand risk, as opposed to just measuring it. The ability to define scenarios, incorporate subjective views, model exceptional events, etc., in a rigorous manner is extremely satisfactory.
- Billio et al. (2012) perform an Econometric analysis that uses Principal Components Analysis (PCA) to estimate the degree of interconnectedness, and a Granger-causality test to determine its direction (Granger, 1969). Contemporaneous effects are not considered. The authors use OLS to fit a VAR(1) process on every pair of variables in a system of up to a hundred variables.
- Boginski et al. (2003, 2005, 2006): These authors build what they call a "market graph", which is a network representation of a universe of stocks. Each stock represents a node, and nodes are connected when their synchronous correlation coefficient exceeds a given threshold. They conduct the statistical analysis of this graph and show that the degree distribution follows the power-law model. They detect subsets of highly interconnected stocks and subsets of relatively disconnected stocks. They also suggest new data mining techniques for classifying stocks based on this market graph, which provide a deeper insight into the internal structure of the stock market.
- Lautier et al. (2011a, 2011b, 2012a, 2012b): These authors build graphs in order to investigate integration and systemic risk in derivative markets on multiple asset classes. Like Boginski et al. (2003, 2005, 2006), they also use synchronous correlation above a given threshold to determine whether an edge exists. Their key contribution is to apply Minimum Spanning Trees (MSTs) to reduce the

[1] All code in this paper is provided "as is", and contributed to the academic community for non-business purposes only, under a GNU-GPL license. Users explicitly renounce to any claim against the authors. The authors retain the commercial rights of any for-profit application of this software, which must be pre-authorized in written by the authors.

complexity of the network. They argue that MSTs represent the shortest and most probable path for the propagation of a price shock. After examining the topology of the MST, they study its evolution over time and its stability.

- *Stochastic Flow Diagrams* (SFDs) are topological representations of dynamic systems, introduced by Calkin and López de Prado (2014). SFDs use directed graphs (also called digraphs), where the direction of flows is determined by lead-lag specifications involving more than two variables. Because of these features, SFD captures the reverberation of a shock as it propagates through the system over time. We will discuss SFD at great length in this paper, and we refer the reader to the above publication for details.

3. Our contribution

The literature cited in Section 2 represents an excellent first step towards the application of Graph Theory to the modeling of the financial system. However, we believe that it can be improved in a number of ways.

Boginski et al. (2003, 2005, 2006) and Lautier et al. (2011a, 2011b, 2012a, 2012b) base their analysis on synchronous correlations. Correlations depend on characteristics of the joint distribution between two related markets. While it seems reasonable to assume that individual return processes are stationary, it is not always true that two returns processes will be jointly stationary. In absence of that stationarity, unconditional correlations do not exist. As Alexander (1999) points out:

> "*Of course it is always possible to calculate a number that supposedly represents correlation, but often these numbers change considerably from day to day, a sign that the two returns processes are not jointly stationary. It is unfortunate that some standard correlation estimation methods induce an apparent stability that is purely an artefact of the method, and the true nature of the underlying correlations is obscured [...] Although correlation has become the ubiquitous tool for measuring comovements in asset returns, its limitations are substantial.*"

Correlation is a poor measure of connectedness, among other reasons because it is a static measure. The literature on the serial dependence of financial returns is extensive (see Hamilton (1999); Campbell, Lo and MacKinlay (1996); Tsay (2010), to cite only a few). In forming a network, we are interested in dynamic behavior, such as lead-lag patterns or cause-effect relationships. Correlation does not imply causation and cannot capture lead-lag patterns. In particular:

- Correlations can be spuriously high in the absence of any relationship between variables (Aldrich, 1995).
- Correlations can be low even in the presence of a strong relationship. For instance, suppose two real-valued random variables $x_t = \alpha + \varepsilon_t$ and $y_t = \beta + \gamma x_{t-1}$, where $t \in \{1, \ldots, T\}$ is the index for the series of observations, ε_t is independent and identically distributed (IID) as a Standard Normal, and α, β, γ are real-valued parameters. The correlation between x and y is $\rho[x, y] = \frac{\sigma[x,y]}{\gamma\sigma^2[\varepsilon]} = 0$. Although x and y are deterministically linked (y_t is just an affine transformation of x_t), the synchronous correlation between x and y is null, because y_t is lagging x_t. Because information does not propagate instantaneously through the financial system, lags exist, which means that correlations will not properly measure the degree of connectedness between variables.

Billio et al. (2012) take into account lead-lag effects, since they apply a Granger-causality test (Granger, 1969) on each pair of variables. We believe this is a more appropriate approach to identify the connectedness within a dynamic system. However, some aspects could be improved: i) Searching for pairwise Granger-causality in a system of over a hundred variables may miss interactions in which two or more variables jointly determine another one. ii) Granger-causality misses contemporaneous effects, which are likely to play an important role in a highly interconnected system such as the global economy. iii) This econometric approach would benefit from the application of findings and techniques from Topology and Graph Theory.

The networks estimated by Boginski et al. (2003, 2005, 2006) and Lautier et al. (2011a, 2011b, 2012a, 2012b) were undirected graphs, where relations are defined in pairs. This was consistent with their choice of contemporaneous correlations as their measure of connection. However, undirected graphs dismiss all information regarding the direction of the capital flows. Our dynamic time series models determine what elements are leaders and what elements are laggers, and we will use that information to set the direction of

flows. Our first contribution is to apply SFDs to compute and study the structure and dynamics of Macro financial flows.

In a flow system, one component may receive flow from more than one other component. For example, suppose three components a, b and c, where a receives flow from b and c. Although all relationships in a graph are by definition established in pairs, their estimation should not be carried out in pairwise terms ($b \to a,\ c \to a$), or flow will not be additive. In a flow system, each equation determines the value of a variable as a function of multiple other variables (not only one-to-one). Our second contribution is to estimate SFD connections as the solution to a full system of dynamic equations in multivariate form. This approach differs from Billio et al. (2012) in important ways: i) We do not use PCA to determine the degree of connectedness, but to reduce the dimensionality of the analysis to logical subsets of variables, i.e. variables that researchers can still relate to. It is our opinion that this feature extraction procedure, which we denote *supervised PCA,* strikes the right balance between dimension reduction and meaningfulness of the variables employed. ii) Our SVAR(p) specification on n equations is more general than a VAR(1) model on two equations (see Section 5.3 for details). In particular, we are not constrained to one lag, we still allow for contemporaneous effects, and our specification accepts any number of variables, not only pairwise relations. iii) SFDs provide a richer architecture of a dynamic system than a Cartesian mapping. iv) SFD flows are additive, because variables are modelled to receive flow from multiple connections. v) SFD incorporates a time dimension, thus allowing us to estimate the time it takes for a shock to propagate through the financial system.

While SVAR(p) models on n equations are more general than the special case of VAR(1) models on two equations, they also incorporate the challenge of their estimation. The literature on estimation of systems with contemporaneous effects is one of the newest and fastest growing areas in statistical analysis. One problem with these specifications is that solutions are not guaranteed to exist, absent some strong constraints imposed on the model. Our third contribution is to propose an automatic estimation method of SVAR(p) that we call Combinatorial Recursive Estimation (CORE). CORE finds the optimal among the set of computationally feasible solutions, with a minimum number of constraints. However, we caution the reader about hitting every nail with this hammer. Like with other automatic estimation methods, the standard disclaimers apply. If such methods were the solution to every estimation problem, there would be no need for theory, which of course would constitute a preposterous position.

Lautier et al. (2011a, 2011b, 2012a, 2012b) define a distance measure between two elements x,y as $d_{x,y} = \sqrt{2(1 - \hat{\rho}[x, y])}$, where $d_{x,y} \in [0, 2]$. They establish a connection if the distance $d_{x,y}$ is below a given threshold $\bar{d}$, and construct a weighted graph where the weights are given by $d_{x,y} < \bar{d}$. One problem with this non-linear transformation is that it is not variance-stabilizing. The implication is that spurious connections are formed at various rates, depending on the sample length and the correlation threshold. Our fourth contribution will be to establish connections using variance-stabilizing transformations. We will establish a connection when the statistical significance of the fitted time series regressors exceeds a given acceptance threshold. The resulting SFD gives the most likely scenarios of flow propagation.

Our fifth contribution is to decouple the concepts of distance and weight. As explained in the previous paragraph, a connection is established if its underlying (time series) regressor is statistically significance. Its weight, however, is determined by the amount of flow it carries. Thus, a highly statistically significant connection may still carry a low amount of flow, which should not be given extraordinary weight.

Our sixth contribution is to explicitly incorporate the time dimensionality in the design of the graph representation, through the concept of *outbound arc*. Because our connectivity model is dynamic (applying time series models rather than correlation), transitions to lagged components imply the lapse of a time unit. In this way, we can determine the time it will take for the flow to propagate and reverberate across the network.

These comments should not be construed as a criticism towards the above cited papers. They are all remarkable studies, with excellent insights coming from various perspectives. We feel that those studies mutually complement each other, and that combining their different angles and strengths would be useful. For example, Boginski et al. (2003, 2005, 2006) and Lautier et al. (2011a, 2011b, 2012a, 2012b) are rich in Graph Theory, however relatively modest in their time series treatment. In contrast, Billio et al. (2012) is stronger in time series analysis however relatively limited in terms of Graph Theory. With this paper we

would like to evoke the spirit of a wonderful manuscript written by Richard Brualdi: "*The Mutually Beneficial Relationship of Graphs and Matrices*" (Brualdi, 2010).

4. Combinatorial recursive estimation of svar systems

Let $\{y_{i,t}\}$ be a series of real-valued random variables y_i, indexed by time t, with $i = 1, \ldots, n$. Consider a Structural Vector AutoRegression (SVAR) system of p lags on n equations, such as the one characterized by Eq. (1):

$$\begin{aligned} y_{1,t} &= \sum_{j=2}^{n} \varphi_{1,j,0} y_{j,t} + \sum_{j=1}^{n} \sum_{k=1}^{p} \varphi_{1,j,k} y_{j,t-k} + \varepsilon_{1,t} \\ y_{i,t} &= \sum_{\substack{j=1 \\ j \neq i}}^{n} \varphi_{i,j,0} y_{j,t} + \sum_{j=1}^{n} \sum_{k=1}^{p} \varphi_{i,j,k} y_{j,t-k} + \varepsilon_{i,t} \\ &\ldots \\ y_{n,t} &= \sum_{j=1}^{n-1} \varphi_{n,j,0} y_{j,t} + \sum_{j=1}^{n} \sum_{k=1}^{p} \varphi_{n,j,k} y_{j,t-k} + \varepsilon_{n,t} \end{aligned} \quad (1)$$

where $\varphi_{i,j,k}$ is the regression coefficient associated with equation i, regressor j, after k lags. The error terms satisfy the conditions of a generalized white noise, see Hamilton (1994, p.257, 258) for details. Without loss of generality, we shall assume that intercepts are null after centering the variables.

A SVAR system combines contemporaneous and lagged effects. Unlike Vector AutoRegression (VAR) systems on n equations, SVAR systems are not regular in the sense that not all variables appear in all equations. Because the left-hand side variables in each equation are also explanatory variables on other equations, the system is simultaneously determined. The Gauss-Markov theorem no longer holds, and OLS yields biased and inconsistent results, see Greene (2008, p.316). The system may not even have a solution in the absence of some structure that allows its identification, see Hamilton (1994, p.332). Alternative estimation procedures should be considered, such as instrumental variables, reduced-form estimates, 2-Stage Least Squares (2SLS), etc. The interested reader can find a detailed discussion in Hamilton (1994, section 11.6).

In this section we introduce a procedure, called Combinatorial Recursive Estimation (CORE), which may be helpful in some applications. The SVAR system in Eq. (2) can be rewritten in matrix form as

$$\Phi Y = BZ + U \quad (2)$$

where Φ is a nxn matrix such that for $i, j = 1, \ldots, n$:

$$\Phi_{i,j} = \begin{cases} -\varphi_{i,j,0} & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases} \quad (3)$$

and B is the $nx(np)$ matrix that contains the parameters that multiply the lagged variables, $\{\varphi_{i,j,k}\}$, where $k = 1, \ldots, p$. This system is identified if Φ has rank n (completeness condition). A sufficient, non-necessary condition for achieving rank n is to transform Φ into a lower or upper triangular matrix, see Hamilton (1994, p.330). A solution is guaranteed to exist for these systems, because they can be estimated recursively, see Greene (2008, p.358), in a similar fashion to how the Gauss-Jordan algorithm solves a system of equations by elimination. These are known as Recursive Structural VAR systems, and they can be estimated via OLS provided that the disturbances are uncorrelated, see Greene (2008, p.372).

A first attempt of ensuring full rank of Φ is to impose exclusion restrictions on $\frac{n(n-1)}{2}$ elements of Φ, and reorder them in such a way that the result is a lower triangular matrix. What items of $\Phi_{i,j}$ are zeroed is determined by the order of the variables. This means that there are $n!$ (i.e., n factorial) feasible permutations. For example, if $n = 9$, we can form 362,880 triangular matrices, each associated with a feasible solution to the SVAR system! This situation no question contributes to endless debates among scholars regarding competing economic models. We could attempt all possible combinations, and select the one that yields the highest explanatory power. The larger the number of combinations attempted, the higher the risk of overfitting the system, see Bailey et al. (2013a, 2013b).

If the number of endogenous variables n is so large that it is computationally impractical to estimate all possible lower-triangular solutions to the system, we can attempt the following alternative (CORE) procedure:

1. Suppose that, due to computational limitations, we can only compute triangular matrices for Φ of up to rank $\tilde{n} < n$. For example, if the researcher's computational power is limited to $\bar{R} = 50,000$ system estimations, then only up to 8 contemporary parameters can be left free, because $8! < 50,000$ and $9! > 50,000$.
2. Compute the covariance matrix on Y.

3. Run $\tilde{n}$–1 iterations of the covariance clustering procedure introduced in Bailey et al. (2012).[2] The resulting clusters contain the set of variables responsible for the greatest increase in the covariance's condition number, and therefore exhibit the highest interdependence. These are the variables on which contemporary effects should be given priority.
4. Estimate SVAR systems on the $\tilde{n}!$ alternative triangular matrices Φ. Statistically insignificant regressors are discarded.
5. Among the SVAR system solutions computed in step 4, select the one with highest explanatory power among those with nearly diagonal disturbance correlation matrix.

Because this procedure creates triangular matrices Φ, a solution is guaranteed among the feasible systems evaluated. These feasible systems have been selected among those where contemporary effects are expected to be strongest, as evidenced by the covariance clustering method (step 3). We believe that this procedure may prove useful in a number of situations. However, like with any automated model selection procedure, we strongly caution against its indiscriminate use. The output may be overfit, in which case it would be helpful to remove some degrees of freedom by imposing additional structure, based on theoretical considerations.

5. The framework

In this section we demonstrate with a numerical example how SFDs can visualize complex flow dynamics. As a practical application, we will study financial flows. This is merely an illustration, and is not proposed as a unique representation of the financial system. The same visualization technique can be applied to represent competing frameworks. It would be interesting to complement this particular example with more detailed and complex implementations, but that would fall outside the general purpose of this publication.

5.1. The data

Our data consist of daily time series of 54 of the most liquid Futures contracts traded worldwide, on every major asset class. A complete list is shown in Table 1. Samples begin on the dates indicated in the table. After aligning them, we obtain a daily time series table spanning from March 18th 2001 to December 31st 2013. Currencies are not present because we have translated all price series in terms of U.S. dollar values per contract, applying the point values and exchange rates relevant to each particular date. The source of our data is Bloomberg, and we have reported the code used for each contract. Contracts are rolled dynamically, meaning that price changes simulate holding a position on the two front contracts with varying weight: We begin with a full weight on the second contract the day that it becomes the new front contract (that is, the day the first contract expires). As the expiration of the new front contract approaches, that weight linearly decreases from one towards zero, while the weight of the new second contract increases from zero towards one.

5.2. Feature extraction by supervised PCA

The open interest of a Futures contract is defined as the number of contracts standing at a particular date. This is a concept similar to the amount of company shares issued, and it measures the size of the Futures market. In our analysis, a large drop in dollar value will have little effect in the financial system if the open interest is small, because very few market participants will be affected by changes in financial valuations. However, a large drop in dollar value will have a substantial effect if the open interest is high, because a large portion of the marketplace may receive a margin call, which may result in credit problems, liquidation of wagers and even bankruptcies. For example, during the 2008 financial crisis there was a large transfer of wealth from net long equity investors to net short equity investors. Investors with high open interest in long equity wagers were forced out of other wagers in order to finance their overall position, thus contributing to the propagation and reverberation of the shock through the financial system. The stock market bottomed around March 12th 2009 and prices made new high-watermarks by March 27th 2013. However, those long equity investors who reduced their open interest before the 2009–2013 rally never recovered their losses. Thus, just looking at price changes does not suffice to understand capital flows. We must combine price changes with open interest information, $\Delta P_{i,t} OI_{i,t}$, where $t = 1, \ldots, T$ indexes time, i is the

[2] That paper contains a detailed description of this clustering algorithm, including Python code. We refer the reader to that publication for further details.

Table 1
Dataset used in the numerical example

Instrument	Class	Code	Exchange	Currency	Point value	Avg. Volume	Avg. Traded (USD)	Start
Corn No. 2	Agriculture	C 1 Comdty	CBT	USD	50	110,264	2,356,897,988	01/03/00
Cotton No. 2	Agriculture	CT1 Comdty	NYB	USD	500	12,784	537,695,040	01/03/00
Sugar No. 11	Agriculture	SB1 Comdty	NYB	USD	1,120	46,385	853,553,309	01/03/00
Coffee	Agriculture	KC1 Comdty	NYB	USD	375	12,340	538,427,078	01/03/00
Cocoa	Agriculture	CC1 Comdty	NYB	USD	10	10,512	292,761,057	01/03/00
Live Cattle	Agriculture	LC1 Comdty	CME	USD	400	29,192	1,575,784,160	01/03/00
Lean Hogs	Agriculture	LH1 Comdty	CME	USD	400	13,728	470,325,848	01/03/00
Soybean	Agriculture	S 1 Comdty	CBT	USD	50	57,835	3,799,036,563	01/03/00
Soybean Meal	Agriculture	SM1 Comdty	CBT	USD	100	26,349	1,125,629,280	01/03/00
Soybean Oil	Agriculture	BO1 Comdty	CBT	USD	600	36,261	856,122,210	01/03/00
Wheat No. 2	Agriculture	W 1 Comdty	CBT	USD	50	43,613	1,328,019,910	01/03/00
Brent	Energy	CO1 Comdty	ICE	USD	1,000	126,905	14,236,202,900	01/04/00
WTI	Energy	CL1 Comdty	NYM	USD	1,000	110,745	11,109,938,400	01/03/00
Henry Hub	Energy	NG1 Comdty	NYM	USD	10,000	107,398	4,691,144,640	01/04/00
UK Nat Gas	Energy	FN1 Comdty	ICE	GBP	310	6,240	224,346,848	01/04/00
Heating Oil	Energy	HO1 Comdty	NYM	USD	420	44,649	5,858,467,275	01/04/00
Gasoil	Energy	QS1 Comdty	ICE	USD	100	50,196	4,793,718,000	01/04/00
T-Note 10Yr	Bond	TY1 Comdty	CBT	USD	1,000	718,634	88,380,782,040	01/03/00
T-Note 5Yr	Bond	FV1 Comdty	CBT	USD	1,000	432,698	51,606,001,382	01/03/00
Euro Bund	Bond	RX1 Comdty	EUX	EUR	1,000	235,435	44,981,242,897	01/03/00
T-Note 2Yr	Bond	TU1 Comdty	CBT	USD	2,000	150,236	33,019,048,549	01/03/00
Euro Bobl	Bond	OE1 Comdty	EUX	EUR	1,000	167,440	28,599,404,180	01/03/00
T-Note Long	Bond	US1 Comdty	CBT	USD	1,000	202,771	26,011,738,724	01/03/00
Japanese 10Yr	Bond	JB1 Comdty	TSE	JPY	1,000,000	16,465	22,493,245,501	01/04/00
Euro Schatz	Bond	DU1 Comdty	EUX	EUR	1,000	134,917	20,449,287,110	01/03/00
Long Gilt	Bond	G 1 Comdty	LIF	GBP	1,000	95,052	16,620,058,650	01/04/00
Eurodollar	Interest Rate	ED1 Comdty	CME	USD	2,500	206,314	50,990,414,478	01/03/00
Euribor	Interest Rate	ER1 Comdty	LIF	EUR	2,500	80,769	27,686,126,670	01/04/00
Short Sterling	Interest Rate	L 1 Comdty	LIF	GBP	1,250	48,802	9,975,532,292	01/04/00
Canadian Bank Bill	Interest Rate	BA1 Comdty	MSE	CAD	2,500	19,065	4,395,679,032	01/04/00
Fed Fund 30d	Interest Rate	FF1 Comdty	CBT	USD	4,167	2,069	861,376,363	01/03/00
S&P 500 E-mini	Equity	ES1 Index	CME	USD	50	622,662	57,175,938,150	01/03/00
Euro Stoxx 50	Equity	VG1 Index	EUX	EUR	10	433,896	18,535,231,375	01/03/00
DAX Index	Equity	GX1 Index	EUX	EUR	25	48,070	15,847,075,745	01/03/00
NASDAQ 100 E-mini	Equity	NQ1 Index	CME	USD	20	94,798	6,769,051,190	03/20/00
Hang Seng	Equity	HI1 Index	HKG	HKD	50	42,668	6,417,087,777	01/03/00
KOSPI200	Equity	KM1 Index	KFE	KRW	500,000	47,247	5,951,602,384	01/04/00
FTSE 100	Equity	Z 1 Index	LIF	GBP	10	52,677	5,821,349,677	01/04/00
TOPIX	Equity	TP1 Index	TSE	JPY	10,000	37,781	4,690,087,633	01/04/00
CAC 40 10 EUR	Equity	CF1 Index	EOP	EUR	10	66,240	3,898,851,227	01/03/00
ASX SPI 200	Equity	XP1 Index	SFE	AUD	25	15,524	1,837,898,775	05/02/00
Swiss Market	Equity	SM1 Index	EUX	CHF	10	15,732	1,439,656,464	01/03/00
IBEX 35	Equity	IB1 Index	MFM	EUR	10	9,886	1,337,886,290	01/03/00
Hang Seng mini	Equity	HU1 Index	HKG	HKD	10	16,511	495,610,373	10/09/00
MSCI Singapore	Equity	QZ1 Index	SGX	SGD	200	3,909	224,718,394	01/03/00
Euro STOXX Banks	Equity	CA1 Index	EUX	EUR	50	14,992	145,215,068	03/19/01
MSCI Taiwan	Equity	TW1 Index	SGX	USD	100	4,181	125,806,290	01/03/00
Copper	Base Metals	HG1 Comdty	CMX	USD	250	36,407	3,080,942,375	01/04/00
Primary Aluminum	Base Metals	LA1 Comdty	LME	USD	25	34,078	1,511,143,356	01/03/00
Zinc	Base Metals	LX1 Comdty	LME	USD	25	15,965	834,349,908	01/03/00
Lead	Base Metals	LL1 Comdty	LME	USD	25	4,270	242,684,758	01/03/00
Gold	Precious Metals	GC1 Comdty	CMX	USD	100	101,035	12,265,620,673	01/04/00
Silver	Precious Metals	SI1 Comdty	CMX	USD	5,000	31,950	3,202,814,384	01/04/00
Platinum	Precious Metals	PL1 Comdty	NYM	USD	50	12,671	873,602,095	01/04/00

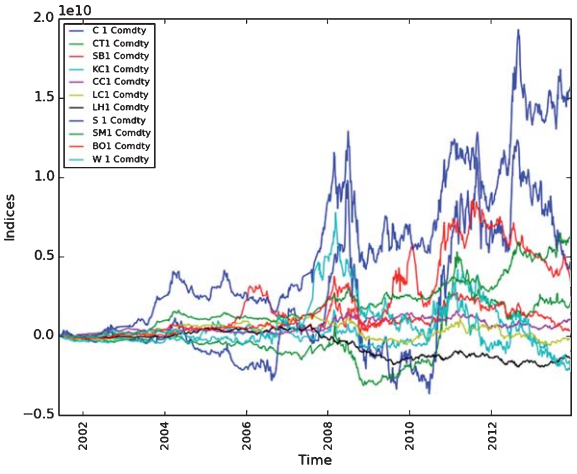

Fig. 1. Cumulative times series of rolled futures contracts, weighted by open interest (Agriculture).

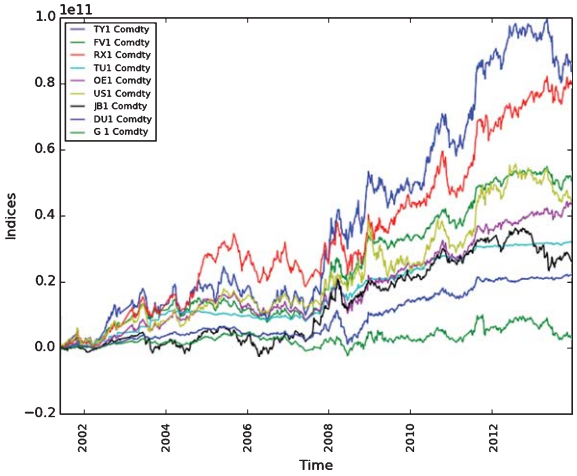

Fig. 2. Cumulative times series of rolled futures contracts, weighted by open interest (Base Metals).

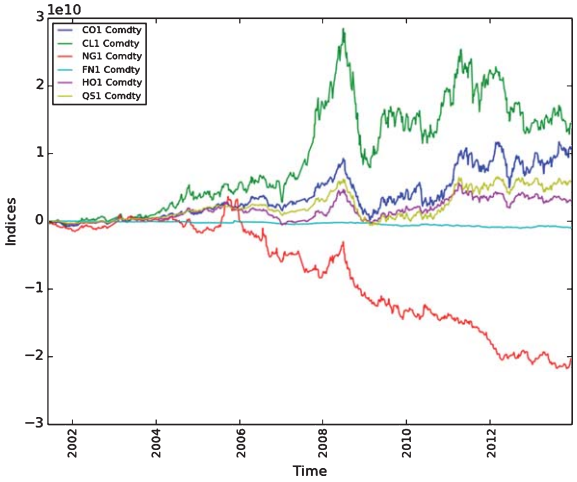

Fig. 3. Cumulative times series of rolled futures contracts, weighted by open interest (Bonds).

security index, $\Delta P_{i,t}$ is the change in U.S. dollar value of a Futures contract i between observations $t-1$ and t, and $OI_{i,t}$ the open interest associated with that contract i at time t. Figures 1 to 7 plot the cumulative time series for $\Delta P_{i,t} OI_{i,t}$, sampled on a weekly basis, grouped by asset class, re-indexed with zero value at the beginning of the series.

As these plots evidence, these series have very different means and standard deviations. We will standardize these variables to prevent that our analysis is influenced by the different scale and levels of the variables involved. Centering the data also brings the practical benefit of giving us a reference for comparing historical values, which will prove useful when we plot results. Consider a set of m random variables $X = \{x_1, \ldots, x_i, \ldots, x_m\}$ and a series $t = 1, \ldots, T$ with

$$x_{i,t} = \frac{\Delta P_{i,t} OI_{i,t} - E[\Delta P_i OI_i]}{\sqrt{V[\Delta P_i OI_i]}} \tag{4}$$

where operator $E[\cdot]$ computes the expected value, and $V[\cdot]$ computes the variance of the random variable between the squared brackets. In combining dollar value and open interest information, the standardized variable $x_{i,t}$ indicates the relative magnitude, direction and relevance of flows impacting contract i.

We would like to extract the main features contained in the X set. We split set X into logical subsets, X_s, with $s = 1, \ldots, S$ and $S \leq m$. We denote m_s as the number of elements belonging to subset s, with $\sum_{s=1}^{S} m_s = m$. With logical subsets we mean subsets that have a fundamental meaning to the researcher, such as asset class, sector, country, currency of denomination, econometric factors, or some combination of those. Then, we can apply PCA to perform a linear mapping of X_s to a lower dimensional space. This linear mapping is such that the variance of the data in the low-dimensional representation is maximized. More precisely, let $C_s = \frac{1}{T} X_s' X_s$ be the correlation matrix computed on subset X_s. The spectral theorem decomposes C_s as

$$C_s = W_s \Lambda_s W_s^{-1} \tag{5}$$

where Λ_s is the eigenvalues matrix and W_s is the eigenvectors matrix. Because C_s is a squared and symmetric

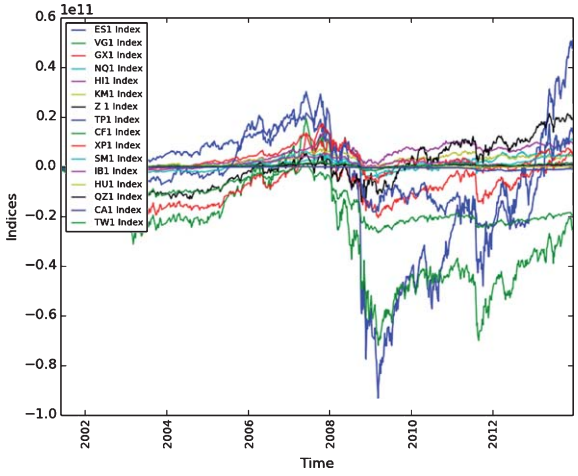

Fig. 4. Cumulative times series of rolled futures contracts, weighted by open interest (Energy).

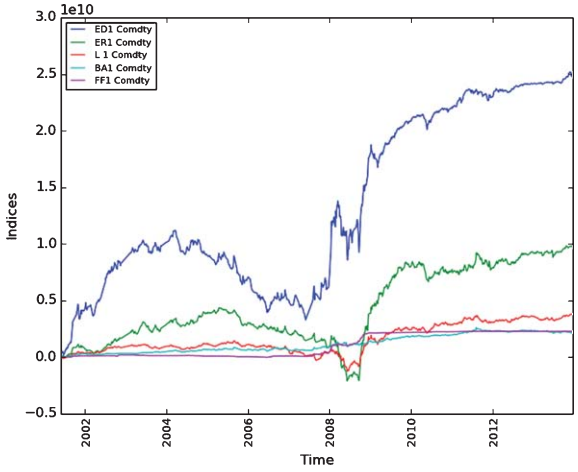

Fig. 5. Cumulative times series of rolled futures contracts, weighted by open interest (Equity indices).

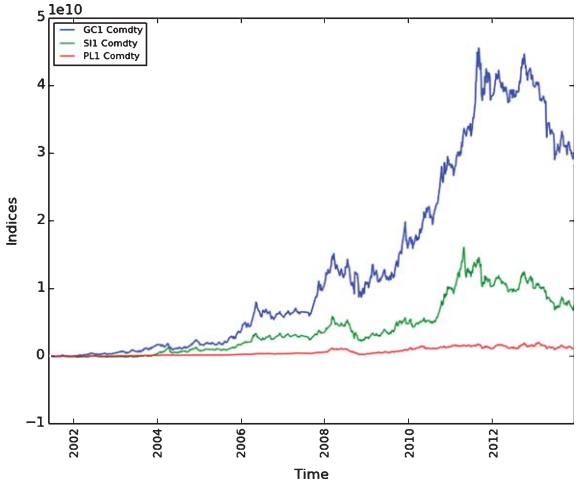

Fig. 6. Cumulative times series of rolled futures contracts, weighted by open interest (Interest rate).

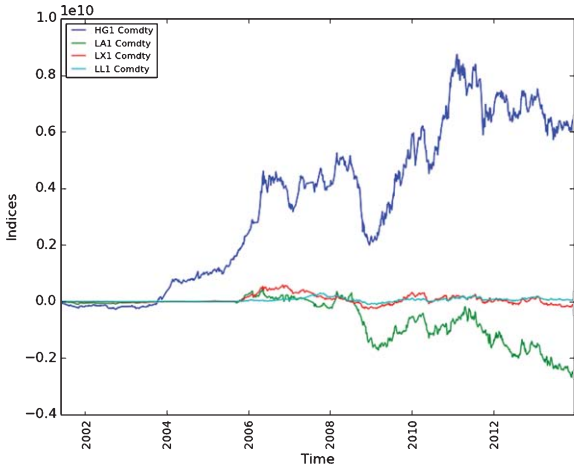

Fig. 7. Cumulative times series of rolled futures contracts, weighted by open interest (Precious Metals).

matrix, the vectors in W_s are real-valued and orthonormal, i.e. $W_s^{-1} = W_s'$, with unit length. Λ_s is a squared diagonal matrix. For convenience, we can reorder the columns in W_s and Λ_s such that $\Lambda_{s_{i,i}} \geq \Lambda_{s_{j,j}}, \ \forall j > i$. Thus, the first column of W_s, denoted W_{s_1}, can be identified as the linear mapping of X_s into the component that spans in the direction of maximum variance present in s. The second column of W_s, denoted W_{s_2}, spans in the direction of maximum variance present in s that is also orthogonal to the previous one, and so on. The time series representative of these extracted features, called principal components, are the projection of X_s into the orthogonal basis W_s, and they are given by $\tilde{X}_s = X_s W_s$. This can be verified by noting that $\tilde{X}_s' \tilde{X}_s = W_s' X_s' X_s W_s = W_s' T C_s W_s = T \Lambda_s$.

Because the diagonal of C_s is a vector of ones, we know that $\sum_{j=1}^{m_s} \Lambda_{s_{j,j}} = m_s$. For each s, a number of components $n_s \leq m_s$ can be chosen such that $\sum_{i=1}^{n_s - 1} \Lambda_{s_{i,i}} < \theta m_s$ and $\sum_{i=1}^{n_s} \Lambda_{s_{i,i}} \geq \theta m_s$, where θ is a real number, $\theta \in [0, 1]$. That is, n_s is the number of components needed to extract a portion θ of the variance contained in X_s. The relevant time series are then given by the first n_s columns of $\tilde{X}_s$.

Finally, we repeat the above feature extraction procedure on each of the S logical subsets. This allows us to form the new set Y of extracted features as the union of features extracted from the various logical subsets,

$$Y = \bigcup_{s=1}^{S} \left\{ \tilde{x}_{s,1}, \ldots, \tilde{x}_{s,n_s} \right\} \tag{6}$$

where $\tilde{x}_{s,j}$ is the jth principal component of logical subset s, and $n = \sum_{s=1}^{S} n_s$ is the size of the new set, $Y = \{y_1, \ldots, y_n\}$. Note that all variables in Y are centered, as a result of the transformation performed in Eq. (5), however they do not necessarily have unit standard deviation, because they are principal components extracted from X_s.

PCA is a standard dimensionality reduction technique (see Muirhead (1982) for an exposition) by virtue of the fourth part of the Fundamental Theorem of Linear Algebra (Strang, 1993). A common critique to this procedure is that it involves a change of basis, thus yielding series that are often far from intuitive (Bailey et al., 2012). But if the reader is familiar with this technique, she may have realized that we have applied it in a manner quite different from the standard. We have carried out PCA on subsets X_s rather than on the entire system X. In doing so, we are preserving the logical structure introduced by the researcher, when she split X into disjoint subsets endowed with certain fundamental attributes. Thus, our implementation of PCA extracts those attributes that the researcher identified as essential to understanding the system, what we may call a *supervised PCA*.

5.3. Feature modelling by core

In a SVAR system such as the one discussed in Section 4, current realizations of one variable, say $y_{j,t}$, affect current realizations of another one, say $y_{i,t}$, with $i \neq j$. Past realizations of one variable, say $y_{j,t-k}$, affect current realizations of $y_{i,t}$, with $i, j \in \{1, \ldots, n\}$ and $k = 1, \ldots, p$. We would like to understand the synchronous as well as asynchronous interdependencies within the system formed by $\{y_i\}$, $i = 1, \ldots, n$. The presence of synchronous interdependencies may be captured by a correlation coefficient, but correlation will miss asynchronous effects. Non-linear responses are not a concern, as long as they can be linearized, through non-linear transformations such as those discussed in McNelis (2005) or Franses et al. (2000). In order to conduct statistical inference, we will assume that the error terms in that specification follow a Normal distribution.

We apply the CORE procedure described in Section 4. The n fitted equations will assign a real value to each estimated parameter in the $nx(np + 1)$ solution matrix formed by the ordered set $\{\hat{\varphi}_{i,j,k}\}$ of estimated regression coefficients, where $i, j \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, p\}$. This would lead to the establishment of connections even where $\hat{\varphi}_{i,j,k}$ may be statistically insignificant. For this reason, during the CORE estimation process we disregard those statistically insignificant connections regardless of the supposed correlation value between variables. In order to do that, given a significance level α, we will test each estimated regressor for statistical significance. Because individual significance tests on a particular regressor are not invariant to the subset of selected regressors, they should be carried out on all possible combinations of subsets of selected regressors. Each equation in Eq. (1) contains $n(p + 1) - 1$ candidate regressors. Thus, testing the significance over the entire system will involve $n2^{n(p+1)-1}$ tests (that is, $2^{n(p+1)-1}$ for each of the n equations). This is a relatively trivial computational task for small systems, but beyond 8 variables it will involve over a million regressions, even if only a single lag is considered ($p = 1$). For large systems, a standard automatic regression selection procedure could be employed, such as Stepwise (Hocking, 1976). However, the reader should be aware that such procedures are prone to overfitting, because the significance test is not incorporating information about the total number of experiments conducted (see Bailey et al. (2013a, 2013b); Foster et al. (1994) and Donoho et al. (1994)).

6. Empirical results by SFD

The results from the estimation above can be used to build this system's SFD. Calkin and López de Prado (2014) provides the Python code that carries out the task of building a SFD from the matrix containing the system's parameters estimates. The resulting system can be simplified by reducing the value of α and repeating the regressor selection process. In other words, as $\alpha \to 0$, only the most statistically significant connections will survive the selection process. Low values of α yield the SFD of most likely propagation of a flow. We can therefore associate various scenarios with a degree of statistical confidence.

6.1. Weekly flows

In this section we study short-term financial flows by modeling price adjustments on the series listed in Table 1, sampled on a weekly basis. We group the time series into logical subsets, defined in terms of asset class. The feature extraction threshold is set to $\theta = 0.5$, meaning that we are satisfied with extracting the key features accounting for over 50% of the variance of each asset class. In the case of futures on Interest Rates, Bonds, Equity indices, Energy, Precious Metals and Base Metals, a single principal component from each of those asset classes is sufficient to explain over 50% of their variances. We label each of these principal components with the name of the asset class, followed by a "1" that indicates that they are associated with their first principal component. Only in the case of Agricultural futures we need to take into consideration two principal components, which we label "Agriculture1" and "Agriculture2", where the former is dominated by grains and the latter by cattle. Table 2 shows the composition of the extracted features. Figures 8–14 plot the resulting indices. For example, Fig. 12 plots the cumulative time series for Equity1. At first sight it may seem surprising that the series ends at zero, since the Standard & Poors 500 index was at historical highs around the end of 2013. There are three reasons for that: First, as indicated in Section 5.2, the series in Y are centered. By construction, the cumulative series on Y start and end at zero. In other words, centering Y has the effect of de-trending cumulative Y. Second, Fig. 5 shows that ES1 Index is indeed at a historical new high around the end of 2013, but that is not the case of most other indices. For example, European indices are far from recovering the pre-crisis price levels. Third, because we are taking open interest into account, a price recovery cannot compensate for the 2008 losses if the size of the futures market is smaller during the 2009–2013 recovery.

We model these principal components through a SVAR system with one lag. Lagged variables are labeled with the suffix "_1". Individual regressors are subjected to Wald's test of statistical significance, with $\alpha = 0.05$. Table 3 reports the estimated betas of the SVAR solution, and Table 4 the t-Stats. Table 5 reports the adjusted R-squares, probabilities of the regressions' F-stats and standard deviations of the regressions' residuals.

Figure 15 draws the resulting SFD, which summarizes the most critical information about the system.

Table 2
Composition of logical subsets (weekly)

Code	Base Metals1	Code	Energy1
HG1 Comdty	0.5198	CL1 Comdty	0.4837
LA1 Comdty	0.4834	CO1 Comdty	0.4894
LL1 Comdty	0.4698	FN1 Comdty	−0.0770
LX1 Comdty	0.5248	HO1 Comdty	0.4921
		NG1 Comdty	0.1571
		QS1 Comdty	0.5037
Code	Precious Metals1	Code	Interest Rate1
GC1 Comdty	0.5913	BA1 Comdty	0.4327
PL1 Comdty	0.5556	ED1 Comdty	0.4970
SI1 Comdty	0.5845	ER1 Comdty	0.4703
		FF1 Comdty	0.3646
		L 1 Comdty	0.4600
Code	Equity1	Code	Bond1
CA1 Index	0.1855	DU1 Comdty	0.3097
CF1 Index	0.2638	FV1 Comdty	0.3697
ES1 Index	0.2559	G 1 Comdty	0.3078
GX1 Index	0.2735	JB1 Comdty	0.2686
HI1 Index	0.2442	OE1 Comdty	0.3615
HU1 Index	0.2346	RX1 Comdty	0.3623
IB1 Index	0.2602	TU1 Comdty	0.2923
KM1 Index	0.2369	TY1 Comdty	0.3733
NQ1 Index	0.2433	US1 Comdty	0.3372
QZ1 Index	0.2579		
SM1 Index	0.2615		
TP1 Index	0.2156		
TW1 Index	0.2279		
VG1 Index	0.2837		
XP1 Index	0.2626		
Z 1 Index	0.2743		

Code	Agriculture1	Agriculture2
BO1 Comdty	0.4037	−0.0389
C 1 Comdty	0.3754	−0.1240
CC1 Comdty	0.2020	0.1307
CT1 Comdty	0.2675	0.0181
KC1 Comdty	0.2525	0.1431
LC1 Comdty	0.1308	0.6304
LH1 Comdty	0.0709	0.6720
S 1 Comdty	0.4352	−0.1783
SB1 Comdty	0.2558	0.1221
SM1 Comdty	0.3762	−0.2218
W 1 Comdty	0.3185	−0.0171

Following the convention described in Appendix 2, we can visually determine:

- The variables (as many as vertices).
- The number of equations (n, as many as elliptical vertices).
- The lagged variables involved (diamond-shaped vertices).
- How variables interact with each other (arcs).
- The direction of the flow (arc's side with pointed or flat arrowhead).

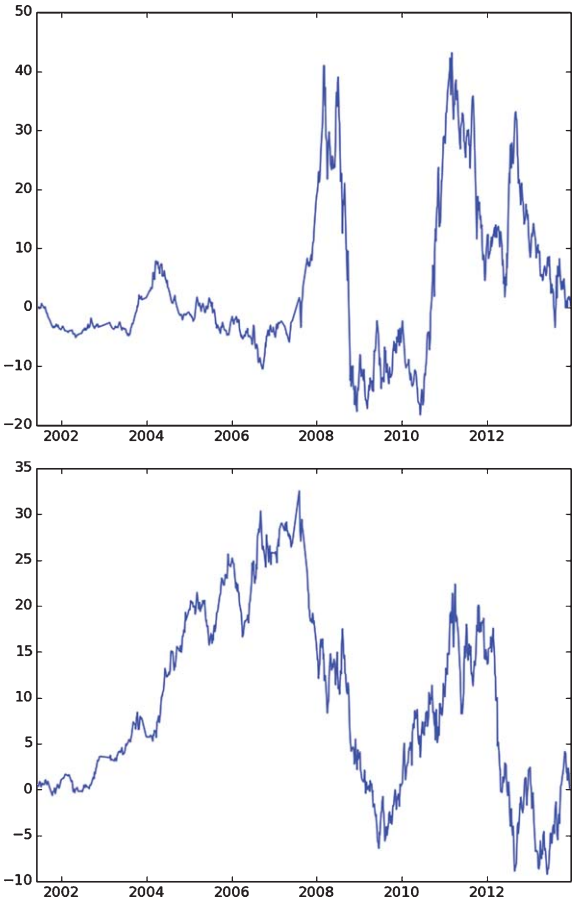

Fig. 8. Cumulative time series for the features extracted from Agriculture (top for Agriculture1, bottom for Agriculture 2).

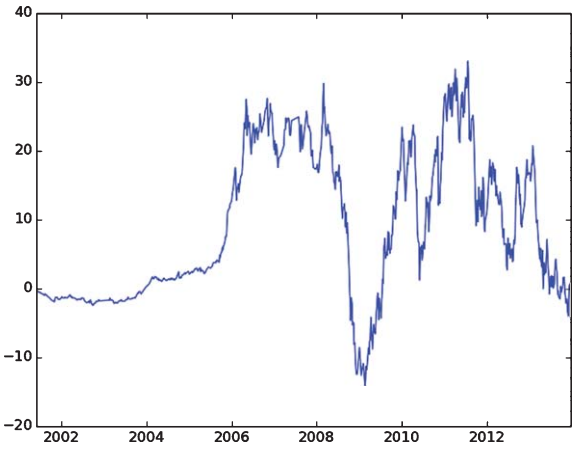

Fig. 9. Cumulative time series for the features extracted from Base Metals.

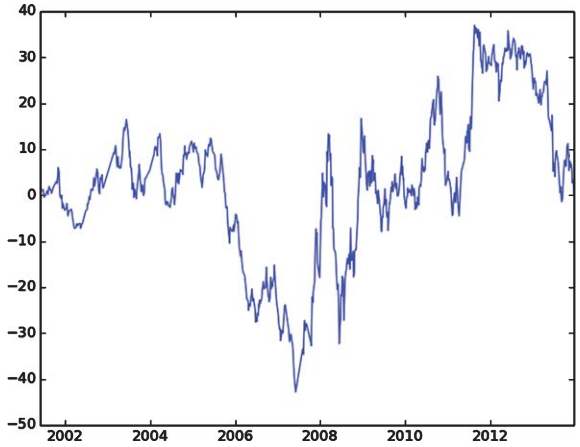

Fig. 10. Cumulative time series for the features extracted from Bonds.

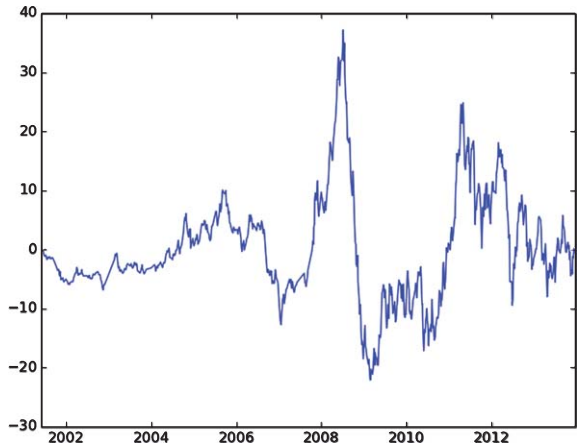

Fig. 11. Cumulative time series for the features extracted from Energy.

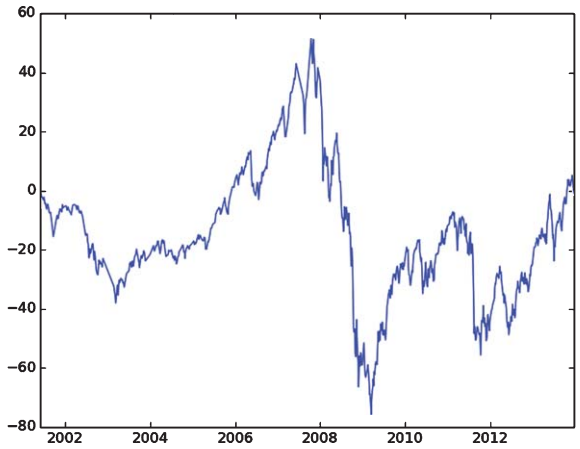

Fig. 12. Cumulative time series for the features extracted from Equity.

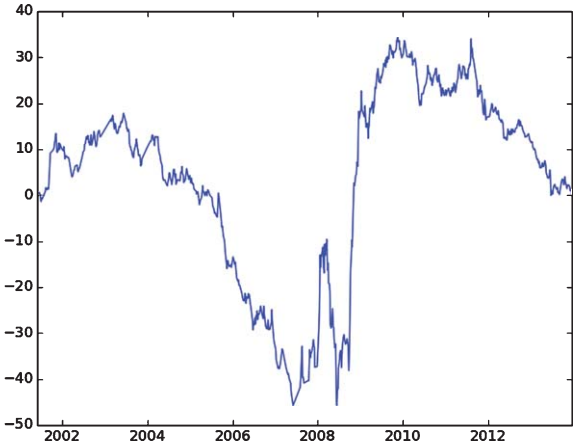

Fig. 13. Cumulative time series for the features extracted from Interest Rates.

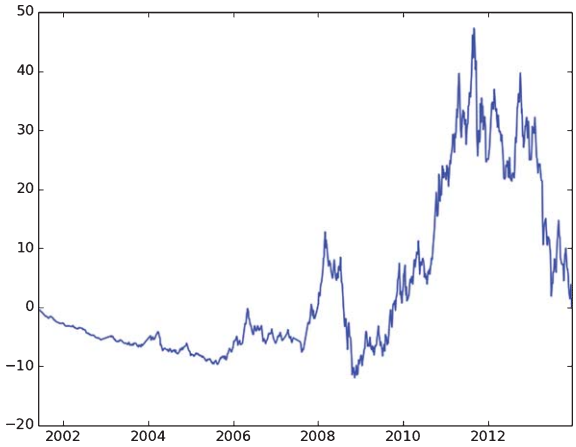

Fig. 14. Cumulative time series for the features extracted from Precious Metals.

- What relationships are contemporaneous (solid arc with flat arrowhead).
- What relationships incorporate a lead-lag effect (solid arc with pointed arrowhead).
- Where memory effects occur (dashed arcs with unfilled pointed arrowhead).
- The strength of the relationships (width of the arc, as an increasing function of the adjusted R-square).
- The sign of the relationship, as indicated by the colour of the arc (red for a negative coefficient and green a positive coefficient).
- The coefficients' p-values are implied by the presence of an arc (if the p-value exceeds α, the arc is not drawn).
- For specifications in differences, the left half of the vertex draws the magnitude of the change, and the right side of the vertex draws the cumulative change.

Figure 15 provides greater insight about the organization of the system than Tables 3, 4 and 5 combined, and yet Figure 15 is based on those same tables. Here comes a key concept we would like the reader to consider:

> *SFDs are more insightful than the standard collection of statistical tables because SFDs shift the focus from the algebraic solution of the system to its logical structure, its topology. Surely, statistical tables are more detailed in terms of reporting estimated values, with a few decimals of precision... However that level of detail also obfuscates how variables interact with each other. Just as Geometry could not help Euler solve the "Seven Bridges of Königsberg" problem, Linear Algebra alone will not help Economists answer many key questions about how financial markets coordinate. Modern Graph Theory in general, and SFDs in particular, can complement Statistical tools to achieve a better understanding of complex dynamic systems.*

SFDs help us understand possible scenarios of flow propagation, what vertices are more critical, what arcs may become overflowed, what shocks could shut the system down, etc. Tables 3, 4 and 5 hide that logical information behind algebraic figures. And this is a relatively simple system. As complexity grows, so does the need to literally see the big picture, rather than being forced to parse through multiple irregular matrices with dozens of entries. As we will see in Sections 7 and 8, SFDs are particularly powerful tools for simulating scenarios.

Although the purpose of this paper is not to discover a particular flow structure, but rather to introduce a new technique for visualizing them, we can use the example in Fig. 15 to briefly discuss this system's architecture. This weighted digraph has diameter 7 and density 0.1758 (recall definitions in Calkin and López de Prado (2014)). Degree correlation (also known as assortativity) is -0.0863, indicating that connections are rather heterogeneous in terms of vertex degrees. The average vertex connectivity is 1.3462, which evidences that the

Table 3
table reports estimations on the contemporary effects, and the bottom table on the lag 1 effects. Lagged variables can be recognized by the "_1" suffix

	Bond1	Equity1	Interest Rate1	Precious Metals1	Energy1	Base Metals1	Agriculture2	Agriculture1
Bond1		−0.2732	0.6545	0.1977	−0.1392		0.1625	
Equity1			−0.4770		0.1751	0.6263	0.1757	0.1506
Interest Rate1				0.2105				
Precious Metals1					0.2052	0.3089		0.0861
Energy1						0.4038		0.2820
Base Metals1							0.1042	0.3846
Agriculture2								
Agriculture1								
	Bond1_1	Equity1_1	Interest Rate1_1	Precious Metals1_1	Energy1_1	Base Metals1_1	Agriculture2_1	Agriculture1_1
Bond1		0.0520						
Equity1					0.1473			−0.1724
Interest Rate1								
Precious Metals1	0.0412							
Energy1								
Base Metals1								
Agriculture2			−0.0601					
Agriculture1	0.0766		−0.1780	0.1210				

Table 4
Estimated t-stats for the SVAR system, 95% confidence level (weekly)

	Bond1	Equity1	Interest Rate1	Precious Metals1	Energy1	Base Metals1	Agriculture2	Agriculture1
Bond1		−11.7543	16.5155	3.9584	−3.3469		2.9107	
Equity1			−8.6769		2.7350	9.0176	2.1441	2.6923
Interest Rate1				5.1513				
Precious Metals1					6.8559	9.5580		3.2928
Energy1						11.1060		9.3916
Base Metals1							2.2686	14.5411
Agriculture2								
Agriculture1								
	Bond1_1	Equity1_1	Interest Rate1_1	Precious Metals1_1	Energy1_1	Base Metals1_1	Agriculture2_1	Agriculture1_1
Bond1		2.4632						
Equity1					2.4794			−3.2014
Interest Rate1								
Precious Metals1	2.2297							
Energy1								
Base Metals1								
Agriculture2			−2.4720					
Agriculture1	2.0854		−3.4504	2.4635				

system is well interconnected. There are 109 simple cycles in this graph, none of them Eulerian or Hamiltonian. It would be extremely difficult for an economist to consider all possible scenarios of flow propagation by merely relying on statistical tables, but thankfully SFDs can carry out those calculations for us.

An analysis of degree of centrality reveals that Equity1 is the most critical vertex, with a centrality coefficient of 0.6923. However, several other vertices come very close, with Agriculture1 at 0.6154, Precious Metals 1 at 0.5385 and Bond1 also at 0.5385. The implication is that weekly flows are relatively decentralized and there is not a single asset class we can point to in order to monitor activity. This of course gives an advantage to Macro traders over asset class specialists.

Although Equity1 has the highest degree of centrality, careful examination tells us that it is not the most influential asset class. In fact, it is one of the most influenced, considering the direction of the flows. A shock on Equity1 would propagate contemporaneously to Bond1, but otherwise Equity1 is impacted by Energy1, Energy1_1, Agriculture1, Agriculture1_1, Agriculture2, Base Metals1 and Interest Rate1. Bond1 is influenced by Equity1, Equity1_1, Bond1, Energy1,

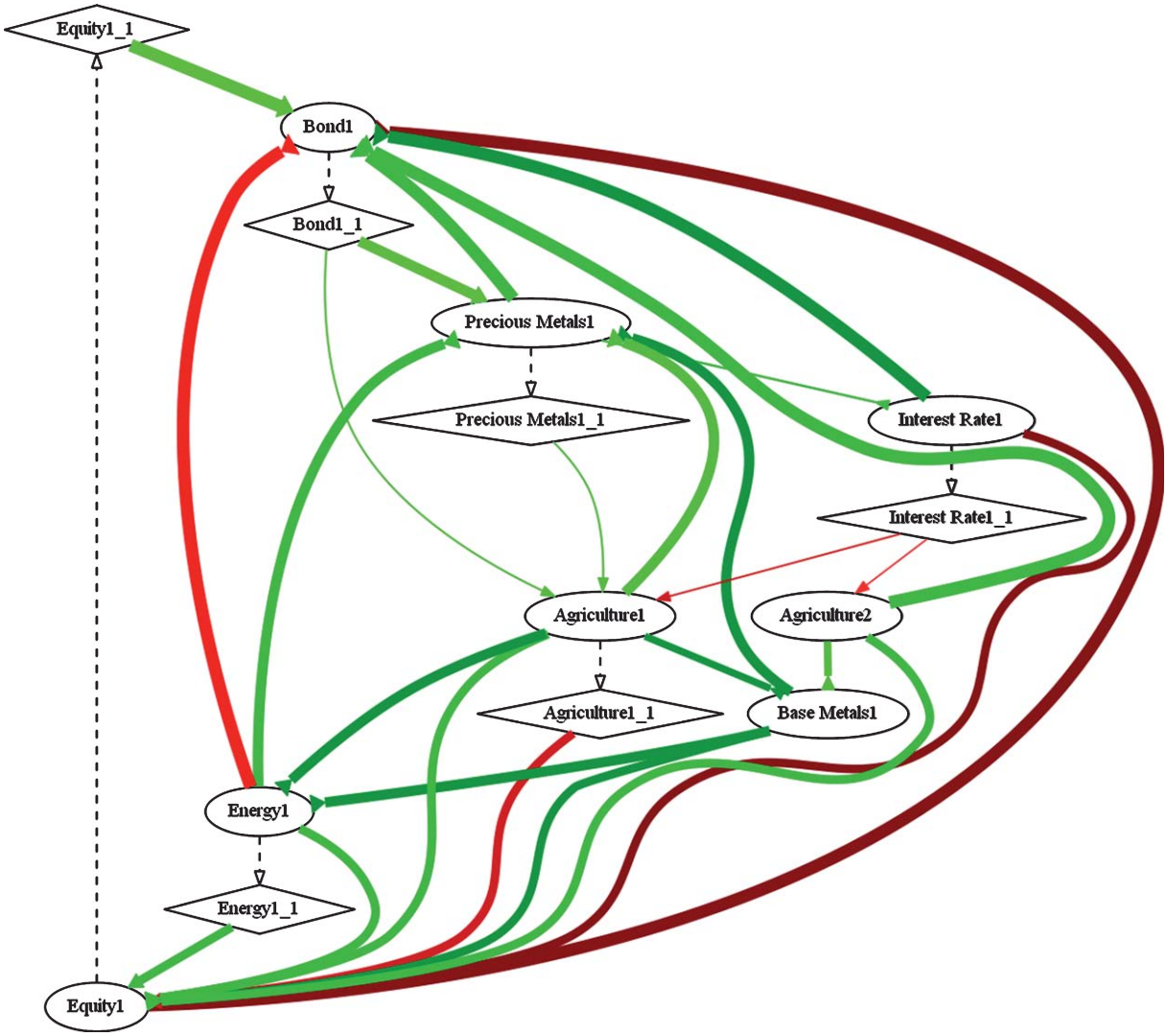

Fig. 15. SFD representation of weekly flows. Elliptical vertices denote a current variable, and diamond-shaped vertices a lagged variable. Green arcs with pointed arrowhead indicate a positive lead-lag effect. Red arcs with pointed arrowhead indicate a negative lead-lag effect. Green arcs with flat arrowhead indicate a positive contemporaneous effect. Red arcs with flat arrowhead indicate a negative contemporaneous effect. The width of the arc displays how strong the relationship is (goodness of fit).

Table 5
Adjusted R-squares, probability of F-stat and standard deviation of residuals (weekly)

Logical Subset	adj-R2	pF	Std_e
Bond1	0.4502	6.44E-96	1.8008
Equity1	0.2761	1.43E-50	2.6398
Interest Rate1	0.0325	3.30E-07	1.7225
Precious Metals1	0.3448	5.59E-69	1.2367
Energy1	0.3400	1.20E-69	1.5004
Base Metals1	0.2200	4.21E-42	1.4929
Agriculture2	0.0067	1.37E-02	1.1741
Agriculture1	0.0164	1.39E-03	2.0294

Precious Metals1, Agriculture2 and Interest Rate1. The only type of arcs that leave Equity1 and Bond1 are outbound arcs, with the exception of the contemporary inbound arc that goes from Equity1 to Bond1. Interest Rate1 influences Bond1 and Equity1.

In contrast, commodities are as central as influential. For example, Agriculture1 influences Energy1, Equity1, Precious Metals1 and Base Metals1. Energy1 influences Equity1, Bond1 and Precious Metals1, and is only influenced by two commodities: Agriculture1 and Base Metals1. The picture that emerges from this

weekly relationships map is a flow system clearly divided between financial assets and commodities, where flows move from commodities to financial assets, and then across the latter. Financial assets reflect the information transmitted by hard assets. Manipulating financial assets is not likely to affect significantly the rest of the system. This evidence may have some implications for the long-standing controversy involving Keynesians and Austrian-school economists, but that is beyond the scope of our paper. We are merely pointing out that SFDs are more likely to address the issues involved in that controversy than the standard statistical and econometric tools used so far.

We can run the historical data through the SFD to visualize how financial flows crossed the network over a particular period of time. An animation of the Weekly Flows model can be found at http://youtu.be/tikYPy7oW-U.

The conclusion is, as it applies to weekly flows, that commodities are the origin of most flows. So when analysts follow the behavior of Stocks and Bonds, they are often looking at the effect of a flow, not its source. The explanation for the behavior of financial assets should be investigated somewhere else, with plenty of trails to search, 109 to be precise! SFDs provide an intuitive and visual tool to recognize the trail followed by a particular flow, and simulate the potential outcomes as the effect reverberates over the following periods.

6.2. *Monthly flows*

As a second example, we will study medium-term financial flows by modeling price adjustments on the series listed in Table 1, sampled on a monthly basis. Otherwise the procedure is very similar to the one presented in the previous section. The feature extraction threshold was set to $\theta = 0.5$. Table 6 shows the composition of the extracted features, which closely resembles what we saw in the weekly case.

Like in the previous case, individual regressors were subjected to tests of statistical significance, with $\alpha = 0.05$. Table 7 reports the estimated betas of the SVAR solution, and Table 8 the t-Stats. Table 9 reports the adjusted R-squares, probabilities of the regressions' F-stats and standard deviations of the regressions' residuals.

Figure 16 draws the resulting SFD. This weighted digraph has diameter 7 and density 0.1428. Base Metals1 exhibits the highest degree of centrality, 0.5714, followed by Agriculture1 at 0.5, Precious Metals1 at 0.4286 and Interest Rate1 also at 0.4286. There are 52 possible simple cycles through this system, however none of them is an Eulerian or Hamiltonian path. This is about half of what we observed in the weekly case. Degree correlation is −0.0416, indicating that connections are again rather heterogeneous in terms of vertex degrees. Average vertex connectivity is 1.3238, similar to the one observed in weekly flows.

Equity1 and Bond1 are less central to monthly flows than it was the case in weekly flows. Like in the weekly case, their influence on commodities is very limited:

Table 6
Composition of logical subsets (monthly)

Code	Base Metals1	Code	Energy1
HG1 Comdty	0.5252	CL1 Comdty	0.4835
LA1 Comdty	0.4783	CO1 Comdty	0.4740
LL1 Comdty	0.4704	FN1 Comdty	−0.0347
LX1 Comdty	0.5236	HO1 Comdty	0.4916
		NG1 Comdty	0.1724
		QS1 Comdty	0.5186
Code	Precious Metals1	Code	Interest Rate1
GC1 Comdty	0.5813	BA1 Comdty	0.4355
PL1 Comdty	0.5534	ED1 Comdty	0.4878
SI1 Comdty	0.5965	ER1 Comdty	0.4716
		FF1 Comdty	0.3777
		L 1 Comdty	0.4553
Code	Equity1	Code	Bond1
CA1 Index	0.2049	DU1 Comdty	0.3037
CF1 Index	0.2640	FV1 Comdty	0.3656
ES1 Index	0.2648	G 1 Comdty	0.3260
GX1 Index	0.2591	JB1 Comdty	0.2853
HI1 Index	0.2325	OE1 Comdty	0.3617
HU1 Index	0.2282	RX1 Comdty	0.3590
IB1 Index	0.2515	TU1 Comdty	0.2949
KM1 Index	0.2446	TY1 Comdty	0.3641
NQ1 Index	0.2593	US1 Comdty	0.3273
QZ1 Index	0.2545		
SM1 Index	0.2564		
TP1 Index	0.2053		
TW1 Index	0.2475		
VG1 Index	0.2731		
XP1 Index	0.2697		
Z 1 Index	0.2710		
Code	Agriculture1		Agriculture2
BO1 Comdty	0.4045		−0.0595
C 1 Comdty	0.4089		−0.0389
CC1 Comdty	0.1952		0.2467
CT1 Comdty	0.2995		0.1169
KC1 Comdty	0.2717		0.2111
LC1 Comdty	0.0381		0.5933
LH1 Comdty	−0.0053		0.6325
S 1 Comdty	0.4348		−0.2052
SB1 Comdty	0.1895		0.1218
SM1 Comdty	0.3774		−0.2339
W 1 Comdty	0.3142		0.1099

Table 7
Estimated coefficients for the SVAR system, 95% confidence level (monthly)

	Bond1	Equity1	Interest Rate1	Precious Metals1	Agriculture1	Base Metals1	Energy1	Agriculture2
Bond1		−0.1872	0.7866					
Equity1			−0.6225	−0.4268		0.9778		
Interest Rate1				0.2736				
Precious Metals1					0.1763	0.3304	0.1788	
Agriculture1						0.3860	0.3862	
Base Metals1							0.4840	0.2677
Energy1								
Agriculture2								
	Bond1_1	Equity1_1	Interest Rate1_1	Precious Metals1_1	Agriculture1_1	Base Metals1_1	Energy1_1	Agriculture2_1
Bond1								
Equity1								
Interest Rate1		−0.1013						−0.2209
Precious Metals1								
Agriculture1			0.1649		0.2076	−0.3241		
Base Metals1	0.1302			−0.1687				
Energy1			−0.2186					0.3299
Agriculture2						0.1624		

Table 8
Estimated t-stats for the SVAR system, 95% confidence level (monthly)

	Bond1	Equity1	Interest Rate1	Precious Metals1	Agriculture1	Base Metals1	Energy1	Agriculture2
Bond1		−3.7109	8.5740					
Equity1			−4.8653	−2.3677		6.1052		
Interest Rate1				3.0529				
Precious Metals1					3.0796	4.8747	2.8400	
Agriculture1						4.3654	4.5643	
Base Metals1							7.5548	2.8853
Energy1								
Agriculture2								
	Bond1_1	Equity1_1	Interest Rate1_1	Precious Metals1_1	Agriculture1_1	Base Metals1_1	Energy1_1	Agriculture2_1
Bond1								
Equity1								
Interest Rate1		−2.3553						−2.0169
Precious Metals1								
Agriculture1			2.1719		2.7982	−3.6864		
Base Metals1	2.7070			−2.1970				
Energy1			−2.6919					2.9224
Agriculture2						2.7331		

Bond1 impacts Base Metals1 through Bond1_1, and Interest Rate1 impacts Agriculture1 through Interest Rate1_1. Among the financial assets, Interest Rate1 is the most influential, impacting Equity1 and Bond1.

The clustering between financial assets on one side and commodities on the other is now even more evident. The cycle Precious Metals1 ⟶ Equity1 ⟶ Equity1_1 ⟶ Interest Rate1 ⟶ Bond1 ⟶ Bond1_1 ⟶ Base Metals1 ⟶ Precious Metals1 is at the core of the system. These are not contemporaneous effects, since several outbound arcs are involved. A shock to any of these assets will reverberate through this cycle over multiple periods. This far from obvious shock propagation path can be anticipated by Macro traders to place their bets regarding the future trajec-

Table 9
Adjusted R-squares, probability of F-stat and standard deviation of residuals (monthly)

Logical Subset	adj-R2	pF	Std_e
Bond1	0.4441	3.71E-20	1.8837
Equity1	0.2860	1.84E-11	2.7271
Interest Rate1	0.0904	6.92E-04	1.6927
Precious Metals1	0.4355	6.27E-19	1.1663
Agriculture1	0.3971	1.04E-15	1.5872
Base Metals1	0.3039	9.37E-12	1.4316
Energy1	0.0882	3.81E-04	1.7690
Agriculture2	0.0411	7.03E-03	1.2514

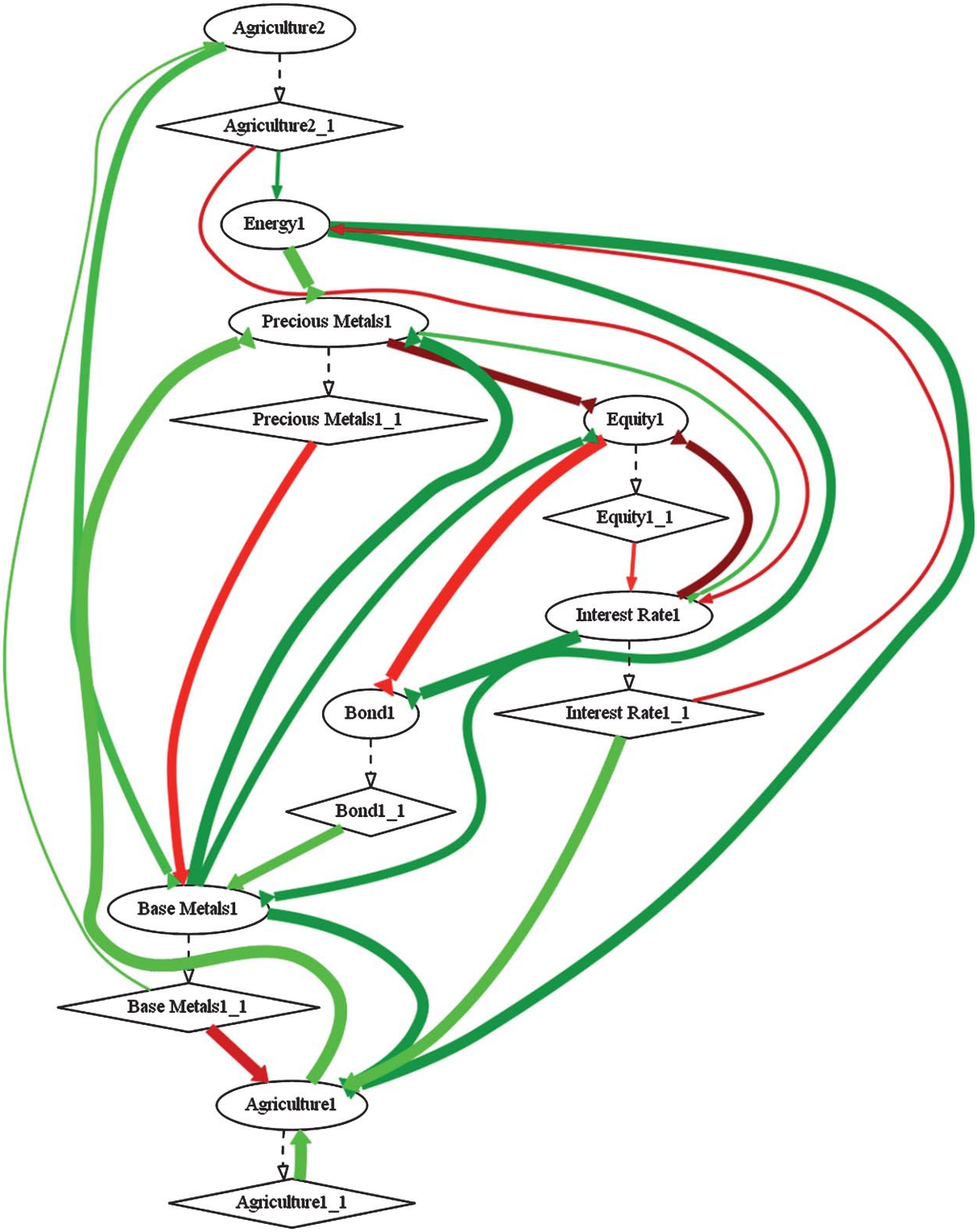

Fig. 16. SFD representation of monthly flows.

tory of the prices involved over the following months. Once again, standard Econometric or Statistical analysis are not designed to recognize this kind of looping mechanisms, much less with the clarity with which Graph Theory exposes them.

6.3. Quarterly flows

As a third example, we will study longer-term financial flows by modeling price adjustments on the series listed in Table 1, sampled on a monthly basis. Following the familiar procedure, the feature extraction threshold was set to $\theta = 0.5$. Table 10 shows the composition of the extracted features, which again resemble the previous cases.

Individual regressors were subjected to tests of statistical significance, with $\alpha = 0.05$. Table 11 reports the estimated betas of the SVAR solution, and Table 12 the t-Stats. Table 13 reports the adjusted R-squares, probabilities of the regressions' F-stats and standard deviations of the regressions' residuals.

Figure 17 draws the resulting SFD. The density of this weighted digraph is 0.1282. Its diameter is infinite, because not all vertices are reachable from another. For instance, Bond1 is a dead end. In the previous cases we saw that Bond1 was becoming less influential, and in the quarterly case it reaches the extreme of being influenced without having an impact on other classes. It would appear as if purchasing bonds may be a rather ineffective way to trigger financial flows in the long run. Central bankers achieve better results by acting on Interest Rate1 and Precious Metals1. The latter is a particularly good asset class to act upon, since it has a degree of centrality of 0.4167. Although Interest Rate1 and Bond1 have the same degree of centrality, the SFD indicates that Interest Rate1 is the one that influences Equity1, Bond1 and Precious Metals1 through Interest Rate1_1. Hence we find another cycle of interest to Macro traders: Precious Metals1 ⟶ Base Metals1 ⟶ Base Metals1_1 ⟶ Interest Rate1 ⟶ Equity1 ⟶ Equity1_1 ⟶ Agriculture1 ⟶ Precious Metals1. Once again, these are not contemporaneous effects, since several outbound arcs are involved. In fact, there are 4 such cycles, less than 20 times the number of cycles we saw in the weekly case, and yet none Eulerian or Hamiltonian. Recognizing cycles is particularly important for Macro investors. As we will see in the following section, some shocks are self-stabilizing, and some are self-reinforced. Understanding both mechanisms provides Macro traders with an opportunity to profit from temporary price misalignments or continuing trends.

Table 10
Composition of logical subsets (quarterly)

Code	Base Metals1	Code	Energy1
HG1 Comdty	0.5282	CL1 Comdty	0.4836
LA1 Comdty	0.5139	CO1 Comdty	0.4743
LL1 Comdty	0.4303	FN1 Comdty	0.0195
LX1 Comdty	0.5214	HO1 Comdty	0.4993
		NG1 Comdty	0.2058
		QS1 Comdty	0.4992
Code	Precious Metals1	Code	Interest Rate1
GC1 Comdty	0.5732	BA1 Comdty	0.4187
PL1 Comdty	0.5642	ED1 Comdty	0.4887
SI1 Comdty	0.5943	ER1 Comdty	0.4545
		FF1 Comdty	0.4076
		L 1 Comdty	0.4616
Code	Equity1	Code	Bond1
CA1 Index	0.1845	DU1 Comdty	0.3322
CF1 Index	0.2610	FV1 Comdty	0.3653
ES1 Index	0.2567	G 1 Comdty	0.3059
GX1 Index	0.2646	JB1 Comdty	0.2641
HI1 Index	0.2437	OE1 Comdty	0.3639
HU1 Index	0.2449	RX1 Comdty	0.3495
IB1 Index	0.2552	TU1 Comdty	0.3045
KM1 Index	0.2478	TY1 Comdty	0.3660
NQ1 Index	0.2449	US1 Comdty	0.3340
QZ1 Index	0.2718		
SM1 Index	0.2619		
TP1 Index	0.1965		
TW1 Index	0.2429		
VG1 Index	0.2790		
XP1 Index	0.2739		
Z 1 Index	0.2514		
Code	Agriculture1		Agriculture2
BO1 Comdty	0.4084		−0.0601
C 1 Comdty	0.3775		0.0122
CC1 Comdty	0.2385		0.1966
CT1 Comdty	0.2945		0.2602
KC1 Comdty	0.3118		0.2562
LC1 Comdty	−0.0073		0.6348
LH1 Comdty	−0.0388		0.5785
S 1 Comdty	0.4327		−0.1797
SB1 Comdty	0.2054		0.0808
SM1 Comdty	0.3307		−0.2004
W 1 Comdty	0.3310		−0.0876

Degree of correlation is 0.0802, indicating that vertex degrees are more homogeneous. Connectivity has dropped to 0.8205, something we could have suspected from the far smaller number of cycles. Although quarterly flows may appear easier to predict due to the lower number of possible cycles, the number of bets one can place per year is also far smaller (4 in the quarterly case, compared to 12 in the monthly or 52 in the weekly case). Hence it pays building a shorter term SFD and

Table 11
Estimated coefficients for the SVAR system, 95% confidence level (quarterly)

	Bond1	Equity1	Interest Rate1	Energy1	Base Metals1	Precious Metals1	Agriculture1	Agriculture2
Bond1			0.7931			0.6225		
Equity1			−1.1299					
Interest Rate1								
Energy1							0.2784	
Base Metals1						0.7021		
Precious Metals1							0.4976	0.2700
Agriculture1								
Agriculture2								
	Bond1_1	Equity1_1	Interest Rate1_1	Energy1_1	Base Metals1_1	Precious Metals1_1	Agriculture1_1	Agriculture2_1
Bond1				0.3356			−0.4700	
Equity1								
Interest Rate1					−0.4144			
Energy1							0.4019	
Base Metals1		0.1625						
Precious Metals1			0.2012					
Agriculture1		0.2082					0.3029	
Agriculture2								

Table 12
Estimated t-stats for the SVAR system, 95% confidence level (quarterly)

	Bond1	Equity1	Interest Rate1	Energy1	Base Metals1	Precious Metals1	Agriculture1	Agriculture2
Bond1			6.0486			3.9506		
Equity1			−5.4055					
Interest Rate1								
Energy1							2.4244	
Base Metals1						7.4297		
Precious Metals1							6.5708	2.3115
Agriculture1								
Agriculture2								
	Bond1_1	Equity1_1	Interest Rate1_1	Energy1_1	Base Metals1_1	Precious Metals1_1	Agriculture1_1	Agriculture2_1
Bond1				2.5193			−3.6139	
Equity1								
Interest Rate1					−2.7663			
Energy1							3.4968	
Base Metals1		3.6463						
Precious Metals1			2.3254					
Agriculture1		2.5710					2.3383	
Agriculture2								

Table 13
Adjusted R-squares, orobability of F-stat and standard deviation of residuals (quarterly)

Logical Subset	adj-R2	pF	Std_e
Bond1	0.6084	1.57E-09	1.5662
Equity1	0.3654	2.00E-06	2.6630
Interest Rate1	0.1195	8.02E-03	1.7078
Energy1	0.3327	2.79E-05	1.5791
Base Metals1	0.6048	1.26E-10	1.0267
Precious Metals1	0.5054	8.90E-08	1.1011
Agriculture1	0.1829	3.26E-03	1.8777
Agriculture2	0.0000	0.00E+00	0.0000

simulating the greater number of possible scenarios of flow propagation.

While the number of arcs is smaller than in the previous two cases, effects are noticeably stronger. Commodities are still very highly interconnected. Agriculture1 impacts Energy1 prices with contemporary and lagged effects, a link explained by the Corn-Ethanol mechanism. Table 14 compiles the degree centrality for each vertex across the three horizons. Degree centrality decreases for most vertices as the horizon increases, evidencing a more diffused structure.

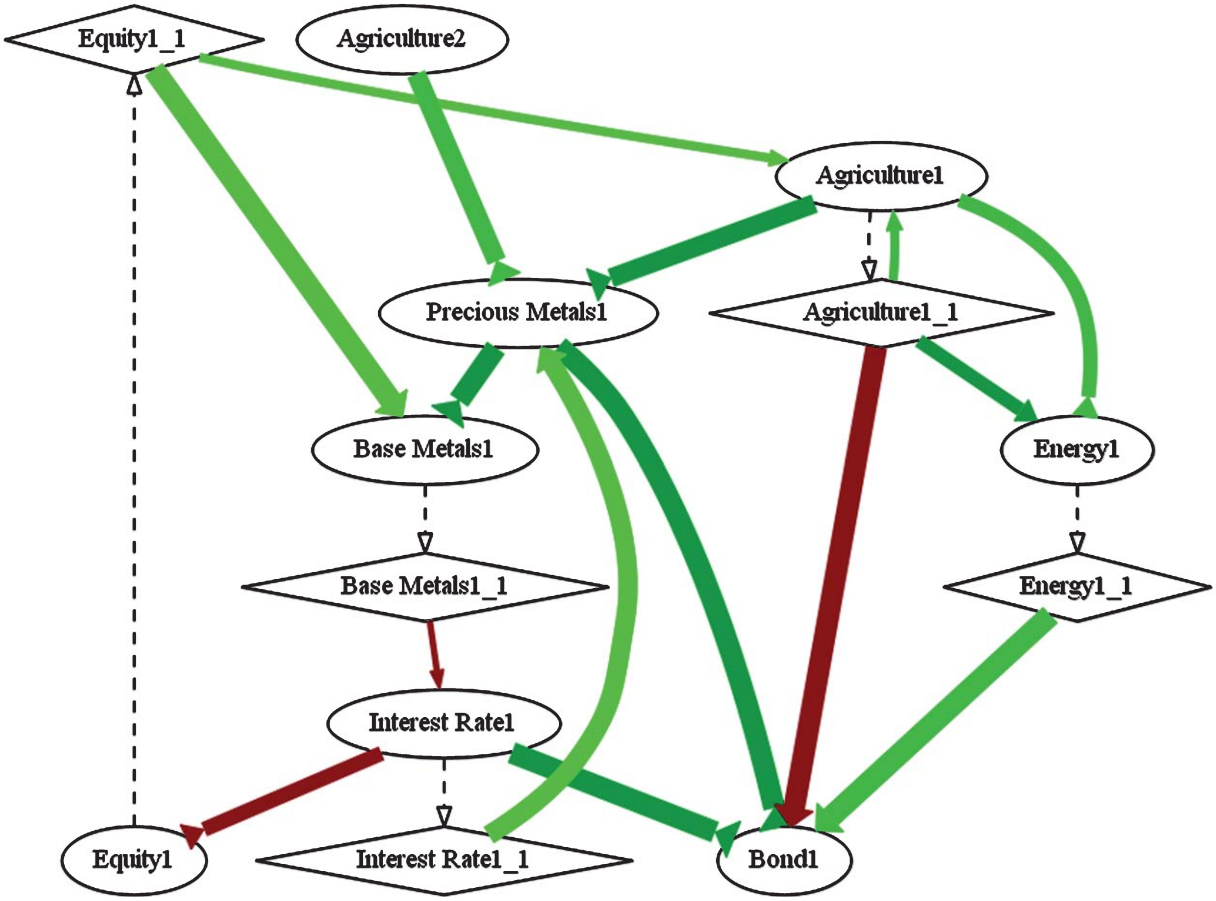

Fig. 17. SFD representation of quarterly flows.

7. Shock simulations

In this section we will illustrate how SFDs can help visualize the outcomes of random shocks to the financial system. Simulating shocks has two important practical applications: i) Given an investment portfolio, we can stress-test it by simulating various shock scenarios, introducing some prior probabilities associated with them. ii) Should a particular shock occur, we can simulate the likely propagation path and derive future equilibrium levels. When those derived future equilibrium levels significantly differ from the current ones, a Macro trading opportunity exists.

Suppose that the Energy1 asset class suffers a 5 standard deviation negative shock on monthly terms (that is, a shock injected into the system through $\varepsilon_{i,0} = -5\sigma_i$, where σ_i is the standard deviation of $y_{i,t}$ and i is the index associated with Energy1). How would the monthly flows system respond? Figures 18–21 illustrate how flows propagate on the four following months.

Table 14
Degree centrality per vertex

	Weekly	Monthly	Quarterly
Equity1	0.6923	0.3571	0.1667
Agriculture1	0.6154	0.5000	0.4167
Precious Metals1	0.5385	0.4286	0.4167
Bond1	0.5385	0.2143	0.3333
Energy1	0.4615	0.3571	0.2500
Base Metals1	0.3846	0.5714	0.2500
Agriculture2	0.3077	0.2143	0.0833
Interest Rate1	0.3077	0.4286	0.3333
Bond1_1	0.2308	0.1429	
Interest Rate1_1	0.2308	0.2143	0.1667
Equity1_1	0.1538	0.1429	0.2500
Agriculture1_1	0.1538	0.1429	0.3333
Energy1_1	0.1538		0.1667
Precious Metals1_1	0.1538	0.1429	

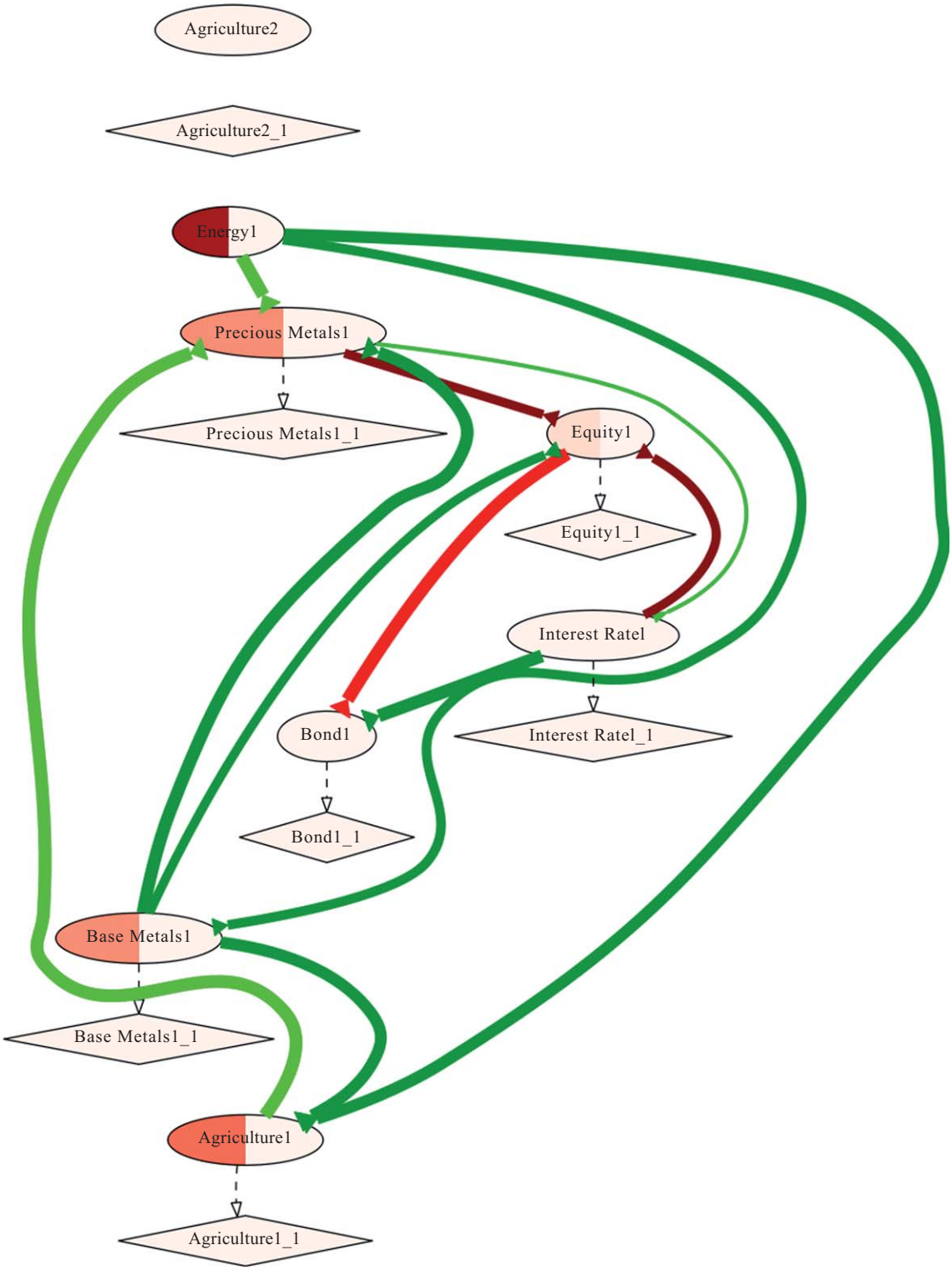

Fig. 18. Response of the financial system to a shock in the Energy1 asset class, after 1 month.

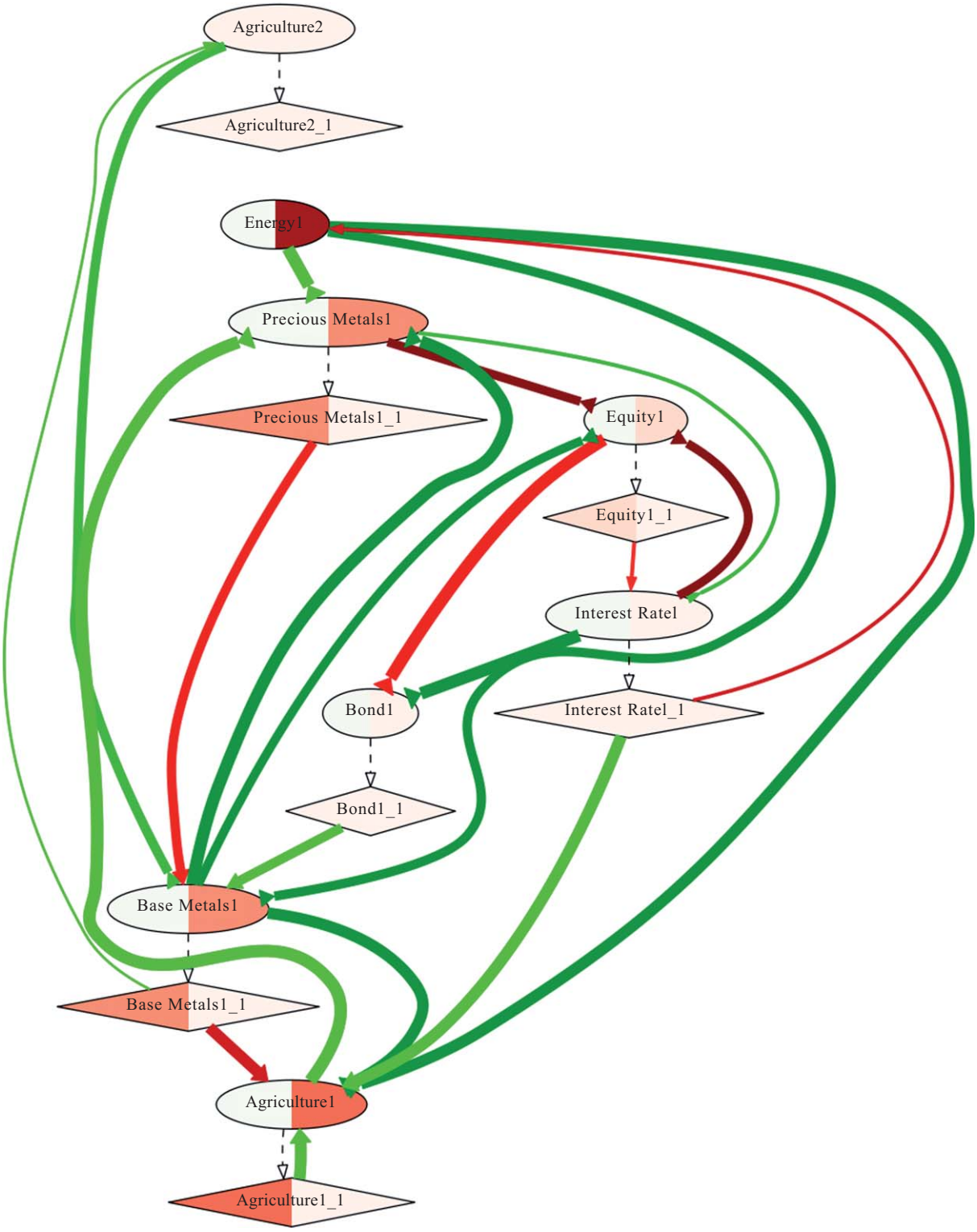

Fig. 19. Response of the financial system to a shock in the Energy1 asset class, after 2 months.

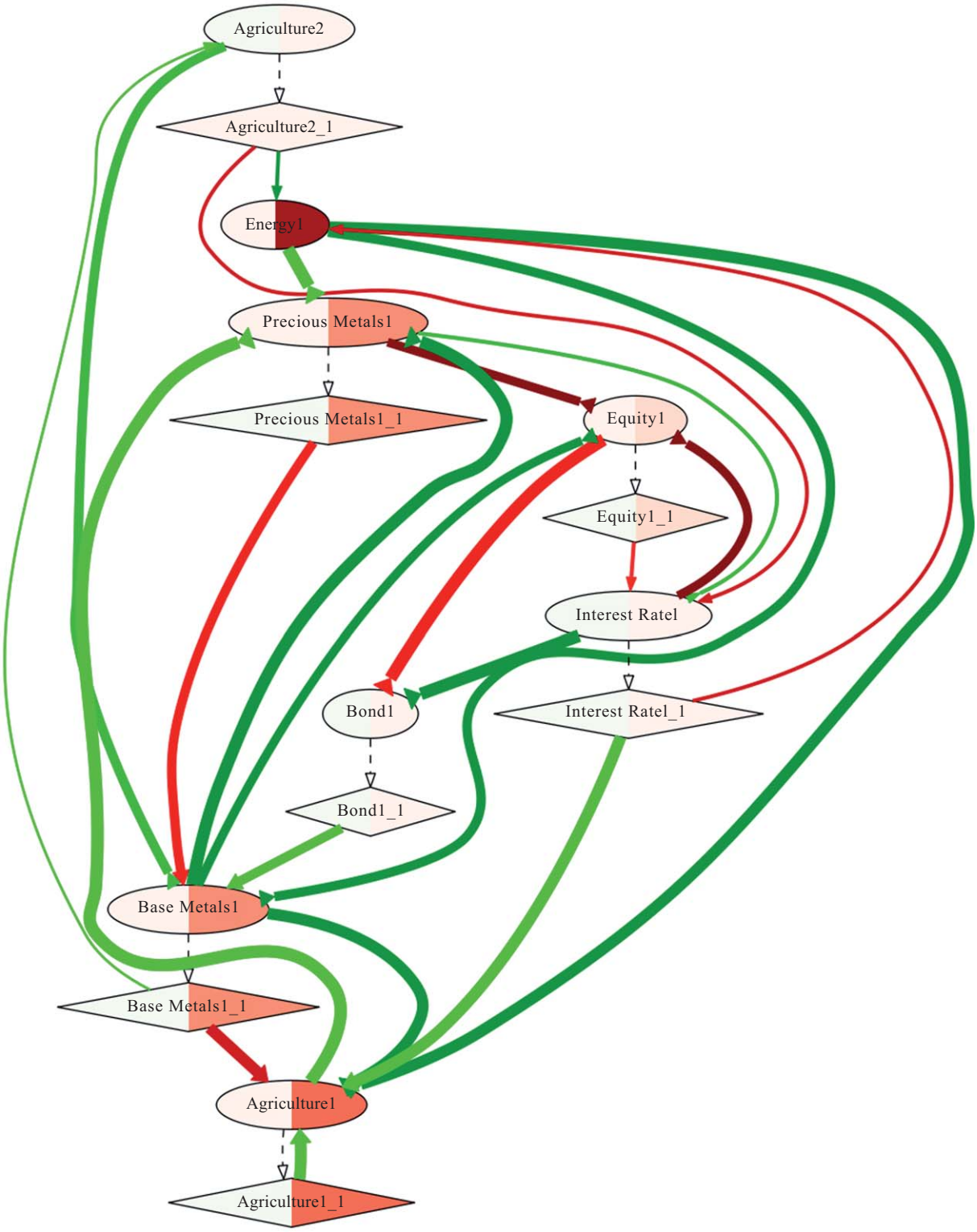

Fig. 20. Response of the financial system to a shock in the Energy1 asset class, after 3 months.

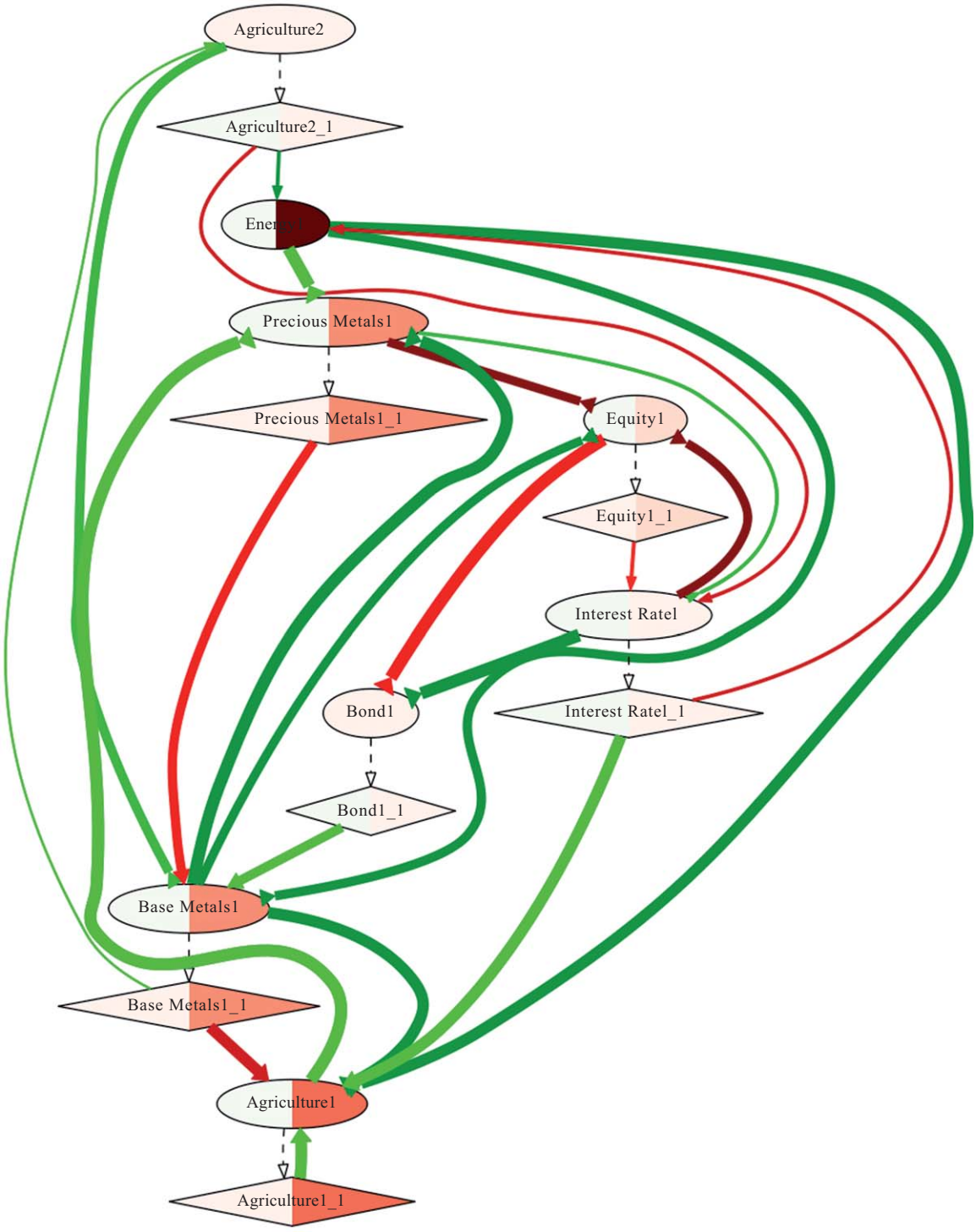

Fig. 21. Response of the financial system to a shock in the Energy1 asset class, after 4 months. For clarity, arcs have been removed when the vertex at the source had zero value.

The immediate effect is deflationary on Precious Metals1, Equity1, Base Metals1 and Agriculture1 (upper left figure). In the second month, as the initial shock propagates through the system, it reverberates back into Energy1, with a stabilizing effect (upper right figure). By the third month, Energy1, Precious Metals1, Base Metals1 and Agriculture1 prices have been negatively impacted (lower left figure), however the system has reached a stable equilibrium. No significant changes occur in the fourth month (lower right figure).

This simulation exemplifies why pairwise correlations are spurious and unstable when estimated on financial variables. An analyst may compute the correlation between Energy1 and Equity1, and conclude that it is positive (a negative shock on the former had an instantaneous negative effect on the latter, see Fig. 18). The reality is, it is not possible to isolate the effect that Energy1 has on Equity1. There is no direct connection between Energy1 and Equity1, and the former affects the latter via intermediaries. A correlation or a VAR(1) model on Energy1 and Equity1 may attempt to isolate the relationship between both variables, but the results may be spurious and misleading, because it will miss the effects that other variables had on these two. A full system of equations is required to simulate the dynamics of Macro financial flows.

Suppose now that the Agriculture1 asset class suffers a 5 standard deviation negative shock on monthly terms. Figures 22–25 illustrate how flows propagate on the four following months.

There is a mild negative initial effect on Precious Metals1, and a positive effect on Equity1 (upper left figure). In the second month, the shock further propagates, reaching Energy1 and Base Metals1, both of which are positively impacted. The lagged vertex Agriculture1_1 feeds the negative shock back to Agriculture1 (upper right figure). By the third month, the shock has spread through the entire system, and reverberation stalls (lower left figure). By the fourth month, a new equilibrium has been reached. Agriculture1, Precious Metals1 and Bond1 have been negatively impacted, while Equity1, Base Metals1, Energy1 and Agriculture2 are positively impacted (lower right figure). The reason for this divergent behavior of Agriculture1 and Agriculture2 can be found in Table 10. Agriculture1 is concentrated on grains, while Agriculture2 is dominated by Live Cattle and Lean Hogs. A shock to the former has triggered a rebalance into the latter.

Suppose now that the Interest Rate1 asset class suffers a 5 standard deviation negative shock on monthly terms. Figures 26–29 illustrate how flows propagate on the four following months.

Contemporaneous effects appear in Equity1 and Bond1, respectively bullish and bearish (upper left figure). In the second month, positive effects spread to Energy1 and Precious Metals1. The original negative shock feeds back to Interest Rates1 through Equity1_1 (upper right figure). The trend continues during the third month, at a smaller rate of change (lower left figure). By the fourth month, a new equilibrium has been reached (lower right figure). The end effect is negative on Interest Rate1 and Bond1, while positive on Equity1, Energy1 and Precious Metals1. The rest of asset classes are essentially unchanged by this shock. This further illustrates the point that financial assets form a sub-system of their own, with little influence on hard assets.

8. Simulating structural changes

The empirical findings presented in Sections 6 and 7 are the result of a historical analysis. We have used time series to uncover the past structure of the financial system. We could assume that this structure will remain essentially unchanged in the near future, in which case we can run simulations on the estimated parameters, as we did in Section 7. However, this structure is likely to evolve over time. Nothing in the SFD methodology prevents us from starting with an estimated system, which is then modified by the researcher in order to incorporate his fundamental beliefs regarding the future organization of the financial system. The practical applications include the stress-testing and Macro trading uses mentioned in Section 7.

Suppose that, for fundamental reasons, we believe that Agricultural prices will experience another 5 standard deviation negative shock like the one simulated in Section 7, only this time we believe that the sign of the inbound arc from Agriculture1_1 to Agriculture1 will be flipped (negative, instead of positive). How will the system respond to this shock? Figures 30–34 report the results.

At the end of the first month, the effect is exactly the same to what we saw in Fig. 22–25 (upper left figure). Results begin to diverge during the second month, due to the opposite reverberation we imposed when we flipped the sign of the inbound arc. Now Agriculture1 receives a positive change from the negative

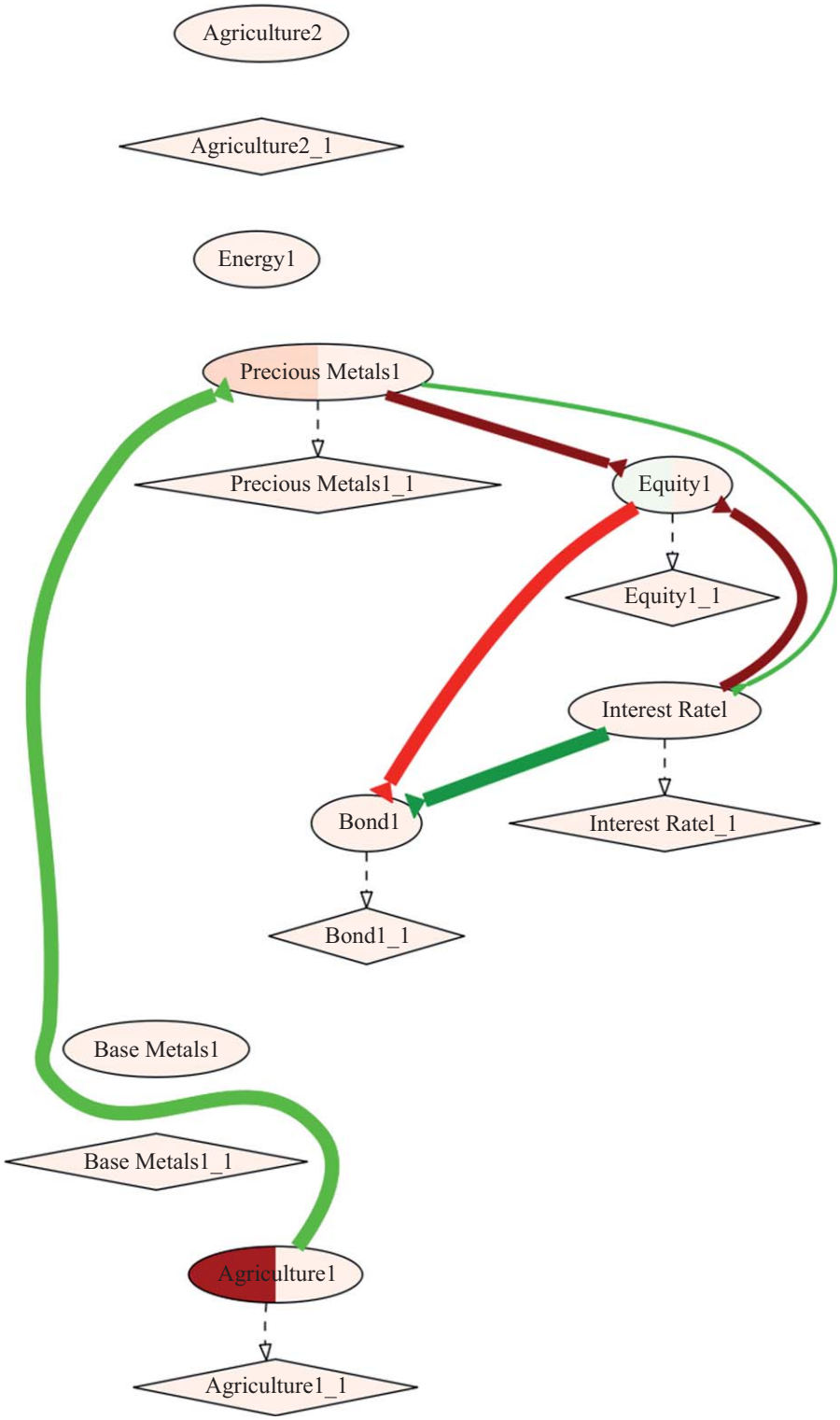

Fig. 22. Response of the financial system to a shock in the Agricultural1 asset class, after 1 month.

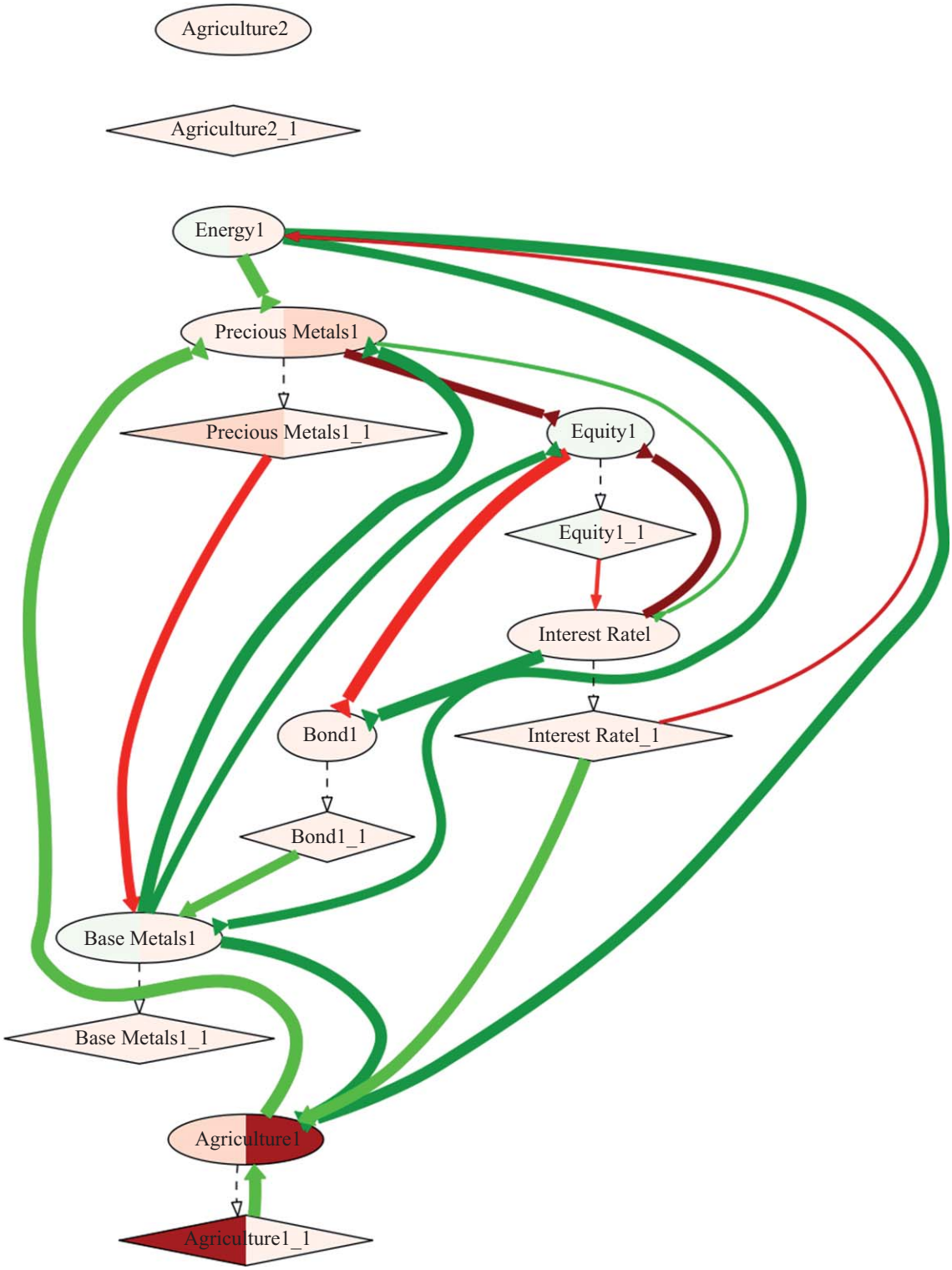

Fig. 23. Response of the financial system to a shock in the Agricultural1 asset class, after 2 months.

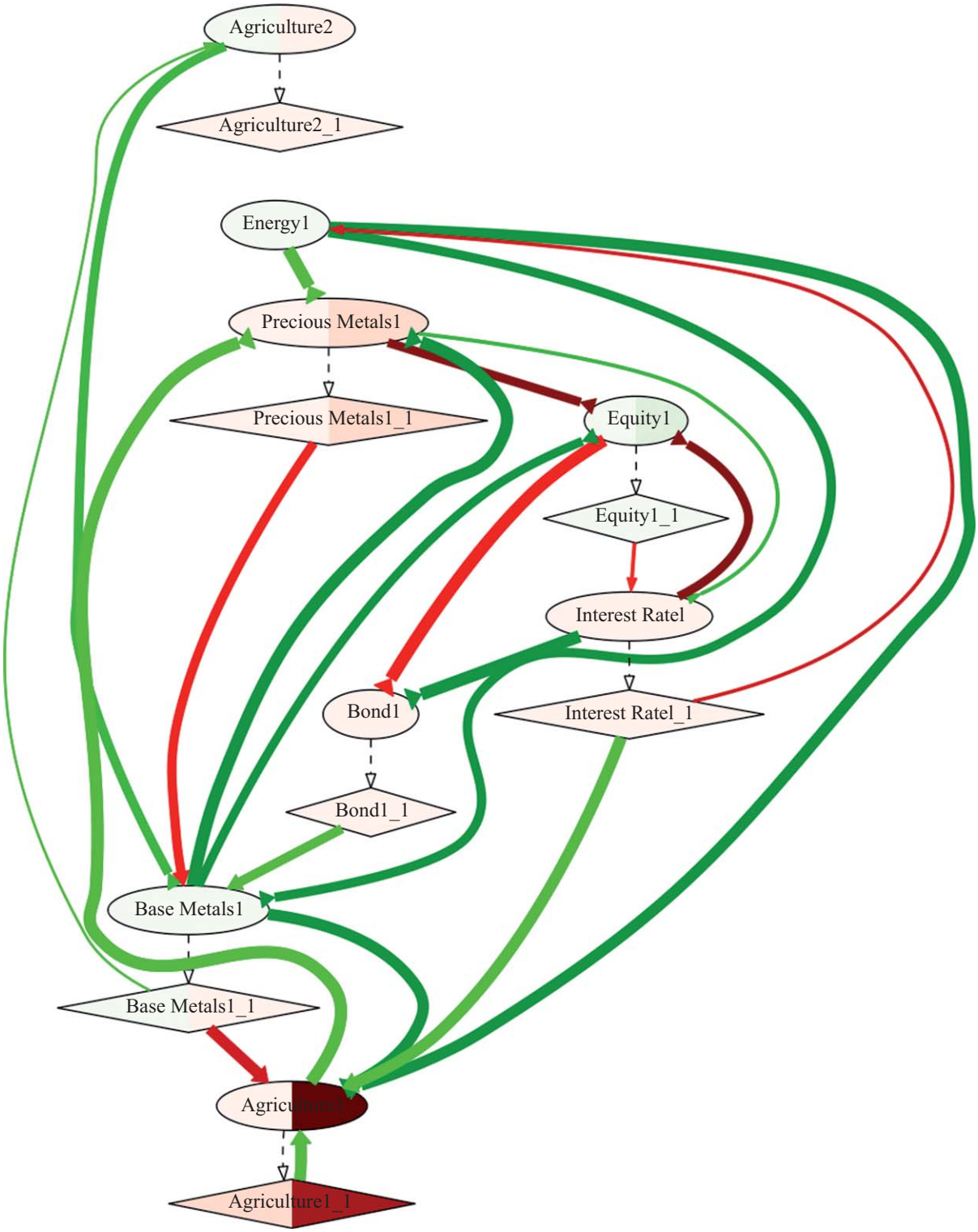

Fig. 24. Response of the financial system to a shock in the Agricultural1 asset class, after 3 months.

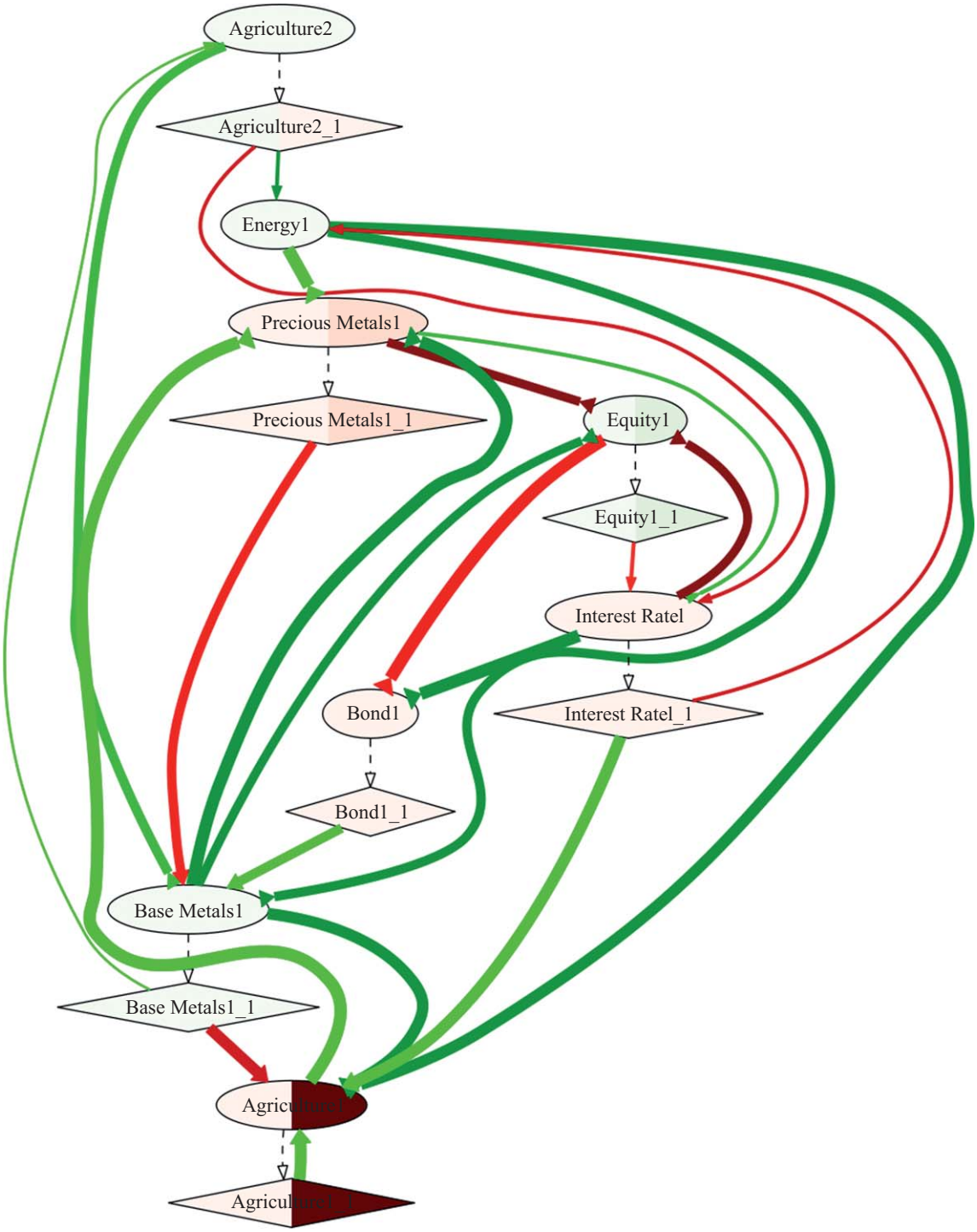

Fig. 25. Response of the financial system to a shock in the Agricultural1 asset class, after 4 months. For clarity, arcs have been removed when the vertex at the source had zero value.

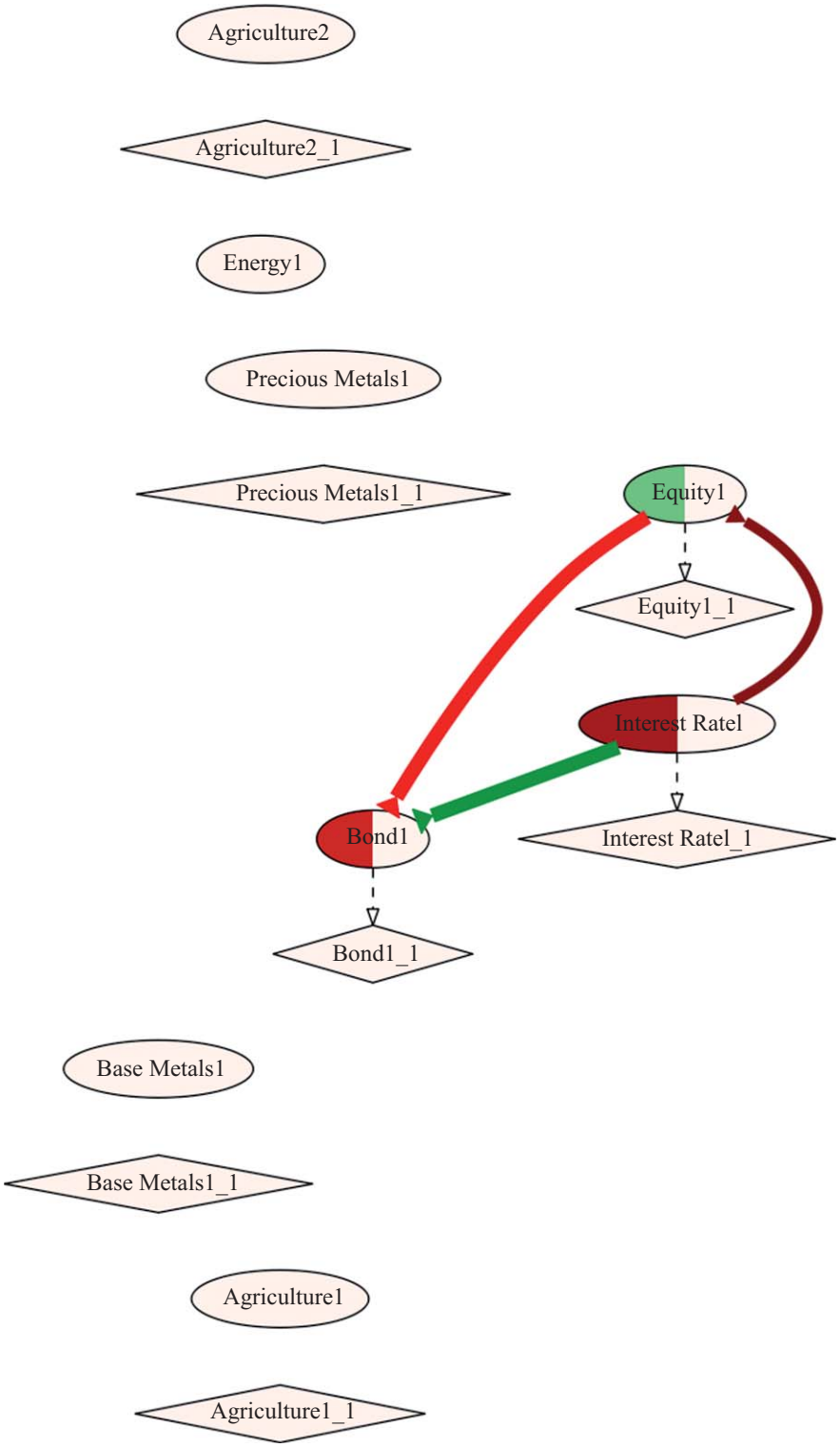

Fig. 26. Response of the financial system to a shock in the Interest Rate1 asset class, after 1 month.

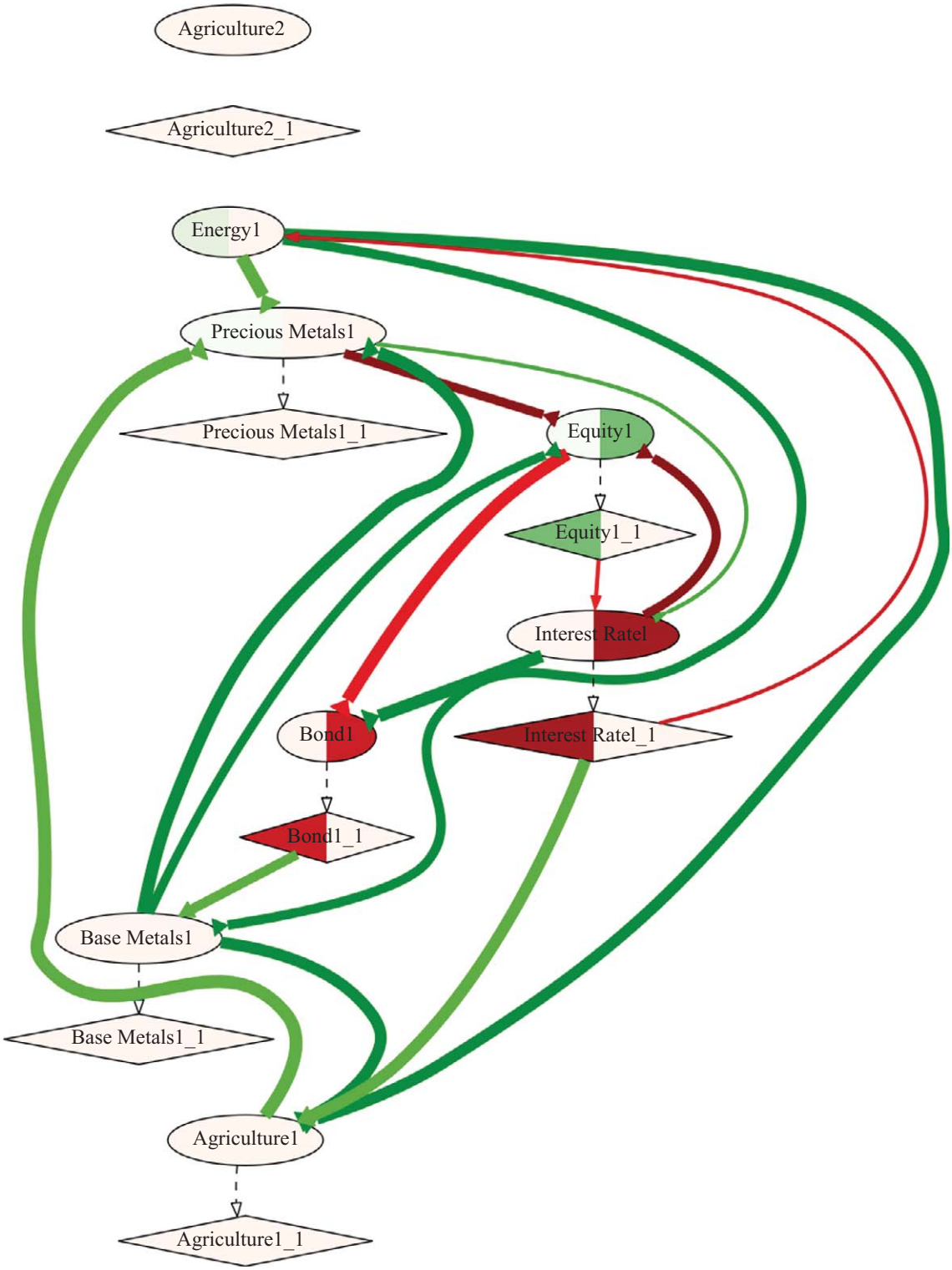

Fig. 27. Response of the financial system to a shock in the Interest Rate1 asset class, after 2 months.

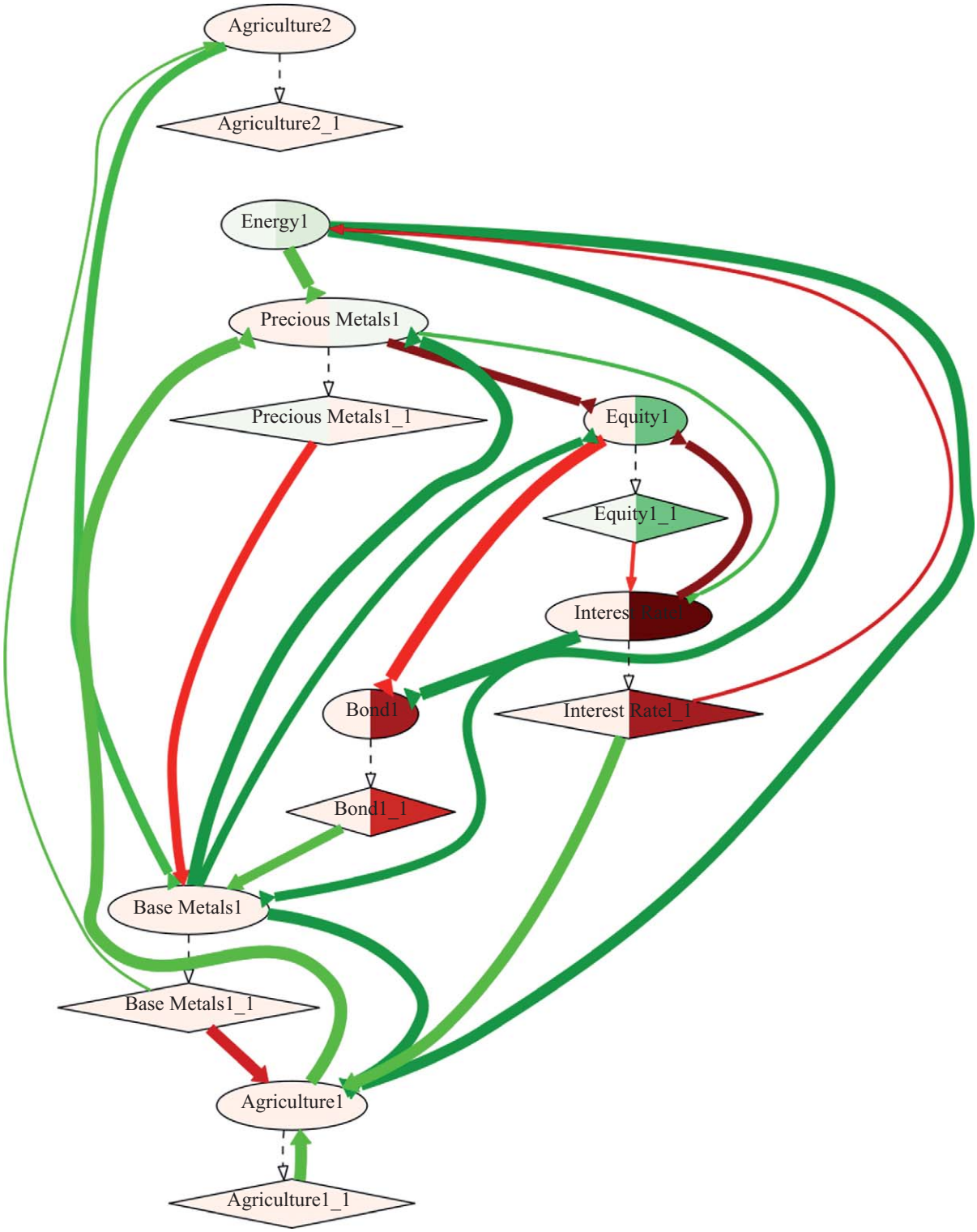

Fig. 28. Response of the financial system to a shock in the Interest Rate1 asset class, after 3 months.

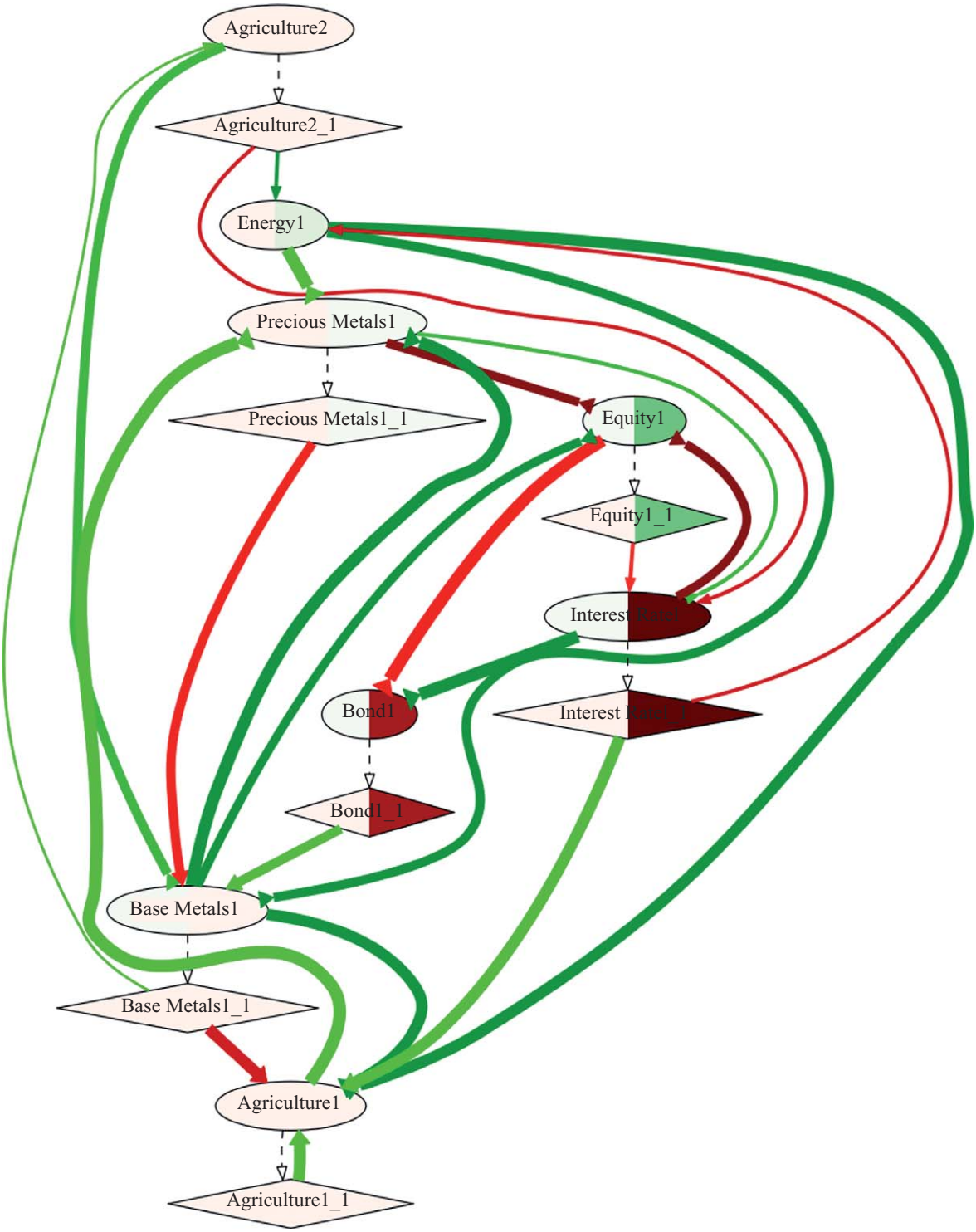

Fig. 29. Response of the financial system to a shock in the Interest Rate1 asset class, after 4 months. For clarity, arcs have been removed when the vertex at the source had zero value.

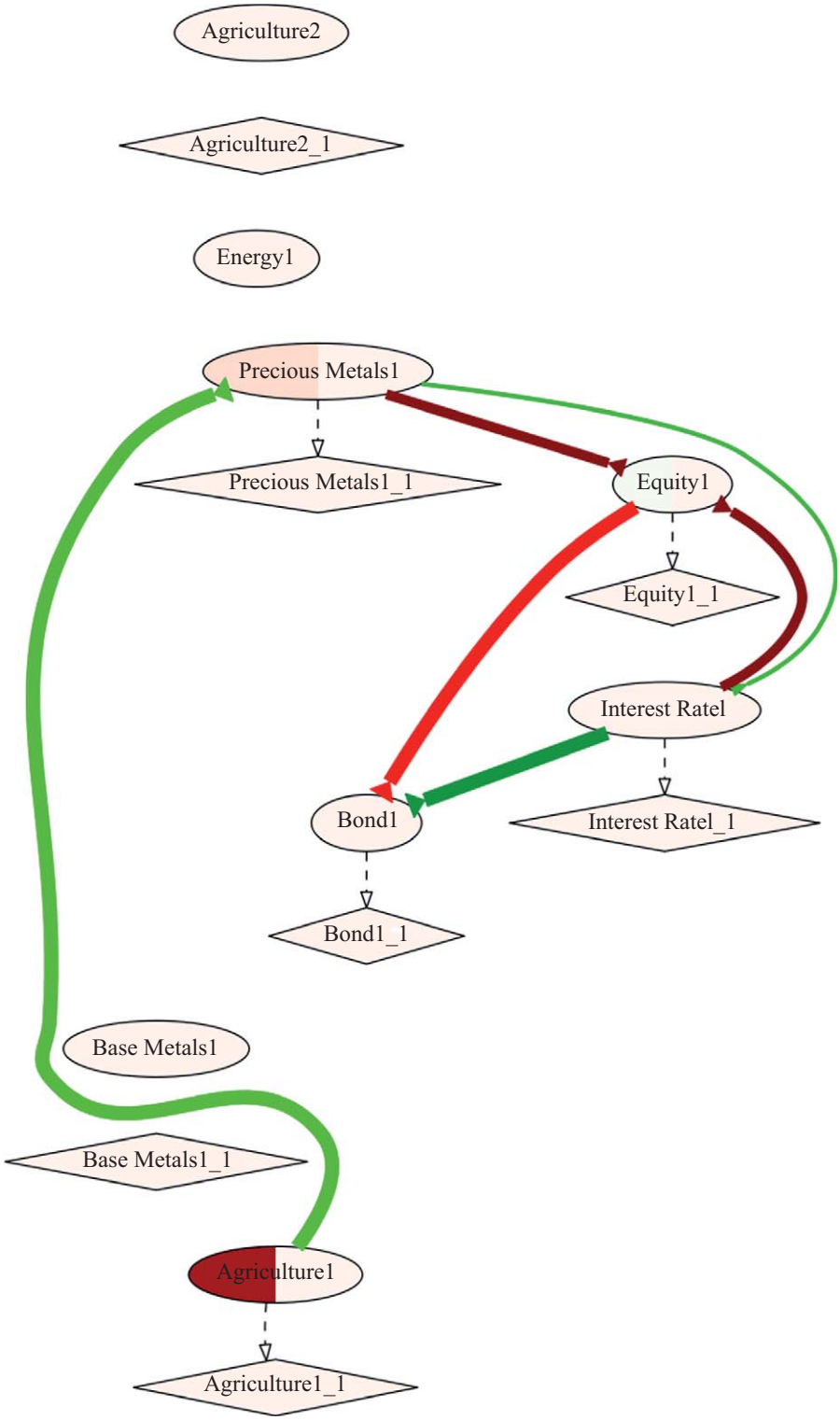

Fig. 30. Response of the financial system to a shock in the Agricultural1 asset class, with structural change, after 1 month.

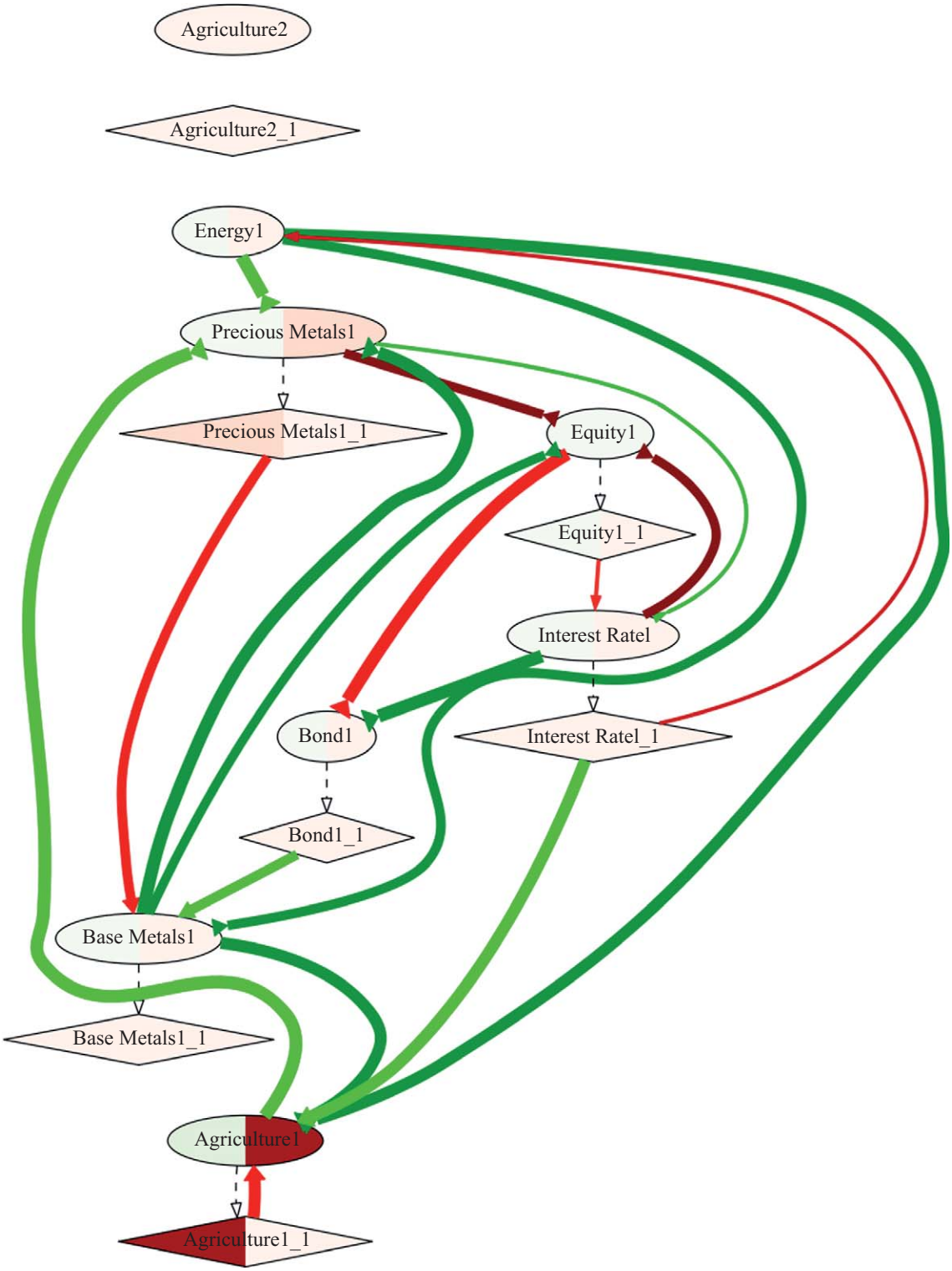

Fig. 31. Response of the financial system to a shock in the Agricultural1 asset class, with structural change, after 2 months.

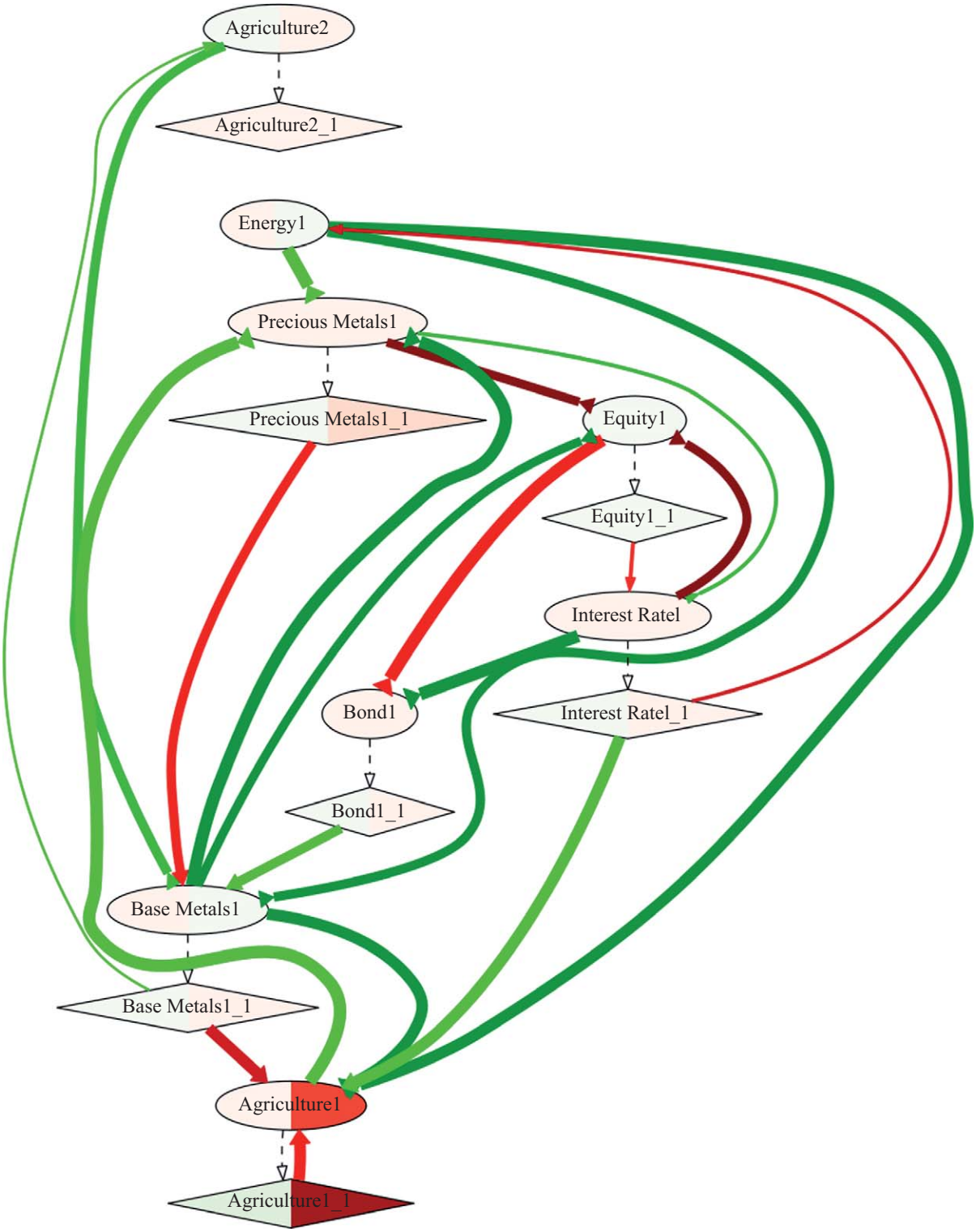

Fig. 32. Response of the financial system to a shock in the Agricultural1 asset class, with structural change, after 3 months.

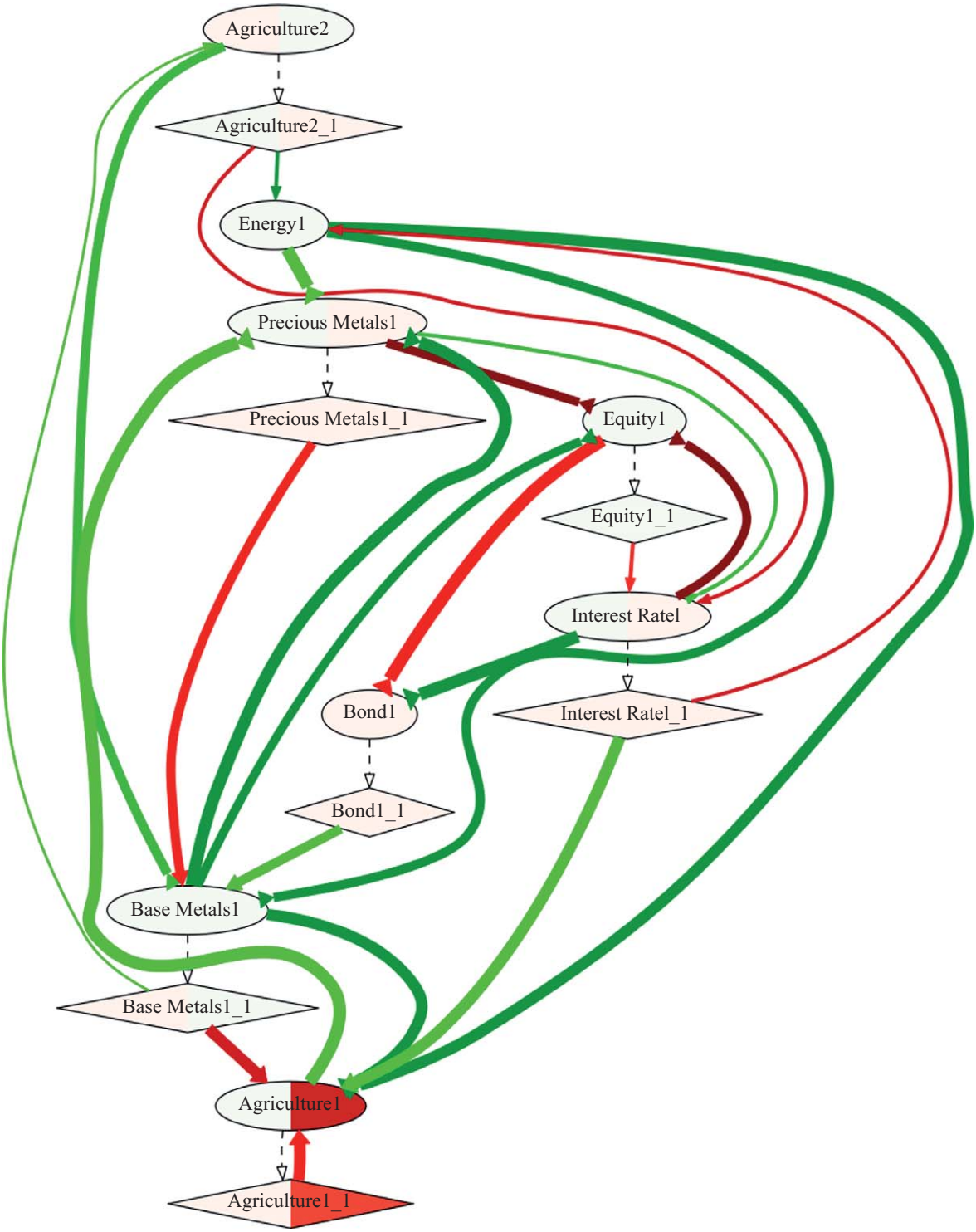

Fig. 33. Response of the financial system to a shock in the Agricultural1 asset class, with structural change, after 4 months. For clarity, arcs have been removed when the vertex at the source had zero value.

inbound arc coming from Agriculture1_1, mitigating the initial shock (upper right figure). During the third month, Agriculture1 receives a negative change from Agriculture1_1, but by now this is too weak to make a difference (lower left figure). The equilibrium is reached around the fourth month (lower right figure). The end result is that the shock has had a very limited impact overall, even on Agriculture1 itself.

9. Conclusions

In 1735 Euler showed that Geometry could not answer simple questions regarding a system's organization, hence founding the subjects of Topology and Graph Theory. We can apply the same reasoning to argue that Econometrics, which is anchored in Linear Algebra and Stochastic Calculus applications (and thus Euclidian Geometry), will not suffice to study Economic and Financial systems.

> *The assumptions that underlie standard Econometric models are relatively simplistic, which may explain why Finance and Economics have not made much progress over the last few decades. Put bluntly, Economists have already learned as much as could possibly be learned from inverting a covariance matrix.*

Recent Mathematical discoveries have revolutionized every aspect of the Sciences, and yet very few have permeated into Economics and Finance. A case in point is the technologies used by Google, Amazon, Visa, American Express, IBM Watson, etc. to forecast demand and manage inventory. Economists are largely ignorant of those mathematical tools, which have obvious applications to their field of study. Instead, much effort is dedicated to fight over perennial controversies, like the one between Keynesians and the Austrian school, using the same tools that have failed to provide a conclusion over the last 70 years. Modern mathematical techniques which are applicable to a wide spectrum of problems encountered by finance professionals include Machine Learning, Graph Theory and Game Theory.

In this paper we apply *Stochastic Flow Diagrams* (SFDs), a novel mathematical methodology that can help us visualize the complex network of flows that constitute the global financial system. SFDs are topological representations of a system of equations in differences, such as those encountered in time series models. Our method combines elements of Graph Theory and inferential statistics to visualize the structure of a complex system, allowing for an intuitive interpretation of its state and future course. The SFD method takes into consideration the dynamic properties of the system, determining the direction of the flows in terms of lead-lag and causality effects. SFD connectivity is determined by statistical significance of the graph's arcs, which are weighted based on the flow carried through the arcs involved. Because SFD maps a dynamic system, it incorporates a time dimensionality, where crossing each arc represents a unit of time elapsed.

Visualization techniques have proven extremely valuable in advancing theoretical research in scientific disciplines, particularly in theoretical Physics. Easier visualization can help policy makers, investors and economists describe their models without having to resort to lengthy formulaic representations. They can help hedge fund managers identify early trends, Central Bankers design intervention tools, policy makers detect regime changes, among many other applications. As it relates to global macro trading, they can help monitor how financial flows propagate across various investment assets.

We have shown that a small number of SFD attributes is able to describe a wide range of time series systems. In the case of Macroeconomics, SFDs allow researchers to monitor flow dynamics, as the values adopted by the random variables change over time. We have also seen that estimating pairwise relationships between variables within a system leads to spurious results. The reason is, understanding the behavior of any single system constituent requires the understanding of the whole. The implication is that pairwise relationships estimated on financial variables, such as correlations, will often yield spurious results.

One important practical application of this approach is to be able to visualize the state of the system, and anticipate the possibility of overflows. Another application is the possibility of simulating the propagation of shocks emanating from alternative vertices. The use of SFDs is not limited to historical estimates. Researchers can incorporate their priors and simulate the dynamics of the system, even if these priors are forward looking scenarios rather than historical ones. We believe that SFDs can prove valuable in enlightening perennial controversies in Macroeconomics, such as the one involving Keynesians and Austrian-school economists.

Acknowledgments

We thank the Managing Editor of Algorithmic Finance, as well as three anonymous referees for their suggestions and support. We are indebted to Alberto Musalem (Federal Reserve Bank of New York) for his insightful comments. We are also grateful to Tony Anagnostakis (Moore Capital), David H. Bailey (University of California, Davis), José Blanco (UBS), Jonathan M. Borwein (University of Newcastle, Royal Society of Canada), Peter Carr (Morgan Stanley, NYU), Marco Dion (J.P. Morgan), David Easley (Cornell University), Matthew D. Foreman (University of California, Irvine), Jon Kleinberg (Cornell University), Jeffrey Lange (Guggenheim Partners), Frank Leitner (Tudor Investment Corporation), Attilio Meucci (KKR, NYU), Riccardo Rebonato (PIMCO, University of Oxford) and Luis Viceira (HBS).

Disclaimer

The opinions expressed by the authors of this article do not necessarily reflect the views of Berkeley Lab or Guggenheim Partners. No investment advice of particular course of action is recommended by this article.

Appendices

A.1. Distance and variance-stabilizing transformations

Suppose two random variables *x,y* distributed as a bivariante Normal with correlation $\rho[x, y]$. The estimated correlation coefficient $\hat{\rho}[x, y]$ is computed over a sample of T independent and identically distributed (IID) observations. A distance measure is then defined as $d_{x,y} = \sqrt{2(1 - \hat{\rho}[x, y])}$, where $d_{x,y} \in [0, 2]$ is a real value. This distance is used to create graph connections between elements x, y if and only if $d_{x,y} < \bar{d}$, where $\bar{d}$ is a user-defined threshold.

One issue with this non-linear transformation is that it is not variance-stabilizing. This means that the variance of $d_{x,y}$ is not constant for all values of $\hat{\rho}[x, y]$. In particular, the variance of $d_{x,y}$ is much smaller for higher values of $\hat{\rho}[x, y]$. This is a problem, because x and y are only considered connected if $d_{x,y}$ is below a given threshold $\bar{d}$, where the estimation error of $d_{x,y}$ is not constant. The implication is that the number of spurious connections grows ever faster as $\bar{d}$ is increased.

Fisher (1915) proposed a transformation of the correlation estimate that is variance-stabilizing. Fisher's transformation tells us that

$$f[\hat{\rho}, T] \sim \mathcal{N}\left(\frac{1}{2} Ln\left[\frac{1+\rho}{1-\rho}\right], \frac{1}{T-3}\right) \tag{7}$$

with an inverse

$$f^{-1}\left[E[f[\hat{\rho}, T]]\right] = \frac{e^{2f} - 1}{e^{2f} + 1} \tag{8}$$

returning $\hat{\rho}$. Note that the variance of $f[\hat{\rho}, T]$ is constant at $\frac{1}{T-3}$, regardless of the value of ρ (thus the variance-stabilizing property of Fisher's transformation). We can apply Fisher's transform on $d_{x,z}$ and see how this distance's estimation errors impact the acceptance of connections:

1. For a given distance threshold $\bar{d}$, compute the implied correlation $\hat{\rho} = 1 - \frac{\bar{d}^2}{2}$.
2. Compute $E\left[f[\hat{\rho}, T]\right] = \frac{1}{2} Ln\left[\frac{1+\hat{\rho}}{1-\hat{\rho}}\right]$ and $V\left[f[\hat{\rho}, T]\right] = \frac{1}{T-3}$.
3. Given a significance level $\alpha < \frac{1}{2}$, estimate the upper confidence band as $u\left[\hat{\rho}, T, \alpha\right] = \frac{1}{2} Ln\left[\frac{1+\hat{\rho}}{1-\hat{\rho}}\right] - Z_\alpha \frac{1}{\sqrt{T-3}}$, where Z_α is the critical value of the Standard Normal associated with α.
4. Recover the correlation value associated with the upper confidence band, as $\hat{\rho}_u = f^{-1}\left[u[\hat{\rho}, T, \alpha]\right] = \frac{e^{2u}-1}{e^{2u}+1}$.

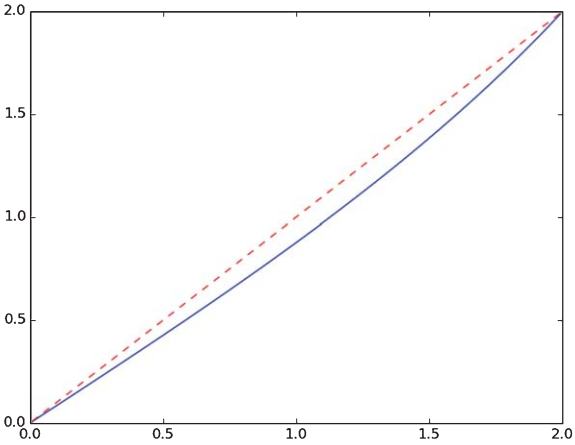

Fig. 34. Fisher's transformation on the proposed distance. The x-axis represents $\bar{d}$, while the y-axis represents $\bar{d}_u$. The dotted line applies Fisher's transformation in the ideal case where there is no estimation error ($T \to \infty$). The continuous line applies Fisher's transformation for $\alpha = 0.05$ and $T = 100$. If $d_{x,y}$ was a variance-stabilizing transformation, the gap between the two lines would be constant.

5. Recover the distance associated with the upper confidence band, as $\bar{d}_u = \sqrt{2(1-\hat{\rho}_u)}$

Results are shown in Fig. 34. The x-axis represents $\bar{d}$, while the y-axis represents $\bar{d}_u$. The dotted line applies the Fisher transformation in the ideal case where there is no estimation error ($T \to \infty$). The continuous line applies the Fisher transformation for $\alpha = 0.05$ and $T = 100$. If $d_{x,y}$ was a variance-stabilizing transformation, the gap between the two lines would be constant. However, the gap reaches a maximum around $\bar{d} = 1.22$. The implication is that this approach forms connections with varying and non-monotonic errors, depending on the selected threshold $\bar{d}$ and the number of observations T. Snippet 1 displays the Python code that reproduces these calculations.

```
#!/usr/bin/env python
# On 20131224 by lopezdeprado@lbl.gov
import numpy as np
import scipy.stats as ss
import matplotlib.pyplot as pp
from math import log,exp
#-------------------------------------------------------------------------------
def main():
    t,alpha=100,.05
    dist0=np.linspace(0.01,1.99,100)
    z=ss.norm.ppf(alpha)
    dist1=[]
    for dist0_ in dist0:
        corr0_=1-dist0_**2/2.
        u_=.5*log((1+corr0_)/(1-corr0_))-z/(t-3)**.5
        corr1_=(exp(2*u_)-1)/(exp(2*u_)+1)
        dist1_=(2*(1-corr1_))**.5
        dist1.append(dist1_)
    pp.plot(dist0,dist1)
    pp.plot(dist0,dist0,'--r')
    pp.show()
    return
#-------------------------------------------------------------------------------
if __name__=='__main__':main()
```

Snippet 1. – Fisher's transform on the proposed distance.

A.2. SFDs glossary

In this section we will summarize how to read a SFD, and justify the symbology chosen. A small number of attributes is able to represent a wide variety of system architectures. These attributes can be further expanded to signal additional features in complex dynamic systems.

A.2.1. Vertices

Vertices have two shapes:

- **Elliptical**, for a *current variable*. There is only one in AR specifications, and multiple in VAR and SVAR systems.
- **Diamond**, for a *lagged variable*.

Vertices can have two colors:

- **Red**: The variable has a negative value (darker as more negative).
- **Green**: The variable has a positive value (darker as more positive).

If the variable expresses a change in value rather than a cumulative value (e.g., in an equation in differences), then the vertex is divided in two halves, where the left half is colored according to the change in value and the right according to the cumulative value.

The label inside the vertex indicates the variable's name, and in the case of a lagged variable, also the order of the lag.

A.2.2. ARCS

Arcs can have three shapes:

- **Dashed arc line, with unfilled arrowhead**: This denotes an *outbound arc*, i.e. an arc that carries flow to a diamond-shaped vertex (lagged variable). They appear in all Time Series models. This is a convenient representation, because these arcs can only have a unit weight, hence no colour is needed to convey the weight or width to convey the goodness of the fit.
- **Solid arrow, with pointed arrowhead**: This denotes an *inbound arc* involved in a lead-lag effect (connecting a diamond vertex with an elliptical vertex), e.g. in AR processes and VAR systems.
- **Solid arrow, with flat arrowhead**: This denotes the *inbound arc* involved in a contemporaneous effect (connecting two elliptical vertices), e.g. in SVAR systems.

Inbound arcs can have two colors:

- **Red**: Negative weight (darker as more negative).
- **Green**: Positive value (darker as more positive).

The width of an inbound arc is a function of the goodness of fit associated with that equation.

References

Aldrich, J., 1995. Correlations genuine and spurious in Pearson and Yule. Statistical Science 10(4), 364–376.

Alexander, C., 1999. Correlation and Cointegration in Energy Markets. In Managing Energy Price Risk, Risk Books, 2nd Edition.

Bailey, D., López de Prado, M., 2012. Balanced baskets: A new approach to trading and hedging risks. Journal of Investment Strategies, 1(4), Fall. Available in SSRN, http://ssrn.com/abstract=2066170.

Bailey, D., Borwein, J., López de Prado, M., Zhu, J., 2013a. Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance. Working paper. Available at: http://ssrn.com/abstract=2308659

Bailey, D., Borwein, J., López de Prado, M., Zhu, J., 2013b. The Probability of Backtest Overfitting. Working paper. Available at: http://ssrn.com/abstract=2326253

Billio, M., Getmansky, M. Lo, A., Pelizzon, L., 2012. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics. 104, 535–559.

Boginski, V., Butenko, S., Pardalos, P., 2003. On structural properties of the market graph. In: Nagurney, A. (Ed.), Innovations in Financial and Economic Networks. Edward Elgar Publishers, Aldeeshot, pp. 29–45.

Boginski, V., Butenkob, S., Pardalosa, P., 2005. Statistical analysis of financial networks. Computational Statistics & Data Analysis. 48, 431–443.

Boginski, V., Butenko, S., Pardalos, P., 2006. Mining market data: A network approach. Computers & Operations Research, pp. 3171–3184.

Bollobás, B., 2013. Modern Graph Theory, Springer.

Bondy, J., Murty, U., 1976. Graph Theory with Applications, Elsevier Science.

Brualdi, R., 2010. The mutually beneficial relationship of graphs and matrices. Conference Board of the Mathematical Sciences, Regional Conference Series in Mathematics, Nr. 115.

Calkin, N., López de Prado, M., 2014. Stochastic flow diagrams. Algorithmic Finance. 3(1-2), 21–42.

Campbell, J., Lo, A., MacKinlay, A., 1996. The Econometrics of Financial Markets, Princeton University Press.

Donoho D., Johnstone, J., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika. 81(3), 425–455.

Dunham, W. 1999., Euler: The Master of Us All. Mathematical Association of America.

Durrett, R., 2007. Random Graph Dynamics, Cambridge University Press.

Easley, D., Kleinberg, J., 2010. Networks, Crowds and Markets: Reasoning about a Highly Connected World. Cambridge University Press.

Easley, D., López de Prado, M., O'Hara, M., 2011. The Microstructure of the flash crash: Flow toxicity, liquidity crashes and the probability of informed trading. Journal of Portfolio Management. 37(2), 118–128. http://ssrn.com/abstract=1695041

Easley, D., López de Prado, M., O'Hara, M., 2012. Flow toxicity and liquidity in a high frequency world. Review of Financial Studies. 25(5), 1457–1493. http://ssrn.com/abstract=1695596

Easley, D., López de Prado, M., O'Hara, M., 2013. High Frequency Trading. Risk Books. http://riskbooks.com/book-high-frequency-trading

Einstein, A., 1941. Science, philosophy and religion: A symposium. Conference on Science, Philosophy and Religion in Their Relation to the Democratic Way of Life. New York.

Fisher, R., 1915. Frequency distribution of the values of the correlation coefficient in sample of an indefinitely large population. Biometrika. 10(4), 507–521.

Foster, D., George, E., 1994. The risk inflation criterion for multiple regression. Annals of Statistics. 22(4), 1947–1975.

Franses, P., van Dijk, D., 2000. Non-linear Time Series Models in Empirical Finance, Cambridge University Press.

Granger, C., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 37(3), 424–438.

Greene, W., 2008. Econometric Analysis, Prentice Hall, 6th Edition.

Hamilton, J., 1994. Time Series Analysis, Princeton University Press.

Hocking, R., 1976. The analysis and selection of variables in linear regression. Biometrics. 32.

Jackson, M., 2010. Social and Economic Networks, Princeton University Press.

Kaiser, David., 2005. Physics and feynman's diagrams. American Scientist. 93, 156–165.

Lautier, D., Raynaud, F., 2011a. Energy derivative markets and systemic risk. Technical report, French Energy Council.

Lautier, D., Raynaud, F., 2011b. Statistical properties of derivatives: A journey in term structures. Phys. A, Stat Mech Appl. 390(11), 2009–2019.

Lautier, D., Raynaud, F., 2012a. Systemic risk in energy derivative markets: A graph theory analysis. Energy J. 33(3), 217–242.

Lautier, D., Raynaud, F., 2012b. High dimensionality in finance: The advantages of the graph theory. In: Batten J, Wagner N (eds). Derivatives Securities Pricing and Modelling, Emerald Publishing, June.

McNelis, P., 2005. Neural Networks in Finance: Gaining predictive edge in the Market, Elsevier.

Muirhead, R., 1982. Aspects of Multivariate Statistical Theory, John Wiley & Sons, New York.

Rebonato, R., Denev, A., 2014. Portfolio Management under Stress, Cambridge University Press. 1st Edition.

Sala-i-Martin, X., 1997. I just ran two million regressions. American Economic Review. 87(2), May.

Strang, G., 1993. The fundamental theorem of linear algebra. The American Mathematical Monthly. 100(9), 848–855.

Strang, G., 2010. Linear Algebra: Left and right inverses; pseudoinverse, MIT Course Number 18.06. Available at http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/lecture-33-left-and-right-inverses-pseudoinverse. Transcript available at http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/MIT18_06S10_L33.pdf

Tsay, R., 2010. Analysis of Financial Time Series, Wiley, 3rd Edition.