2

Aligning Models and Data for Systemic Risk Analysis

Roger M. Stein

Abstract The recent financial crisis has brought to the fore issues of understanding and reducing systemic risk. This focus has precipitated exploration of various methods for measuring systemic risks and for attributing systemic risk contributions to systemically important financial institutions. Concomitant with this stream of research are efforts to collect, standardize and store data useful to these modeling efforts. While discussions of modeling approaches are pervasive in the literature on systemic risk, issues of data requirements and suitability are often relegated to the status of implementation details. This short chapter is an attempt to deepen this discussion. We provide a 2×2 mapping of modeling strategies to key data characteristics and constraints that can help modelers determine which models are feasible given the available data; conversely, it can provide guidance for data collection efforts in cases where specific analytic properties are desired. The framework may also be useful for evaluating, at a conceptual level, the trade-offs for incremental data collection. To provide background for this mapping, we review the analytic benefits and limitations of using aggregate vs. micro-level data, provide background on the role of data linking and discuss some of the practical aspects of data pooling including concerns about confidentiality. Throughout the chapter, we include examples from various domains to make the points we outline concrete.

2.1 Introduction

Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit

[a] The views expressed in this article are the author's and do not represent the views of current or former employers (Moody's Corporation, Moody's Research Labs, Moody's KMV, Moody's Investors Service) or any of their affiliates. Accordingly, all of the foregoing companies and their affiliates expressly disclaim all responsibility for the content and information contained herein.

> is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination; that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue.
>
> R. A. Fisher

This chapter discusses some practical considerations in data collection and model development for systemic risk analysis. We focus on *data characteristics* and *data requirements* for risk models and how these can influence options for research on and deployment of systemic-risk analytics.

Recent discussions of systemic risk have stimulated increased research focus on various modeling approaches. However, modelers often relegate issues of data requirements and suitability to the status of implementation details. At the same time, concerns about systemic risk have also led to governmental and industry efforts to collect and store data. In these cases, the focus has often been on standardization and operations rather than on the analytic tools that might be developed from this data (though this is the ultimate objective)[1]. From a practical perspective, there is a striking disconnect between the push for data collection on the one hand and the relative lack of attention on the part of modelers to the mechanics and semantics of the data required to build effective models on the other.

Plans for large-scale data collection efforts are emerging, and in some cases, these efforts are already underway. When such projects come to fruition, a much richer view of systemic relationships and risks will be possible than has been before. However, today, it is often difficult for individual organizations even to integrate data from different divisions within *their own* firms, let alone to integrate data across multiple firms.[2]

It will take time and effort to create practical standards for assembling larger data sets for research on the broader financial system, and to then implement those standards fully enough to permit data pooling. In the interim, productive, if incomplete, data collection and systemic-risk modeling projects are feasible in advance of the more fulsome solutions. These more modest efforts can begin immediately.

As an example to motivate our discussion, consider how we might answer the following question:

[1] In our discussions of the analytic properties of data, we are not focusing on *domain knowledge* relating to how, e.g., options are traded or how the sinking fund on a municipal bond works. While this type of domain expertise is required to ensure that complete and useful data is collected and stored, our focus is on a different type of domain knowledge – that relating to the statistical and computational operations that different types of data permit or preclude and what the minimum data requirements are for specific analytic applications.

[2] This is not unique to financial institutions. For example, in a 2010 survey of 443 senior finance executives at large firms (CFO Publishing LLC, 2011), respondents reported that the single most common challenge to improving the effectiveness of the company's finance function was that the IT systems at their firms were outdated, inflexible or unable to support new requirements. 43% of respondents identified IT system limitations in this context. Respondents were not only drawn from the financial services industry. The sample included a cross-section of commercial sectors.

How would the default of a US state on its general obligation (GO) debt[3] *affect large financial institutions?*

This question might arise in the context of a stress test[4] or as part of a more focused analysis on the part of a regulator or policy-maker who may have detailed information on a specific possible scenario.

Financial institutions have many forms of exposure to municipalities. They may, for example, hold large positions in municipal bonds or be susceptible to ripple effects from shocks to the tax-exempt sector that would inevitably follow such a default.

For purposes of our simple example, we will consider only one very limited form of exposure that a financial institution (FI) might have to US municipalities: financial guarantees provided to municipalities at the state or local level. Such guarantees commonly take the form of letters of credit, backstop liquidity agreements or bond insurance. A default on the part of one of the bond issuers would cause the draws on the guarantees, which would deplete the capital of FI. Network analysis (Chapter xx) offers one approach to investigating this question. Thought this is only one form of exposure, it is useful in demonstrating some of our themes without requiring extensive background. (Later we will give examples of more robust stress tests.)

One strategy for assembling data to answer the question of the impact on of a default on draws on FI facilities would be to examine the exposures of each financial institution to each state and local municipality. However, this could be wasteful since there are many financial institutions and most do not provide guarantees to most municipalities. It would also require the collection of detailed portfolio data for each financial institution.

An alternative, more efficient design would start with the state and local bonds, examining all counterparties that provide guarantees to them, along with the size of the exposure. Once these were properly identified and linked, the exposures could be aggregated and a first order approximation to the network of exposures generated from this.

An example of such an analysis is shown in Figure 2.1.

It is instructive to note that the data requirements for this analysis were sparse. To calculate all quantities we required only:

- CUSIPS of all GO bonds

[3] A *general obligation* bond is one that derives its credit quality from the ability of a municipality to make timely payments of principal and interest from its own cash flows. These cash flows typically come from tax receipts and other sources of governmental income. In contrast, other forms of municipal debt, e.g., revenue bonds, are typically issued to fund specific projects such as bridge construction, etc. These bonds are not generally backed by the issuing municipality, but instead rely on revenues from the underlying project for repayment.

[4] This might be similar in flavor; for example, to the $10 \times 10 \times 10$ approach proposed in Duffie (2010) or it could be part of a "thought experiment" stress scenario as described in Stein (2012).

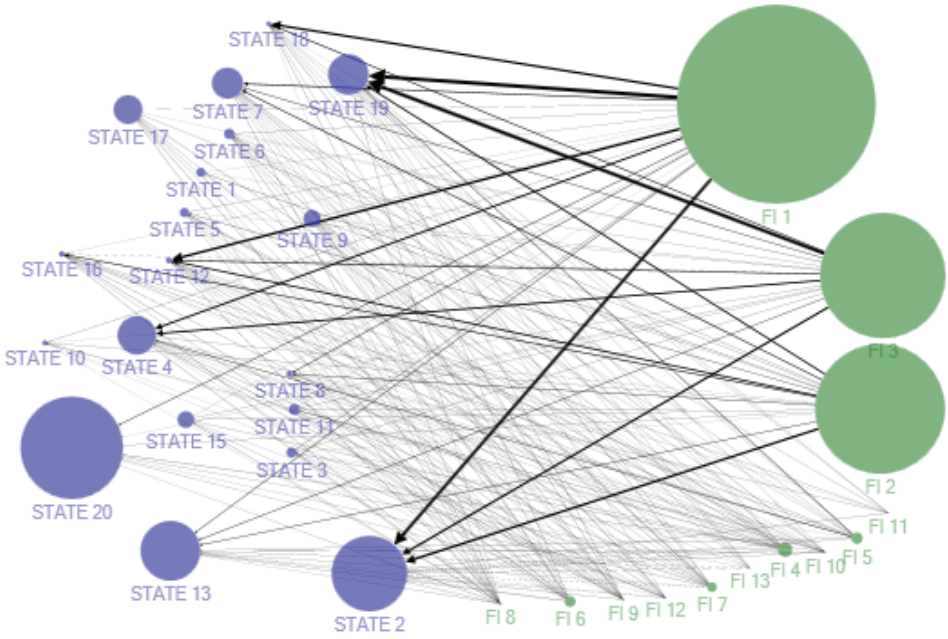

Figure 2.1 Network diagram showing largest exposures of key FIs to US states and their local (city, county) governments (fictitious sample data). In the figure, the states are represented on the right, while the FIs are represented on the left. The size of each node represents the total guarantee amount for states or the total guarantees written (for FIs). The weight of the edges of the graph that connect states to guarantors indicates the size of the guarantee. (The data used to generate the network in this example do not represent those of real exposures or financial institutions.)

- Issuer ID
- State ID
- Guarantor ID
- Guarantee type
- Guarantee amount

Now, consider the vast quantities of data that would not be required: in addition to not requiring portfolio information on each FI, we did not need the full annual financial statements for each state and local government, for example. We also did not require time series of bond prices or CDS. While detailed terms and conditions for the bonds and guarantees would certainly improve our analysis, we were able to get a reasonable overview without them.

Upon seeing this analysis most readers can imagine dozens of follow-up questions that would be interesting to ask. For example, a more involved analysis might contemplate the capitalization of the FIs before and after the default of the state,

or go on to calculate the conditional probability of default for the FI subject to the state defaulting, and so on. Each of these would add to the required data set. Some of these could be answered by simply joining additional fields onto the results of the original simple query, while some would require much more involved operations or even additional data collection. Clearly, the analytics themselves could be made more complex as well, for example, accommodating feedback loops and contagion.

Thus, we see a trade-off that is typical to data analytic problems. In the near term, the challenge for modelers will to find the appropriate balance between data that can yield useful results in reasonable time (and without extensive manipulation), and analytics that are too general or simple to provide actionable information.

It is always the case that researchers will not know, in advance, the full scope of the data requirements for ad hoc investigations. Collecting and organizing as much data as possible is one long-term solution to this problem. But it *is* a *long-term* solution. It is also a solution that, by its nature, will continually evolve as new analytic techniques are developed and as financial markets themselves evolve. However, there is much that modelers can do in the interim using the data and tools at hand.

The bulk of the chapter deals with how different data collection and access approaches relate to the analytic issues associated with modeling systemic risk. The objective of the chapter is to examine the requirements of some broad classes of modeling techniques and provide some background on the trade-offs of using data at various levels of aggregation and anonymity.

As a side benefit, we provide some guidance on initial data collection and integration projects in advance of the completion of larger-scale and more fulsome data standardization and pooling efforts. Beyond this, however, we hope that the framing of these issues will inform some of the more robust efforts to construct large-scale comprehensive data repositories.

We provide a loose framework for thinking about trade-offs in this space. The key dimensions we consider are the level of aggregation and the level of linkability of the data. Aggregation and linkability are important because they influence directly the types of models researchers can build and the ease with which data can be collected for them. We develop a conceptual map relating these dimensions, expressed as a 2×2 matrix. Curious readers can jump forward to Figure 2.4 in Section 2.4 to see this matrix, which we have also populated with some example model types. The remainder of the chapter discusses these dimensions for systemic-risk modeling.

Section 2.2 describes how data aggregation impacts model performance. The statistical and econometric properties of micro-level vs. aggregate data have been studied extensively, and we briefly review some of the more useful results. This dis-

cussion helps answer questions such as "How important is it, in terms of accuracy, to be able to model mortgage portfolios at the loan level?"

In Section 2.3 we discuss the related topic of *linking* various independent pieces of data to form a more coherent and holistic view of risk. Linking may involve aggregating data across organizations (e.g., "What is the total exposure of Asian banks to US RMBS?") or it may involve connecting entities within a single organization hierarchically (e.g., "What is the impact of a default of Company XYZ on the other members of the corporate family?"). This is central to some forms of analysis. Because it becomes increasingly feasible to link together data from different sources as the unit of analysis becomes finer, the level of aggregation also affects modeling options from this perspective. We highlight some linking issues and how they can affect model choice and quality.

In Section 2.4 we present a 2×2 framework for matching modeling approaches to data availability or, conversely, for defining data collection efforts based on modeling requirements. This section deals with trade-offs that modelers and data experts can make. Since these are often subjective, this section gives a conceptual view on how these trade-offs can be framed and how to formulate answers to questions such as "Given that we can easily compile data set x what modeling options do we have?" or "In implementing a network model of counterparty CDS exposures, what data issues should we consider?"

Finally, because it will be necessary to pool data from various public and private sources in order to develop a comprehensive view of some forms of systemic risk, in Section 2.5 we consider some of the practical concerns for institutions contributing data to consortia and for those organizing data collection. Paramount among these are challenges relating to privacy and anonymization. This section helps outline the issues that are central to answering questions such as "If we wish to collect detailed hedge-fund exposures at the instrument level from individual institutions, to what degree can we protect the confidentiality of the information?"

The appendix contains an example of systemic risk dashboards with discussions of the data requirements for each of the analytics used.

2.2 Data aggregation and statistical inference: At what level of detail should data be collected?

All equal, detailed data is more useful that aggregate data (if for no other reason than aggregates can be created from the detail, but not the reverse). However, detailed data collection, organization and storage are expensive, and the expense increases with the level of detail. Making such trade-offs is key to a systemic data strategy.

For example, should position-level data be stored for each portfolio in a bank or

should the aggregate exposures to major asset classes in each of those portfolios' be reported? While the former provides much richer detail, the latter is far less costly to produce, store and maintain and it raises fewer confidentiality considerations. A reduction in data collection effort comes at a cost; there are trade-offs in both the scope and precision of analyses done at different levels of detail.

For many problems, micro-level analysis provides a richer information set through which to study systemic risk (provided modelers also have the ability to aggregate when convenient). This is particularly so when measuring exposures to asset classes that exhibit high levels of non-linearity and/or heterogeneity in their behaviors or have path dependent payoffs. A body of empirical evidence and statistical theory has emerged in support of the notion that aggregation across such asset classes can often mask important relationships in the underlying assets: see Kelejian (1980).

However, the desire to model at the micro-level implies both substantial computing power and exacting database design. This can be expensive and in some cases, this level of detail may be unnecessary. In this section, we describe briefly some of the topics and results that have come out of the study of aggregation[5]. A fuller literature review can be found in a number of articles we cite and their references[6].

The most basic form of the aggregation problem has been termed the *ecological fallacy* and has been well studied in the political science and epidemiology literature. The kernel of this concept is that *group summary statistics* (e.g., means within a specific subset of the population) typically cannot serve as proxies for the attributes of *individuals* within the group[7]. Said differently, the relationship between two aggregate measures is not generally the same as the relationship between the equivalent two micro-measures. So for example, a model relating the various average levels of credit score, LTV and loan coupon rate of California borrowers to the average default rate in California, generally does not tell us much about the behavior of an individual borrower with a given credit score, LTV and coupon rate.

A more relevant form of aggregation relates to the advisability of using individual level vs. aggregate data to estimate the *aggregate* outcome (example, using aggregate portfolio statistics to estimate the default rates for the portfolio). Here much of the discussion involves determining the degree to which the micro-data and models are subject to error. In principle, it is possible that the estimation error on *individual* observations and models is higher than that of the aggregate because

[5] An earlier, more extensive, version of this literature discussion first appeared in Chinchalkar & Stein (2010).

[6] Note that we focus here on more general statistical aggregation issues. From an economics perspective, a stream of literature deals with production functions, consumer behavior and other topics and entails the study of *index numbers*. This research involves specific types of aggregation and economic models. For some of the more recent developments, see, for example, Barnett, Diewertb, & Zellner (2011) and the articles contained in the special issue that it introduces.

[7] See Robinson (1950) for one of the earliest mathematical discussions of this topic.

aggregating the micro-data allows the errors to cancel out[8], for example, if the relationships are linear.

In the early 1980s, perhaps as a result of the increasing sophistication of the non-linear models used in finance and the availability of computing resources to estimate them, researchers began to focus on aggregation of non-linear micro-models (see Kelejian, 1980) which continues an area of active research[9]. These results have led to a view, that, when feasible, using micro-data will lead to more accurate models and inferences, particularly in the presence of heterogeneous populations and non-linear dynamics (see Blundell & Stoker, 2005).[10] This may be particularly important in credit risk management settings, where estimates of the higher moments of loss distributions are of interest (Hanson, Pesaran, & Schuermann, 2008).

Example 2.1 (Equity Options) To make this more concrete, consider the example of estimating the value of a portfolio of 200 equity call options. Each option has a different strike price, a different expiration date and references a different underlying firm's equity price. Even with knowledge of the *average* strike price, the *average* expiration date, the *average* historical volatility, etc. for the options in the portfolio, it would be quite difficult to estimate the value of this portfolio of options using only this aggregate information. It would be similarly challenging to run a stress test (e.g., the S&P 500 drops by 20% in a day) based on such summary information. Examining an historical time series of the option portfolio's value might provide only limited insight into how it would behave in the future, given the non-linear payoffs that characterize the options.

Example 2.2 (Mortgages) Chinchalkar & Stein (2010) give an example of mortgage loan portfolios constructed from a universe of US non-conforming prime mortgage loans data. The authors used two common credit factors as measures of mortgage default risk: FICO (the borrower's consumer credit score) and the combined loan to value ratio (CLTV). They demonstrate that it is not hard to systematically create portfolios that appear to be identical based on their aggregate

[8] In one of the earliest analyses, Grunfeld & Griliches (1960) demonstrate that in linear settings in which there is substantial error in estimating the micro-models, aggregation may produce a benefit by smoothing over data and estimation noise, which may be sufficient to offset or exceed the aggregation error. Also see Ainger & Goldfeld (1974).

[9] For example, van Garderen, Lee, & Pesaran (2000) explicitly study the topic of aggregation of non-linear micro functions for prediction and conclude that except in the special case in which the aggregate and micro-equations produce identical results, "... if reliable disaggregate information is available, it should be utilized in forecasting, even if the primary objective is to forecast the aggregate variables." Blundell & Stoker (2005) provide a review on the state of the art on the issue of aggregation when there is heterogeneity across micro-units and then provide some special cases where data can be aggregated in a non-linear setting, though these cases are not typical.

[10] Blundell & Stoker observe in this paper that "Heterogeneity across individuals is extremely extensive and its impact is not obviously simplified or lessened by the existence of economic interaction via markets or other institutions. The conditions under which one can ignore a great deal of the evidence of individual heterogeneity are so severe as to make them patently unrealistic ... There is no quick, easy, or obvious fix to dealing with aggregation problems in general."

summaries but that experience quite different realized default behavior due to their individual characteristics. Because the joint distributions of the factors are not practically recoverable from aggregations of this sort, the conditional behavior of the factor interactions is lost. The authors also extend the example to show that if the underlying mortgages are securitized, the differences in analysis (on pools with very similar aggregate characteristics) become amplified.

In principle, with sufficient data, it should be possible to capture the appropriate levels of granularity within aggregate-level summaries. One possible approach is to exhaustively describe all conditional relationships within the data. For example, if there were k factors in a model each with m levels, we would create $m \times k$ cohorts. However, in general adequately including all such interactions would require so much data as to be impractical in most cases.

There is a rich literature on statistical methods for aggregate data. For example, much of the literature on time-series econometrics deals with characterizing and forecasting aggregate time series (see Elliott, Granger & Timmermann (2006) or Enders (2009) for a less technical treatment). We differentiate between these problems and problems in which the unit of analysis exhibits the heterogeneity, nonlinearity and path dependence. (Returning to Example 2.1, few researchers would realistically suggest forecasting even the value of an option on an equity index by estimating a time series model on the historical option prices. A more natural approach might be to model the underlying equity index and then apply an options pricing model.)

However, from a practical perspective, it is often cognitive and computational issues, rather than statistical ones that can dictate the optimal level of aggregation. For example, it is computationally far less demanding to compute on spot foreign exchange rates than it is to do so on the individual FX trade that occur over the course of a year. Cognitively, even were such calculations within the computing power of an analyst, visualizing all of the individual cross-currency FX trade could be overwhelming. The challenge, then, is to find the level of aggregation that retains the quantitative properties of the phenomenon under study, while not overly taxing the resources required to analyze them.

2.3 Data linkage

Much of the study of systemic risk involves the study of relationships among different financial entities. This requires that modelers be able to combine data from different sources. However, it is often the case that data in different data sets may not share common identifiers. This can impede such analysis. Traditionally, data linkage operations allow additional data from a one data set to be joined with

the records in another, when useful identifiers are not available. Record linkage techniques are commonly for marketing, census analysis or medical research applications. For example, a database of demographic information on consumers in one database might be joined with in-store purchase information from a second database for purposes of determining the relationships between demographics and purchasing patterns.

Much of the research in record linkage involves the properties and efficiency of automated algorithms for identifying common entities, when no common keys are present. Early work (see Dunn (1946) or Fellegi & Sunter (1969)) focused primarily on linking personal information and eliminating duplicate records for census or medical research applications, though more recent work has also focused on applications such as marketing and data mining. Recently, there has been an emphasis on research emphasis on using text fields to match records. For a detailed discussion of record linkage approaches see Winkler (2006) or Herzog, Scheuren, & Winkler (2007).

Record linkage is important statistically because of the sometimes-substantial impact that broken or mistaken links can have on the analysis of a data set. For example, Abowd & Vilhuber (2005) show that even small rates of linkage errors can have material impact on statistical results[11]. In fact, a substantial segment of the applied use of record linkage analysis is to remove duplicate records when merging multiple databases.

Though augmenting and expanding the content of individual records or aggregating multiple records from the same entity into a single summary record is also useful in the context of systemic risk, this may not be the main objective systemic risk modeling. In some cases, the primary objective is to create connections *between* entities and to permit more granular "drill-through" analysis.

For example, it can be useful to perform a stress test on multiple bank portfolios while ensuring that the same, e.g., CDO security, behaves in a correlated way across these institutions. One way of doing this is to stress the assets underlying structured securities, link these back to the CDO securities themselves and then calculate the impact of these stresses on the CDO transactions that hold them. Similar analysis can be done for RMBS (by first stressing the underlying mortgages, calculating the cashflows and losses for each mortgage, and then running this through a cashflow waterfall), etc. Because some payoffs (e.g., for basket swaps, structured tranches, etc.) are non-linear in the underlying assets, this type of analysis is not

[11] In this study, the authors analyzed a dataset containing about a billion quarterly employment records in which the error rate of the recorded social security number was estimated to be between one and two percent per quarter, resulting in inaccurate matches for these records. The authors find that even this small incidence of matching error could impact estimates of employment and job creation substantially.

feasible unless the underlying assets can be linked to specific derivatives transactions and these, in turn, linked to specific portfolios within financial institutions.

As another example, a source of stress in financial markets is the dependence of one institution on another counterparty, which, in turn may also have counterparty exposures to still other institutions, creating the potential for cascading defaults and contagion across institutions as the failure of one requires its counterparties to raise capital in an adverse environment.

Applications in, say, marketing may place emphasis on expanding the *number of fields* in a single record to provide more factors through which to identify interesting patterns. However, applications in systemic risk tend to be more focused on linking hierarchical information: either making use of relationships between objects observed at *different levels of granularity* or between *similar levels of different hierarchies* (e.g., as in the case of understanding how the portfolio holdings of Bank A relate to the portfolio holdings of Bank B, a large counterparty to A). Unfortunately, this type of linking has not historically been as amenable to automated methods.

The need to link disparate data sources and to be able to identify common entities within them has led to a number of industry and governmental efforts to create a single global set of *legal entity identifiers* (LEI). For example, the first formal call for proposals from the Office of Financial Research in the US was for an organization to develop and maintain an LEI registry that would serve as a public utility for the financial markets (Office of Financial Research, 2010).

Though such a universal LEI is seemingly fundamental, as of the end of 2010, there was no consistent industry standard for identifying either financial instruments or institutions. While numerous schemes existed for identifying corporations (various tickers, 6 digit CUSIPS and so forth) these were neither uniformly used nor uniquely applied. Events such as corporate mergers and divestiture, reorganizations and bankruptcies often confounded the use of common identifiers throughout the lifecycle of a financial firm. The situation was no better for the identifiers used for individual transactions and instruments (CUSIP, ISIN, SEDOL, ...). These different standards provided redundant coverage in some cases, but no coverage in others, even for some standard and widely used instrument types (e.g., commercial paper). One goal of the OFR's LEI initiative is to address this gap and provide a single identifier for financial entities.

A number of market participants view LEIs as only a first step and are working on more expansive data models. Parallel and sometimes complimentary proposals have been offered as the industry iterates towards a more fulsome and detailed data model for financial entities and instruments (see Flood (2009), Gross (2010), EDM Council (2011), Bennet (2011)). Some of these efforts are discussed more fully in Chapter xx.

If successfully executed, these projects will enable more expansive research on systemic risk (and more generally on financial institutions and instruments), which is currently hobbled by the use of ad hoc mechanisms for identifying common institutions and instruments across data sets. Such inaccuracies affect not just the ability to formulate problems and conduct research, but also the reliability of results based on analyses of current data.

Importantly, however, *labeling* an entity with a unique identifier is quite different from *understanding* how that organization or security fits into the broader financial system. Because of the many interactions among market participants and the various ways in which individual securities may be used and reused within the financial markets, some modeling frameworks additionally require both ontology and the *reference data* to populate it. The ontology describes how various entities can relate to each other logically and legally, while the reference data populates this description with real instances. Chapters xx of this volume describes some of these efforts.

Example 2.3 (Counterparty Exposure) To illustrate the complexity involved in defining relevant relationships, consider the (highly stylized) view in Figure 2.2 of some of the entities and links that could be required to determine the counterparty exposure of one entity to another or the aggregate exposure to a particular entity within an investment portfolio. In the figure, we see that answering the question,

"How much exposure does Portfolio A have to Company XYZ?"

requires traversing a number of linkages. In this example, we might begin by first examining direct obligations of Company XYZ in the portfolio (bonds, loans, CDS, etc.) by linking the specific instruments to their issuers. We might next examine whether there are other instruments to which XYZ is counterparty (e.g., swap contracts, etc.). We might also wish to check for the presence of XYZ in the collateral pools of CLO tranches that are held. Depending on our analysis, we might also include transactions on which XYZ provides some other form of credit enhancement (e.g., letters of credit, guarantees, etc.), though these might be weighted differently; or we might we might elect to repeat the process for the subsidiaries of XYZ[12].

For other questions, different types of links and configurations might also be useful for systemic risk analysis. Consider the linking structure that might be required to answer the following question:

"What is the exposure of non-US banks to California residential mortgages?"

[12] In practice, it is common to record the "ultimate parent" of an entity in order to expedite such aggregation. The ultimate parent is the entity at the top of a corporate family tree. However, this is not always obvious. See Section 2.3.

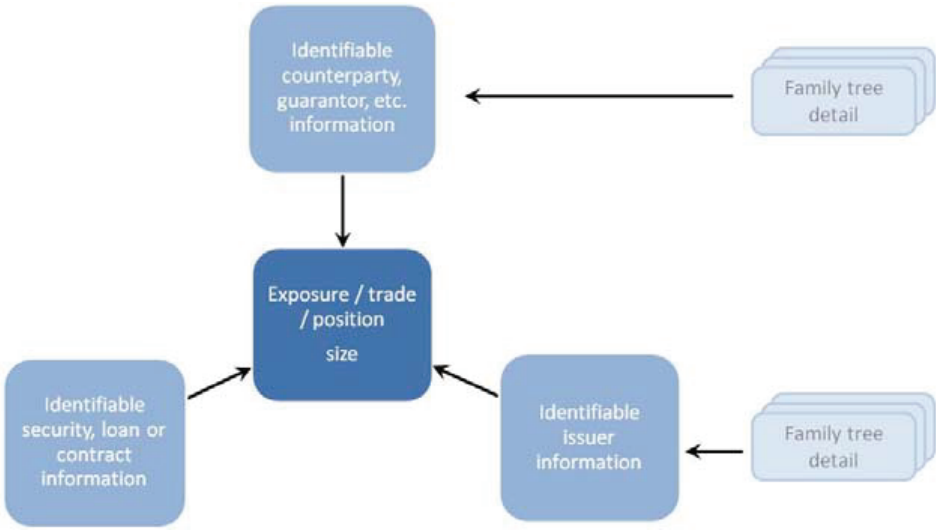

Figure 2.2 Stylized diagram showing some of the linkages required for aggregation and disaggregation of financial data for modeling or reporting. Arrows are shown with respect to flow from higher- to lower-level, though other operations would change the directionality.

In this case, in addition to information on the actual mortgage holdings, RMBS collateral pools, bonds issued by mortgage insurers, etc. would be of interest[13].

Example 2.4 (CDO collateral holdings in same corporate family) In evaluating portfolios of corporate exposures, such as those underlying CDOs, it is common to consider the correlation between two or more firms in the portfolio. This correlation may be estimated in a number of ways and may be characterized as default correlation, asset correlation, cash flow correlation and so on. It is common to use historical data as a first step in estimating correlations between firms or in estimating a firm's loading on a set of common factors used to induce correlation.

However, consider the case in which two firms in different industries share a common parent. For simplicity, consider the case in which the default of the parent makes it more likely that both of the subs will also default. The sharing of a common parent may increase the correlation between the two firms in a manner that is not necessarily obvious from the historical behavior of either firm.

How would we study the degree to which corporate CDOs, as an asset class, contain multiple firms with common corporate parents? To do this analysis, we would require:

(a) the portfolio holdings of each CDO;

[13] One could even extend this analysis to include, e.g., exposures to the debt of intuitions which themselves have large exposures to these asset classes, such as bond guarantors or mortgage guarantors.

(b) a common identifier for individual securities that was shared across all portfolios in (a);
(c) a common identifier for issuing firms that was used across all portfolios in (b); and
(d) a family tree hierarchy that also used the same set of common issuer identifies as in (c).

Designing databases and analytic systems to efficiently perform the types of navigation and analysis in our examples requires substantial planning and expertise. However, it also requires that the databases be populated and maintained reliably. While a number of institutions do maintain reference data and linkage information for subsets of the financial industry, there does not currently appear to be a complete mapping of the hierarchies of individual institutions and entities. Mapping such relationships and maintaining hierarchical information such as this can be time consuming, particularly if common identifiers are not used by convention. In the absence of standardized identifiers, much of this work must be done manually.

As an alternative to such manual processes, new approaches to inferring relationships between financial firms from publicly available data have emerged. This type of entity reference data extraction is similar in some sense to automated record linkage, though rather than discovering common records in different databases, the objective is to discover the more complicated relationships between entities in order to derive reference data. For example (Hernandez et al., 2010) describe a text extraction application that constructs a network of business relationships from public filings. The authors demonstrate the approach by extracting a network of lending relationships from SEC and FDIC filings. Such automated construction and maintenance of corporate hierarchies, legal identifiers and counterparty relationships may become increasingly important as the demands for linkable information grow.

2.3.1 Challenges in defining key relationships

Examples 2.3 and 2.4 took as given that corporate family tree relationships were well defined. However, such analysis may be complicated by differences in the definition of a relationship between two firms.

In general, it is straightforward to identify the affiliation between a parent company and a wholly owned subsidiary. However, from a risk analysis perspective, this is not the only set of relationships that may be important. For example, a portfolio may contain debt issued by a partially owned subsidiary of another issuer (in the portfolio) or transactions for which a firm in the portfolio is a guarantor or, in the case of structured securities, a servicer.

These more subtle affiliations may also induce codependence. Defining the de-

Table 2.1 *Percentage of money market fund industry survey respondents who would aggregate each relationship type with the parent as a "single exposure."*

Type of entity	% that would report as same name exposure
Parent	100%
Wholly owned sub	94%
Guarantor	78%
Partially owned sub	69%
Servicer on ABS	13%

gree of affiliation can be difficult and market participants do not always agree on these definitions. For example, Shilling (2007) surveyed money-market fund professionals to determine their views on the definition of entities that should be consolidated for purposes of reporting a "single name exposure." The results, shown in Table 2.1, below, suggest that in some cases respondents differed considerably in the degree to which they considered various types of relationships to constitute a "single name" risk exposure. For example, about a third of participants would not consolidate a partially owned subsidiary, while the majority would.

If such disagreements are indicative, it is unclear that there is a simple algorithmic solution to the problem of defining ownership relationships in such cases. Furthermore, in different contexts, the definitions are subject change. Managing context specific semantics will be a challenge and suggests the need for far more flexible data systems and ontologies.

2.4 Aligning data and models

Ultimately, large and extensive investments in IT infrastructure and staffing expertise will be necessary to perform the most detailed analysis. Robust data architectures and fast storage and retrieval mechanisms are important and essential components of any reasonable strategy for aggregating data on systemic risk.

Given the scale of such investment, it is useful to consider how best to deploy resources and capital for compiling data on systemic risk. One way to organize and prioritize infrastructure and design activities is to do so according to the analytic questions risk managers, policy makers and regulators wish to answer. There are many approaches to thinking about and modeling systemic risk and while the most involved of these require long-term planning and careful attention to data structure, hardware performance and common identifiers, a substantial proportion of modeling approaches do not require the full complement of IT infrastructure for researchers to begin producing actionable output.

For exposition, we delineate two dimensions in thinking about data-model fits.

The first is the level of (dis)aggregation that a modeling approach requires of the data and the second is the degree of linkability that the model demands. We discuss these dimensions in the context of model selection and provide a 2×2 matrix that shows an approach to using these dimensions to align data and models. To make the 2×2 framework more tangible give examples of how a number of modeling approaches would map onto the matrix.

The framework brings together the topics we discussed in Sections 2.2 and 2.3. Though closely related, aggregation and linkability remain distinct dimensions. It is clearly infeasible to link exposure-level records if data is aggregated at the portfolio level. However, it nonetheless possible for data at the detailed exposure-level data to be either linkable or not, depending on the level of anonymization and the extent to which common identifiers have been implemented. An example of the availability of detailed micro-level data without linkability is the mortgage data now provided by most US RMBS trustees. This data contains detailed loan-level information on the characteristics and performance of the individual mortgages that underlie specific RMBS transactions, but there is no common borrower identifier so the data cannot be directly linked to other information about the borrower. As a result, it can be difficult to identify loans with second liens or to link to information about other credit lines.

To fully represent the interplay between these dimensions – level of aggregation, level of anonymization and degree of linkability – would require a three-dimensional construct. However, much of the volume of the 3D space would be empty (e.g., any time a level of aggregation exceeds the unit of analysis, a lower bound on anonymization is determined and an upper bound on linkability is implied). For simplicity, we abuse slightly the dimensionality and reduce the presentation to the two dimensional graphical one shown in Figure 2.3. (We also defer discussions of anonymization and confidentiality to Section 2.5.)

In the figure, the level of detail and linkability increases as we move from the left to the right to and the level of detail increases as we move from bottom to top. The corresponding volume of data sets also increases as the level of detail grows. The examples shown in the figure pertains to aggregations of individual portfolio positions or trades up through institution-level holdings and finally to market-level summaries. Other applications would follow similar lines, but could differ in their details.

At the finest level of granularity, shown in the top right, are fully identified position-level information on the size of the trade or position, and the various parties involved in the transaction (e.g., issuers, guarantors, swap counterparties, etc.). As we move down and to the left, we strip away either detail or linkability (or both). The off-diagonal quadrant represent data sets that are partially aggregated

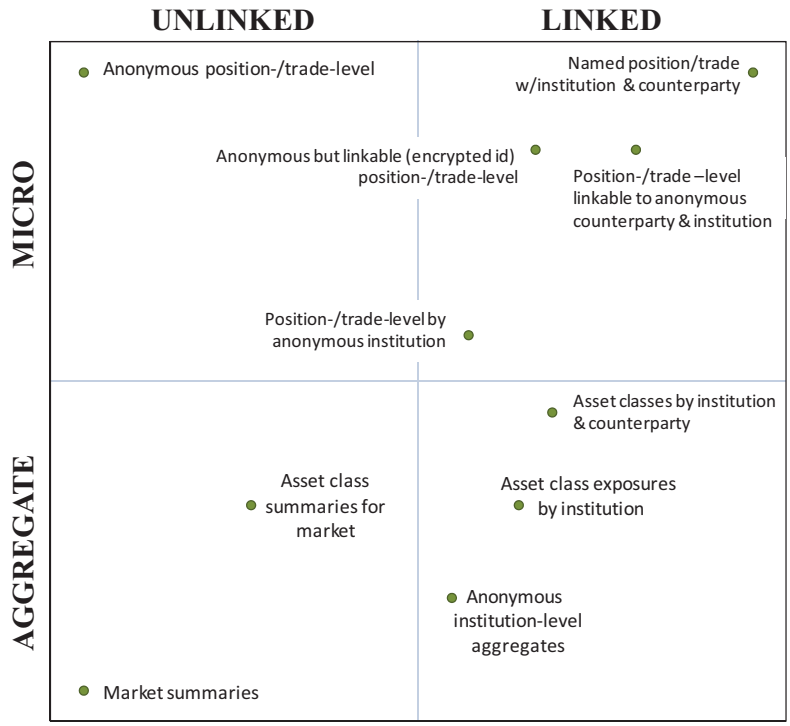

Figure 2.3 A 2×2 framework describing levels of data aggregation and linkability with examples. The descriptions in the lower left are the most aggregated and least linkable. Those at the top right are the highest resolution and potentially most linkable.

or linkable, but not fully flexible in this regard. Finally, the lower left corner of the plot holds market-level summaries.

The level of structure in a particular modeling approach often determines to a significant degree the type of data that are required to implement it. The level of aggregation at which a model operates therefore provides guidance on the types of data, and the linkability of that data, that is required. The converse is also true: if a researcher has access to a given set of data, the attributes of the data naturally suggests some practical boundaries on approaches for modeling systemic risk.

In thinking through both modeling strategies (given data) and data collection strategies (given a modeling objective), it can be useful to evaluate the dimensions of aggregation and linkability. To aid this perspective, Figure 2.4 provides a version of the 2×2 matrix for balancing data constraints and modeling objectives with some examples of modeling approaches for systemic risk. The approaches have been mapped onto the matrix. These examples are neither comprehensive nor

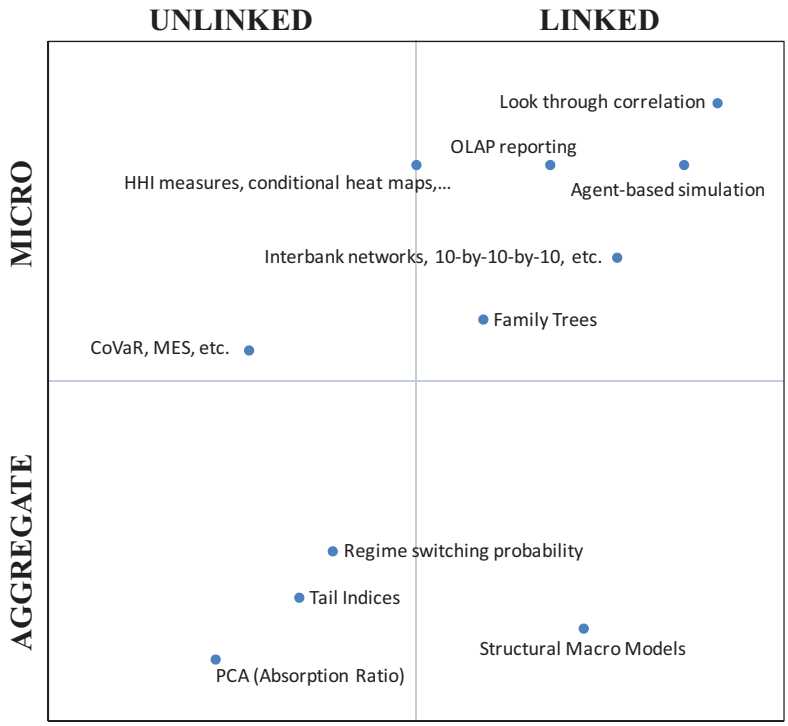

Figure 2.4 The 2×2 framework as used for matching data to models, with examples. Unlike some 2×2 formulations, it is not always possible to establish the analytic dominance of one quadrant over another.

exhaustive, but they do give some sense of how the framework can be used to think about data and models[14]. For an extensive review of modeling approaches for systemic risk; see Bisias, Flood, Lo & Valavanis (2012).

While the framework is useful as a heuristic tool, unlike some 2×2 formulations, it is not always possible to establish the financial or operational dominance of one quadrant over another. We can say that, in general, data collection in the lower left quadrant is easier to implement than the upper right, the situation is murkier with respect to, say, the lower right quadrant and the upper right. For example, it is may be far more difficult to implement a structural equation model of the macroeconomy than to create analytics for traversing family trees.

There is also no clear dominance in terms of the analytic utility of one model

[14] Modelers will almost certainly argue with the classifications of some of the methods. However, the purpose of the 2×2 is not to create a universal standard. Rather it is intended to aid practitioners and researchers in formulating strategies for modeling and data collection that contemplate these dimensions. In this context, a bit of debate is productive. Though the matrix resembles a scatterplot, the placement of the various modeling approaches is ultimately subjective.

over the other. Both the effort required to implement different approaches and the resulting value the resulting tool generates depends a great deal on the details of the applications and implementations themselves. Thus, while the matrix may be useful in thinking through these issues, there is still broad scope for subjectivity in matching data and modeling approaches.

Example 2.5 (Monitoring the potential for systemic events) To demonstrate the framework, we examine two monitoring approaches: that of Kritzman, Li, Page, & Rigobon (2010) and that of Duffie (2010). Both of these methods offer tools for understanding the potential for increasing of systemic risk in the financial system resulting from couplings between market participants.

Consider first the Kritzman, Li, Page, & Rigobon (2010) approach. In this case, recall that various time-series of market returns for different market segments are decomposed into principle components. The measure of interest, which the authors term the *absorption ratio* (AR), is computed as the amount of total variance in the movements of the set of market segments that is explained by the first n eigenvectors (in their example $n \approx N/5$, where N is the total number of asset classes or market segments under study). Conceptually, the measure relies upon the notion that periods during which the movements of many asset classes are explained by a small set of common factors are periods in which there may be a tighter linkage between different asset markets and thus higher potential for a system-wide event.

This approach is quite general and makes minimal demands on the data. In fact, it can be implemented entirely using publicly available data. Furthermore, a researcher or regulator does not need specific knowledge about the practices or behaviors of the institutions that participate in a market. The cost of the parsimony is that the information the AR provides is narrowly focused. Although the measure gives signals about linkages between markets and can also be extended to examine the systematic importance of individual institutions (Kinlaw, Kritzman & Turkington 2011), it provides relatively less guidance on which institutional relationships may be at the most influential in transmitting the risk and why.

Now consider Duffie (2010). Under this approach, a regulator requires that the most significant N financial institutions report their exposure to their largest K counterparties under each of M asset-specific stress scenarios, where N, K and M are not too large (e.g., $O(10)$). (Institutions would choose the K counterparties stress-scenario by stress-scenario, based on their exposures.) Reporting entities also provide analysis of their own sensitivity to the stress scenarios[15]. Once the results of each scenario have been computed by each institution, the regulator then aggregates these results, scenario-wise, to get a snapshot of the state of the finan-

[15] The exposure measures for both self- and counterparty-stresses include cash-flow impacts as well market value impacts both before and after collateral.

cial system "one tick after" the scenario takes place. One goal is to identify key asset classes or counterparties that may be important systemically or key events that could lead to system-wide disruption.

In contrast to Kritzman, Li, Page, & Rigobon (2010), Duffie (2010) demands private data from institutions and requires that it be linkable, in some fashion, so that aggregate exposures to common counterparties may be identified for key asset classes[16]. In exchange for the higher demands on data and institutional knowledge, however, Duffie (2010) provides more specific information about the underlying drivers of systemic risk and the institutions that are tied most directly to them. The regulator can further use the information to identify new entities, previously not among those being monitored, that may be important systemically.

The fact that one technique is easier to implement or provides more specific information, respectively, does not imply that one is "better" or "worse" than the other[17] Rather, it highlights the need to match data and modeling techniques in a deliberate fashion so that the analysis makes the most use of the available data and the data is collected in an efficient and directed manner.

2.5 A brief comment on confidentiality, anonymization and the role of consortia

It would be an omission to conclude a chapter on systemic risk data without discussing one of the central operational challenges that arises in data pooling and consortia: data governance and access. Determining appropriate governance for combining and accessing information from various individual institutions can be exceedingly difficult, particularly in instances in which data contributors are competitors.

By construction, measuring systemic risk requires that researchers and analysts consider risk across institutions, markets and asset classes because it can be difficult for individual institutions to understand the many relationships and interconnections between market participants, and it is impossible for any institution to do this across all relevant market segments.

For example, most broker-dealers can observe their repo exposures to their own clients and can perhaps even infer relationships between their clients and other broker-dealers, but it would be hard for them to observe the full portfolio holdings of the clients and how these relate to the repo exposures. Similarly, currency dealers may have excellent insight into the micro-structure of currency markets and how their clients are hedging or speculating in them, but they might not be able to

[16] In addition, the regulator needs to have a sense of how the institutions are conducting stress tests and to be comfortable with the approaches.

[17] To the contrary, one could imagine using both in a complimentary fashion to inform each other.

easily characterize their clients' overall direct and indirect exposures to a particular currency.

Data pooling through governmental or industry consortia is an obvious institutional response to this problem; without such pooling, is difficult to envision a fulsome picture of systemic risk emerging. However, consortia can be difficult to implement due to issues of both cost and confidentiality.

Confidentiality concerns emerge since an implication of data pooling is that other institutions (i.e., competitors or clients) can gain greater insights into an FI's holdings, trading behavior, client relationships and risk management processes. This can undermine the core business models of some market participants, which rely on proprietary research and client confidentiality, among other things.

Because of this tension – the desire for detailed analysis of market linkages on the one hand, and the need to protect client and firm confidentiality on the other – commercial institutions and regulatory agencies will likely spend a good deal effort working through governance mechanisms for how data is collected and to whom the data is made available.

Data collectors may also need to provide mechanisms for anonymizing data from different market participants and data vendors. Because of the advantages of using micro-level, linkable data for some types of analytics (see Section 2.2 and Figure 2.4) and of linking data from one institution to data of another (see Section 2.3 and Figure 2.4), there are clear advantages to preserving linkability.

Anonymization and privacy protection is a rich, and often application-specific, domain that we do not attempt to discuss fully here. Interested readers can find more extensive discussions of anonymization algorithms and metrics in e.g., Fung, Wang, Chen, & Yu (2010) and the references therein[18]. In this section, we focus on the implications of some of the ways in which data may inadvertently disclosed despite having its identifiers obfuscated.

Anonymization is related to both aggregation and linkability. (There is less identifiable information in an aggregate summary of portfolio holdings for a sector than there is in an aggregate portfolio summary for each firm and there is less identifiable information in an aggregate portfolio summary for a firm than there is in the position-level data that underlies it.)

The simplest form of anonymization involves recoding the unique identifiers for institutions and securities so that their identities are obfuscated. This might be done in a linkable fashion, in which their commonality is preserved[19], or it could be

[18] Fung, Wang, Chen, & Yu (2010) focuses on anonymization more for general data mining applications rather than solely statistical ones. This perspective is useful as it contemplates table linkage and the use of background information. The article also briefly discuses multi-party data pooling, high-dimensional transaction data anonymization and threats to privacy when researchers are able to repeatedly query a database.

[19] Note that even this is not trivial. Such a mapping operation also requires that there be both (1) a uniform coding of market entities and (2) the existence of a trusted third-party capable of maintaining in confidence

Table 2.2 *Distribution of banks by region and size (hypothetical)*

	NE	Mid	SE
Small	30	15	12
Med	22	12	13
Large	10	2	5

Table 2.3 *Mean % of assets in toxic asset class (hypothetical)*

	NE	Mid	SE
Small	1.2	5.2	7.2
Med	13.5	16.5	11.5
Large	15.3	25.3	21.3

done in a more destructive manner in which the identifier of each entity is simply removed and entities can no longer be linked.

However, even with true anonymization of identifiers or aggregation of data, challenges remain. Most market participants, given a sufficient number of facts about an institution or position, are often able to surmise its identity. This is particularly so of firms that compete with each other. Anonymization procedures that can mitigate such inferences can be designed to avoid this type of disclosure, though this remains difficult.[20]

Example 2.6 (Anonymized stress-test results) For a (trivial) example – one in which a *single* fact is sufficient to discover the identity of an obfuscated financial entity – imagine a report providing information on the results of the average exposures to a very high profile "toxic" asset class for a subset of banks on the East Coast of the US.

To provide a sense of the characteristics of the banks in the report, Table 2.2 shows the distribution of banks by region and size. (In the example, we use the labels "Small," "Med" and "Large" but assume that these map to specific asset-size ranges of the banks that are defined elsewhere.)

Clearly, if a reader were the CFO of a Large Mid-Atlantic bank, she would know the exposure of her own bank and thus, since there are only two firms in that category in the report, she could combine the tables to solve for the exposure of the other large bank in her region (a likely competitor). Furthermore, any of the other

the "skeleton-key" (e.g., junction table) that maps each of the institutions' and instruments' true identities to that uniform coding of identifiers that can be linked. (An example of how such a trusted third party relationship might is the operating model of the Lincoln Lab at the Massachusetts Institute of Technology, which performs a conceptually similar function in the defense domain.)

[20] See Agrawal and Srikant (2000), Ashwin, et al., (2007) and Aggrawal and Yu (2008) for more details. Castro (2011) provides a discussion of recent research on protecting tabular data from such inferences.

statistics on the banks in the sample that were similarly segmented by size and region would be susceptible to the same decoding.

Basic anonymization techniques generally fall into one of two categories: those that remove key data elements, either selectively or globally (*data reduction* or *suppression* techniques), and those that that somehow change data elements to greater or lesser degrees (*data perturbation* techniques)[21].

Data reduction techniques can range from global suppression of fields such as securities' CUSIP numbers or a mortgage holder's tax IDs to using record-level case-specific rules for suppressing fields in records where they might lead to a loss of anonymity. In some cases, entire records may be deleted.

Data perturbation approaches involve changing in some fashion the data in one or more records, without deleting the data outright. A simple form of perturbation involves adding white noise to the values of a specific field or fields. More involved methods involve swapping fields among records and so-called, *micro-aggregation* or *generalization* in which individual records' values are replaced for a specific field with an aggregate or more general value (e.g. the mean of a cluster of similar records or the State rather than ZIP code). Both swapping and micro-aggregation can be done in a manner that preserves, to varying degrees, the marginal distributions of key variables in a data-set, though joint distributions may or may not be preserved.

However, the type of decoding shown in Example 2.6 involves the use of one reported data set to infer the identities of the firms reporting anonymized data in another data set. Exploiting contextual information to infer the identity of anonymized entities is a topic of great interest in the privacy literature. It implies that two anonymized data sets may be used to decode each other, even though the information is protected in each one individually.

The use of background information (e.g., in our previous example, "There *are* only two large banks in the Mid-Atlantic region so two banks in the table represent the full universe.") that is not explicit in either data set makes the anonymization problem more involved since the links between the data sets need not be explicit (e,g,, Sweeney, 2002, Narayanan & Shmatikov, 2008). An active a stream research in data privacy and anonymization is concerned with developing techniques that trade-off the dual objectives of preserving anonymity and maintaining the statistical integrity of the original data under these conditions. Along with them, metrics for measuring both the risk of disclosure and the loss of information due to anonymization have also continued to evolve (e.g., Dwork, 2008, Wang & Liu, 2011).

[21] We leave aside the approach of generating synthetic data (based on the statistical properties of the original data), which may also be used for anonymization. Though in some settings this may be useful, it can be difficult to generate synthetic that resemble real data closely enough for many applications.

Because of the sensitivity to disclosure on the part of the various institutions that generate and maintain data that is valuable in understanding systemic risk, regulators and consortia will find it useful to engage in both bi-lateral and collective discussions with industry participants to determine the implications of different confidentiality regimes.

As a means to further mitigating the risks of accidental disclosure (or intentional discovery), one suggestion has been to lag sensitive information for some reasonable period of time before disclosing it. However, even with such protections, it will likely be necessary to implement additional safeguards on access to the detailed data to ensure true confidentiality. To this end, it is useful to consider the potential for defining graduated levels of disclosure, similar to security clearance levels, with respect to the level of aggregation, the breadth of data elements and the time delay in reporting, depending on the entity receiving the information and the sensitivity the data.

A significant step forward in anonamization has recently emerged in the use of techniques for computer encryption. For example, techniques have been developed for performing calculations directly on encrypted data. Said differently, it is possible, in principle, for institutions to encrypt their data and submit it to a consortium, and for researchers to combine this encrypted data with other similarly encrypted contributions for analysis, all while the contributions are still in an encrypted state.

For example, so-called *fully homomorphic encryption* may eventually enable data encryption in a manner that still permits statistical analysis while retaining the anonamization and encryption. Though earlier forms of these techniques have been explored for the past several decades (see Rivest, Adleman, & Dertouzos, 1978), only recently has a single approach been developed that is fully homomorphic over addition and multiplication – the basic building blocks of most statistical operations – (Gentry, 2009). There are questions as to whether these techniques could be made computationally practical, though recent research suggests that in some settings this can be done (see Dijk, Gentry, Halevi, & Vaikuntanathan, 2010) in principle, though this approach remains in development.

Recently, practically implementable techniques have been introduced to permit statistical analysis of private financial data without requiring the unencrypted data to be revealed. The first authors to do this, Abbe, Khandani and Lo (2011) propose an alternative approach that takes advantage of the structure of certain statistical functions of interest in monitoring systemic risk. The authors make use of secure multiparty computation protocol that they adapt for monitoring systemic risk using financial data of the sort required by regulators and policy makers. This groundbreaking paper demonstrates the practicality (and tractability) of the method, by using individually encrypted real estate lending data from three large banking institutions to compute the total amount of outstanding loans linked to real estate.

Using the authors' approach, this is done without requiring the unencrypted data for any of the individual institutions to be revealed.

2.6 Conclusion

> What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.
>
> Herbert Simon

In this chapter, we have tried to outline some of the modeling trade-offs that result from data choices, and conversely, some of the data requirements implied by different modeling approaches. Our purpose in doing this is to encourage researchers and data experts to enter into more active dialogs to set research agendas and priorities.

The key themes in this chapter are the following stylized observations:

- Data organization and collection efforts for modeling systemic risk can benefit substantially by considering the analytic application context for data. Analysis of the fit between model characteristics and data constraints can guide collection and enable near-term development of useful tools.
- In whatever form data efforts begin, in many cases, there will be significant focus on the degree to which micro-level analysis of linkable risk exposures is desirable (e.g., for heterogeneous, path-dependent assets) versus coarser aggregates (e.g., for broad market flows).
- Constructing data sets for systemic risk measurement will require collaboration and pooling (through consortia and policy) by financial institutions. A byproduct of such pooling is the creation of significant confidentiality concerns that pose substantial challenges. These must be addressed in a manner acceptable to both industry and oversight bodies.

The ultimate goal of data collection for systemic risk analysis is the development of highly flexible, highly detailed and densely linked data repositories. However, the design and implementation of this type of data store will require careful planning and coordination. Designers will need to negotiate standards for various ontological conventions and identifiers as well as overcome difficulties that many financial institutions currently have in extracting and combining the data that each generates in the course of doing business. Because analytic needs will continue to evolve, there are still open research and operational questions on how systems and ontologies may be designed to accommodate this evolution without requiring full-blown redesign.

Fortunately, we can still do much with today's through much less ambitious

pooling. We have tried to provide some background and a framework for thinking about these nearer-term projects.

Researchers, analysts, regulators and policy makers need not make the false choice between either deferring systemic risk modeling efforts until standardization and collection is better formed, on the one hand or to undertaking only the most cursory analysis, on the other.

Rather, research can begin in the middle-ground and yield actionable results, provided organizations are willing to accept the extra cost and inefficiency that accrue from iterating over multiple versions of databases (and, in some cases, model implementations) as the fuller richer data repositories come on-line. We favor this strategy. While the costs of duplicative efforts and discarded models and code are high, we believe that the cost of waiting, both in terms of risk exposure and lost momentum, is higher.

Appendix: An example of a systemic risk dashboard with annotations for the data required to construct it

This appendix contains a sample wire-frame mock-up of a dashboard for systemic risk. The example may be useful in demonstrating the types of data that would be required for various practical modeling efforts. To this end, in addition to showing examples of the analytics in such a dashboard, we also provide some detail on the data that would be required to construct each of the measures: see Figures 2.5 and 2.6.

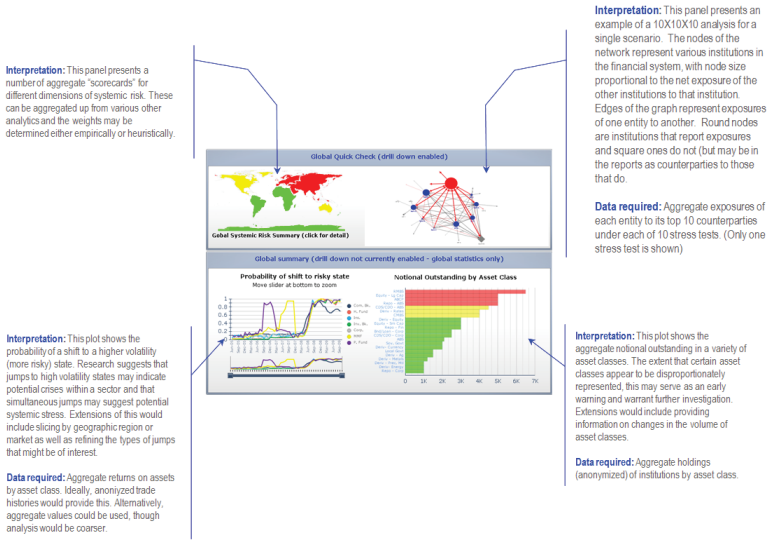

Figure 2.5

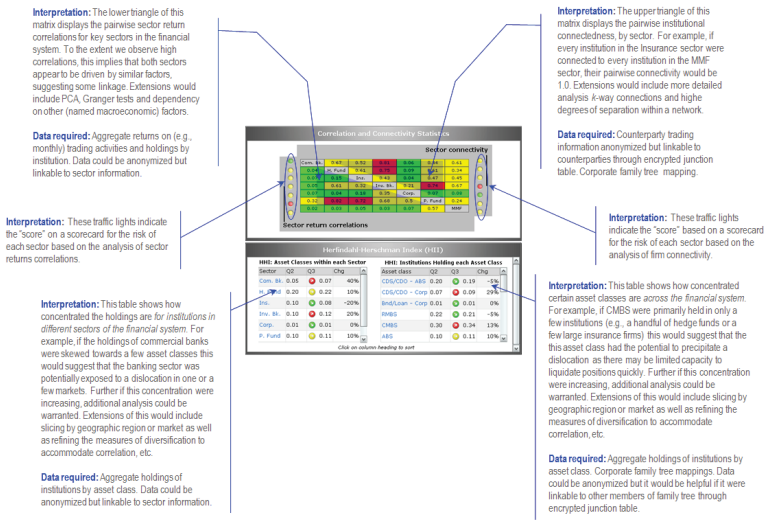

Figure 2.6

Acknowledgements I wish to thank Emmanuel Abbe, Shirish Chinchalkar, Felipe Jordäo, Andrew Kimball, Andrew Lo and Samuel Ring for their comments on earlier versions of this chapter. I received valuable feedback from the participants in the Measuring Systemic Risk Conference at the Federal Reserve Bank of Chicago on a presentation of some of these topics. I also had very useful conversations on a number of the issues in this paper with Darrell Duffie, Mark Flood, Francis Gross, Mark Kritzman, Joe Langsam and Yaacov Mutnikas. H.V. Jagadish provided extensive and very useful comments on a previous draft of this chapter.

References

Abbe, E. A., Khandani A. E. & Lo, A. W. (2011). Privacy-preserving methods for sharing financial risk exposures. MIT Sloan School of Management.

Aggarwal, C. C. & Yu, P. S. (eds.) (2008). *Privacy-Preserving Data Mining: Models and Algorithms*. Springer-Verlag.

Abowd, J. M., & Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers. *Journal of Business and Economic Statistics* **23** (2) 133–165.

R. Agrawal and Srikant, R. (2000). Privacy-preserving data mining. In *SIGMOD 2000: Proceedings of the International Conference on Management of Data*, 439–450. ACM.

Ainger, D. J., & Goldfeld, S. M. (1974). Estimation and prediction from aggregate data when aggregates are measured more accurately than their components. *Econometrica* **42** (1) 113–134.

Barnett, W. A., Diewertb, W. E., & Zellner, A. (2011). Introduction to measurement with theory. *Journal of Econometrics* **161** (1) 1–5.

Bennet, M. (2011). *The EDM Council Semantics Repository – Considerations in Ontology Alignment*. EDM Council.

Bisias, D., Flood, M., Lo, A. W. & Stavros V. (2012). A survey of systemic risk analytics. Working Paper. Office of Financial Research, Washington DC.

Blundell, R., & Stoker, T. M. (2005). Heterogeneity and aggregation. *Journal of Economic Literature* **XLIII** 347–391.

Castro, J. (2011). Recent advances in optimization techniques for statistical tablular data protection. *European Journal of Operations Research*.

CFO Publishing LLC. (2011). A new role for the times: opportunities and obstacles for the expanding finance function. CFO Publishing LLC.

Chinchalkar, S., & Stein, R. M. (2010). Comparing loan-level and pool-level mortgage portfolio analysis. Moody's Research Labs.

Dijk, M. v., Gentry, C., Halevi, S., & Vaikuntanathan, V. (2010). Fully homomorphic encryption over the integers. In *Advances in Cryptology – EUROCRYPT 2010*, LNCS 6110, 24–43.

Duffie, D. (2010). Systemic risk exposures: a 10-by-10-by-10 approach. Graduate School of Business, Stanford University.

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health* **36** (12) 1412–1416.

Dwork, C. (2008). An ad omnia approach to defining and achieving private data analysis. *PinKDD'07: Proceedings of the 1st ACM SIGKDD International Conference on Privacy, Security, and Trust in KDD*, 1–13). Springer-Verlag.

EDM Council. (2011). Semantics Repository. Retrieved from EDM Council web site: `http://www.edmcouncil.org/sec_semantics.aspx`.

Enders, W. (2009). *Applied Econometric Time Series*. Wiley.

Elliott, G., C.W.J. Granger, C.W.J. & Timmermann, A. (eds) (2006). *Handbook of Economic Forecasting*, vol. 1, Elsevier.

Fellegi, I., & Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association* **64** (328) 1183–1210.

Flood, M. (2009). Embracing change: financial informatics and risk analytics. *Quantitative Finance* **9** (3).

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: a survey on recent developments. *ACM Computing Surveys* (CSUR).

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. IN *STOC09* 169–178. ACM.

Gross, F. (2010) International reference data utility: a necessary infrastructure for measuring systemic risk. Presentation at the Measuring Systemic Risk Conference – The Milton Friedman Institute, Chicago. `http://mfi.uchicago.edu/events/20101215_systemicrisk/ppts/12.16.2010_FGross.ppt`.

Grunfeld, Y., & Griliches, Z. (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics* **42** (1) 1–13.

Hanson, S. G., Pesaran, M. H., & Schuermann, T. (2008). Firm heterogeneity and credit risk diversification. *Journal of Empirical Finance* **15** (4) 583–612.

Hernandez, M. A., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I. R., et al. (2010). Unleashing the power of public data for financial risk measurment, regulation and governance. In *WWW2010*.

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer-Verlag.

Kelejian, H. H. (1980). Aggregation and disaggregation of non-linear equations. In *Evaluation of Econometric Models*, J. Kmenta, & J. Ramsay (eds.) Academic Press.

Kinlaw, W. B., Kritzman, M. and Turkington, D. (2011). Toward determining systemic importance. MIT Sloan Research Paper No. 4940-11.

Kritzman, M., Li, Y., Page, S., & Rigobon, R. (2010). Principal components as a measure of systemic risk. MIT Sloan School of Management.

Machanavajjhala, A, Kifer, D. Gehrke, J. & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1** 1.

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of large sparse datasets. (How to break anonymity of the Netflix prize dataset). In *Proc. of 29th IEEE Symposium on Security and Privacy*, 111–125). IEEE Computer Society.

Office of Financial Research. (2010). Office of Financial Research Statement on Legal Entity Identification For Financial Contracts. Retrieved from `http://www.treasury.gov/initiatives/Documents/OFR-LEI_Policy_Statement-FINAL.PDF`.

Rivest, R. L., Adleman, L., & Dertouzos, M. L. (1978). On data banks and privacy homomorphisms. *Foundations of Secure Computation*.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* **15** (3) 351–357.

Shilling, H. (2007). *Money Market Fund Industry 2007 Outlook Survey Results*. Moody's Investors Service.

Stein, R. M. (2012). The role of stress testing in credit risk management. *Journal of Investment Management*, Forthcoming.

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10** (5) 557–570.

van Garderen, K. J., Lee, K., & Pesaran, M. H. (2000). Cross-sectional aggregation of non-linear models. *Journal of Econometrics* **95** (2) 285–331.

Wang, H., & Liu, R. (2011). Privacy-preserving publishing microdata with full functional dependencies. *Data & Knowledge Engineering* **70** (3) 249–268.

Winkler, W. E. (2006). Overview of record linkage and current research directions. Research Report, US Census Bureau, Statistical Research Division, Washington, DC.