2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)

# How do firms transact? Guesstimation and Validation of Financial Transaction Networks with Satisfiability

Christos Tsigkanos[∞], Alessio Arleo[+], Johannes Sorger[*,×], and Schahram Dustdar[∞]
[∞]Distributed Systems Group, TU Wien, Vienna, Austria
[+]Visual Analytics Science & Technology, TU Wien, Vienna, Austria
[*]Complexity Science Hub, Vienna, Austria
[×] IIASA, Laxenburg, Austria

## Abstract

*Knowledge of monetary flow between firms can give a significant advantage both from a profit or research point of view. So-called firm-to-firm transaction networks are valuable in analyzing a market or an economy. However, such detailed and complete data is seldom available. In this work, we aim at supporting economists by reusing available financial information from different sources at different levels of detail and completeness. With our technique, experts' domain knowledge can be fused together with publicly available information to extract a representative, coherent instance of the transaction network. Supporting underspecification is important, as experts may develop partial econometric models. Our technique fills such blanks by systematically guesstimating missing information. Our approach builds upon formal foundations of satisfiability modulo theories and thus obtained transaction networks respect constraints imposed by domain knowledge and input data sources. We outline a taxonomy of general data types in the domain, and we programmatically construct formal predicates describing them. We demonstrate both guestimation of missing information of a transaction network and validation of external, expert-provided models. Finally, we investigate feasibility and performance of the advocated technique over a fragment of the Austrian economy.*

## 1 Introduction

Understanding what moves a financial market and what lies beneath its complex mechanics is crucial. Such knowledge could give a significant edge in investments but also in forecasting and assessing whether or not how specific economic policies could affect regions, sectors or individual companies (or "firms"). There is no scarcity of data concerning the trends of an economy, e.g., from newspapers, websites, and papers; however these sources often only tell a part of the story. Granular information about specific companies are available for purchase or for academic use, but the contained information only describes the individual performance of each firm, thus making it more difficult to extract interesting relations. Financial flows on a macro-economic level, i.e., between industry sectors, are usually publicly available [16, 1]. Such macro-economic trends describe the *high-level* effects of what each firm does as a single "agent" in the financial system within its local environment when it interacts with others. Firm-to-firm interactions can be modeled as outbound cash flows (from one company to the other), each one represented by the money an entity must pay when conducting a transaction with another. They include cash paid to suppliers, services acquired from other companies and taxes paid on income.

Knowledge of all transactions of outbound cash flow within some scope (e.g., a region or country), would form a weighted directed graph, where firms are nodes and edges capture monetary flows between them. This firm-to-firm transaction network can be considered the "missing link" between "micro" (single-firm-related) and "macro" data. However, it is typically not available as companies do not tend to publicly share their transaction data. When such data is available though, it is not clear whether it is complete and/or trustworthy. We therefore aim at supporting economists and financial analysts in their investigation of economies by reusing available financial information from different sources at different levels of detail and completeness.

Economists attempting to build econometric models frequently experience data scarcity. In this case it is considered to be a reasonable alternative to guess rather than to estimate parameters of such models [3]. This notion is captured by the concept of "guesstimation", which is an estimation made by guesswork or conjecture [17]. Such estimation is the first step of any economic empirical research [8], and has been used as a building block for understanding and modeling existing economies [7, 6]. In this paper, we therefore investigate the problem of fusing information sources in order to "guesstimate" (or validate) a representative instance of a firm-to-firm transaction network.

978-1-7281-1337-1/19/$31.00 ©2019 IEEE
DOI 10.1109/IRI.2019.00017

The benefit of our approach is thereby two-fold: first, we provide a technique that fuses different information sources in order to obtain a representative, coherent instance of a firm-to-firm transaction network. Secondly, our technique allows experts, who have independently developed a financial model (as is typical in the domain), to complement and validate their work. This expert-provided model may be partial, i.e., not describing information for (or between) all firms. In this case, missing information is systematically guesstimated. Moreover, the expert-provided model may be faulty, in the sense that the information it contains does not respect the inherent data constraints (such as the exact total monetary outflow of a sector). In this case, our technique serves as a tool for expert model validation.

Our technique is based on the formal foundations of *Satisfiability Modulo Theories* (SAT/SMT [5]). Therefore, generated transaction network respect inherent constraints provided by supplied domain knowledge and input data by construction. In this paper, we study the procedural aspects and implications of creating a model ("How do we find an allocation?"), rather than qualitative issues ("What makes a good allocation?") [9], the latter being out of our scope.

Our contributions are as follows. (i) We provide a taxonomy of typical information sources in the domain (Section 2). (ii) We outline a methodology which enables systematic reasoning of firm-to-firm transaction networks by fusing different domain-specific data sources (Section 3), and (iii) we specify a formal encoding of data types for programmatically constructing predicates describing them (Section 4). Finally, (iv) we demonstrate how the technique advocated can be used for both guesstimation of missing information in a transaction network and for validation of external, expert-provided models (Section 5). We thereby investigate feasibility and performance of the technique advocated over a fragment of the Austrian economy.

## 2 Information Source Taxonomy

In this section, we provide a taxonomy of the different financial data sources that we found relevant in the context of the guesstimation of financial transaction networks. We thereby separate data types into three categories as indicated in the top layer of Figure 1.

**Domain Information.** Data in this category concerns information that is measured and validated (e.g., through government institutions), such as financial reports. We thereby differentiate between two granularity levels, i.e., macro and micro data.

*Macro data.* Macro data refers to information describing the macroscopic effects and trends of the economy over a specific region and/or sector. A group of financial entities within the same sector exhibits a measurable collective transaction flow to entities in another sector. *IO tables* [16] reporting annual sector-by-sector investment are an example for such macroscopic data.

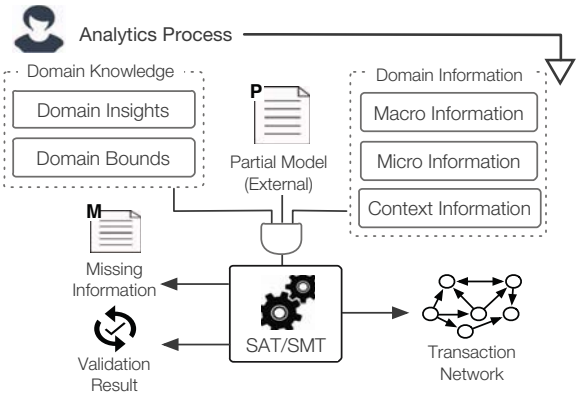


Figure 1: Guesstimation and validation of firm-to-firm transaction networks within an analytics process.

*Micro Data.* In contrast to macro data, micro data describes more fine grained information on specific financial entities or groups of them. Such information includes the location, financial performance parameters (such as cash flow, personnel expenses, etc.) and the operating sector, in respect to a specific time frame. Data on this level of detail is typically cultivated and maintained by financial authorities or policy institutes. The *Sabina* [2] database is an example for microscopic data.

*Context Information.* Context information describes data that is is not directly related to finances; examples may be population data, or the distribution of health-care or education facilities (e.g. universities) over a specific region. This information can be supplied both on a "micro" or "macro" level. Context information exists thus "orthogonally" to finance data, but can be relevant for predicating about relationships to financial information.

**Domain Knowledge.** Data in this category concerns information that, while in theory still being measurable and verifiable, is not collected and verified by official institutions. It rather stems from expert insights and experience and is thus placed in its own category.

*Domain Insights.* Individual insights into specific aspects of an economy can further constraint the problem. Examples for such insights would be knowledge of specific business relations, such as the regular supplier of a certain product to a certain company.

*Domain Bounds.* Such information includes absolute truths that are applicable to the domain analyzed, thus further constraining the domain. Examples of this may be that firms have non-negative numbers of employees, or that companies transact with at least another.

**Modeled Data.** Modeled data concerns information that has been generated through an artificial (and in most cases simplified) model of an economy or a market. It is therefore not measured, but it is verifiable through comparison with equivalent domain information data.

*Expert Model.* Computational economics yields a wide array of knowledge and model guesstimation methods. Our methodology can be used to complement/expand existing firm-to-firm transaction models obtained through third-party means. Such validation typically is part of the model design process; the financial expert might submit models in the early design stages for validation.

## 3 Guesstimation and Validation

In this section, we briefly outline the methodology that our approach supports. Thereafter, we illustrate and motivate its application based on an exemplary use case scenario, i.e., the running example. The lower part of Fig. 1 depicts the three potential use-cases of our approach: the generation of a firm transaction network, completing the missing information in a partial external model, and the validation of an external model. The formal representation for each use-case will be demonstrated in Sec. 4.

**Methodology.** With the overall aim of supporting a design process, our methodology utilizes expert-provided models and knowledge, and targets guesstimation and validation.

*Guesstimation.* Guesstimation describes the process of systematically estimating values that are not specified within the problem, including generating low-level, i.e., firm-to-firm transactions. The minimal required components for the guesstimation process are thereby macro data and micro data relating to the same financial entities, as well as a specification of domain bounds. The macro data thus must be some aggregate of missing micro data transactions, i.e., the set of firms (micro data) has to be contained within the macro data. Domain bounds thereby act as a constraint to produce realistic, representative results.

*Additional Constraints.* As an optional step, in order to increase the accuracy of the generated model, additional information (i.e., domain knowledge and context knowledge) can be drawn upon to create further constraints on possible solutions.

*Expert Model Completion.* A partially-specified model which is externally-sourced can be supplied as additional input, in order to guesstimate any missing information including transactions within the firm-to-firm transaction network.

*Validation.* An expert-provided model may be faulty, in the sense that information it contains does not respect the constraints inherent to the supplied domain information and domain knowledge. In this case, a model under development can be included as additional data source. If constraints or bounds are validated by the additional input, a notification of violation is raised, allowing analysts to incrementally fine-tune their model.

**Running Example.** Consider the exemplary setting of a financial analyst investigating 5 Austrian firms; she seeks to build a model that captures the outbound cash flows between them. Firms generally operate within a sector, by law have tax residence in an Austrian region and have a certain number of employees. By consulting a policy institute, she gathers partial information (*micro data*) about them: companies 1 and 2 are in the technology sector while companies 3 and 4 are in agriculture and manufacturing respectively. Companies 1 and 2 are located in Vienna, company 3 in Salzburg and company 4 in Linz. Companies 1, 2 and 4 have 6, 9 and 23 employees respectively.

She possesses *domain insights*: in Austria, "if a company has more than 2 and less than 10 employees, its combined spending to others is between 40 and 200". By consulting public aggregate tax data (*macro data*), the financial analyst knows the total monetary flows between sectors (i.e., between technology, agriculture, and manufacturing). Furthermore, her knowledge of Austrian economy states that the startup scene is thriving and allows her to formulate constraints based on *context information* – specifically, "if there are more than 3 universities within the same region, local technology firms with at least 3 employees have transactions of greater than 10 monetary units with similar firms in Vienna within the technology sector". Finally, she is developing an econometric *expert model* that – perhaps utilizing domain-specific stochastic or simulation techniques – predicts the transaction flows of companies. However, it does so only partially; the model only yields that the flow between company 4 and 2 is 30, between 1 and 2 is 11, between 4 and 4 is 45, and between 2 and 3, 174.

Notice how the financial expert possesses certain domain information (e.g., aggregate tax data, or startups in Austria) about some partially specified companies in a given setting – there is no information about where firm 0 is located or which sector it belongs to. She also develops an econometric model to construct the company transaction network. However, two *design questions* arise within such a process:

(DQ1) How can missing information be estimated? Estimation entails a mixture of guesswork and calculation, to produce reasonable values for missing information.

(DQ2) Does the (possibly partially specified) econometric model developed by the expert respect the known constraints? In this case, one seeks to evaluate if the expert-provided model is valid.

## 4 Guesstimation with Satisfiability

To tackle guesstimation and validation, after first defining the components of the problem we demonstrate how each category in the taxonomy can be formally represented. Subsequently, we produce a Conjunctive Normal Form (CNF) encoding of the problem that is suitable as input to a solver. The output of the process is a valid transaction network.

Satisfiability modulo theories (SMT) solving consists in deciding the satisfiability of a first-order formula with unknowns and relations lying in certain theories. The formulas

Table 1: Representation of base company characteristics.

| | Domain | Symbol |
|---|---|---|
| **Industry/Sector Type** | $\mathcal{T} = \{tech, agriculture...\}$ | type(c) |
| **Size** | $\mathcal{S} = \{small, medium, large\}$ | size(c) |
| **Operating Area** | $\mathcal{A} = \{vienna, linz...\}$ | area(c) |
| **Employees** | $\mathcal{N}_0$ | empl(c) |
| **Transaction Flow** to t | $\mathcal{N}_0$ | f(c, t) |

we adopt may contain the usual boolean operators, quantifiers over finite sets, as well as integer linear arithmetic operators. In the following, we describe construction of such a formula which integrates available data about an economic domain, along with certain constraints. Constraints encode knowledge about financial actors in the domain, and may be specified in different degrees of abstraction – they may express arbitrary relationships between firms (or groups of them) that must be respected. Information may refer to known domain micro- or macro- information, but also to external models which are to be submitted for validation.

### 4.1 Fusing Data and Domain Knowledge

We assume a finite set of companies $\mathcal{C}$ representing the problem domain. Each company in $\mathcal{C}$ has certain characteristics, which need to be captured in a formal model. Indicative model characteristics are illustrated in Table 1 – these can be modified and extended for each specific application.

For our example purpose, companies have a size classification (derived through a metric), such as $small$, $medium$ or $large$. Similarly, companies belong to the same group if they operate in the same segment of the economy or share a similar business type, or have a tax residence in a specific location. Such characteristics are captured in respective symbols, whereas the value ranges are drawn from the supplied data sets (Table 1). This way, quantitative characteristics of firms can be formally captured, using integer linear arithmetic for their specification, e.g., by specifying the number of employees. Finally, a firm $\mathsf{c}$ may have a monetary transaction flow to another, reflected by an integer value. To describe transactions, a symbol for each pair is required. For a transaction flow to a firm $\mathsf{t}$, we represent this as $\mathsf{f}(\mathsf{c}, \mathsf{t})$.

Note that arbitrary characteristics of firms can be encoded in a similar manner, utilizing, e.g., finite sets or integers as possible values to identify subsidiaries, affiliates, or financial activity types. These characteristics may typically indicate financial measures such as expenditures, credit or values of financial instruments. We limit our description to the characteristics given in Table 1 as they are rather indicative, and useful for the evaluation study in Sec. 5. Recall the taxonomy of Sec. 2; various firm characteristics or more abstract information about their relationships (or groups thereof) can be encoded. This is based upon the defined base symbols (e.g., Table 1) and specified in the following.

**Macro Data**. Recall that macro information captures monetary flows between groups of firms that share a common characteristic. Such information can be encoded in two steps, illustrated in Formula 1 for aggregate flows between firms belonging to two sectors ($\mathcal{T}_s$ and $\mathcal{T}_t$).

$$\sum_{i \in \mathcal{T}_s} \sum_{j \in \mathcal{T}_t} \mathsf{f}(\mathsf{i}, \mathsf{j}) = \mathsf{l}, \text{ where } \mathcal{T}_\mathsf{t} = \{\mathsf{k} \in \mathcal{C} \mid \mathsf{type}(\mathsf{k}) = \mathsf{t}\},$$
$$\mathcal{T}_\mathsf{s} = \{\mathsf{k} \in \mathcal{C} \mid \mathsf{type}(\mathsf{k}) = \mathsf{s}\}. \quad (1)$$

Firstly, the sum of flows that a firm exhibits to others in the target group $\mathcal{T}_t$ is considered. Secondly, the sum of flows of all firms from the source group $\mathcal{T}_s$ towards firms within the target group, yields the total aggregate flow. Notice that such an encoding (as shown in Formula 1) must occur for every pair of firm groups within the macro data; for instance, between every pair of sector investment indicators within an IO table [16].

**Micro Data**. The specification of the micro data encoding simply requires the provision of the concrete values from the supplied data. For our motivating example, information from the policy institute states that company A resides in Vienna: thus, $\mathsf{area}(\mathsf{a}) := \mathsf{vienna}$, etc. Depending on the granularity, *context information* is encoded akin to the specified macro or micro data formulas.

**Domain Insights**. Particular expert insights concerning the investigated scenario can be encoded as additional constraints. Such information may express arbitrary relationships between companies (or groups of them). For our motivating example, expert knowledge tells that firms having between 2 and 9 employees have an outgoing cash flow between 40 and 200; this is encoded in Formula 2. Additionally, arbitrary domain information in respect to context information that an expert may be aware of can be specified as well. Formula 3 states that if the number of universities in a region exceeds 3 (through some predicate $\mathsf{uc}(\mathsf{area})$), then technology companies based in the region have at least 3 employees and transact with an amount of at least 11 monetary units with companies in Vienna in the technology sector.

$$2 \leq \mathsf{empl}(\mathsf{c}) < 10 \iff 40 < \sum_{\mathsf{i} \in \mathcal{C}} \mathsf{c.f}(\mathsf{i}) \leq 200, \forall \mathsf{c} \in \mathcal{C}. \quad (2)$$
$$(\mathsf{uc}(\mathsf{a}) > 5 \wedge \mathsf{area}(\mathsf{c}) = \mathsf{a} \wedge \mathsf{type}(\mathsf{c}) = \mathsf{tech})$$
$$\rightarrow (\mathsf{empl}(\mathsf{c}) > 3 \wedge \sum_{\mathsf{j} \in \mathcal{T}_\mathsf{t}} \mathsf{f}(\mathsf{c}, \mathsf{j}) > 10), \forall \mathsf{c} \in \mathcal{C};$$
$$\mathcal{T}_\mathsf{t} = \{\mathsf{k} \in \mathcal{C} \mid \mathsf{type}(\mathsf{k}) = \mathsf{tech} \wedge \mathsf{area}(\mathsf{k}) = \mathsf{vienna}\}. \quad (3)$$

**Domain Bounds**. Domain bounds capture statements in accord with facts which apply universally. For instance, companies may have only one size (i.e., a company cannot be simultaneously small and medium). Such a statement can be formalized in classical first-order logic (Formula 4 for an exactly one area constraint). Other such facts may capture that the number of employees or transaction flows are not negative; or they may provide reasonable maximum and minimum values for, e.g., the number of branches of a company, to further constrain a generated transaction network.

$$(\mathsf{area}(\mathsf{c}) = \mathsf{a}_1 \oplus \mathsf{area}(\mathsf{c}) = \mathsf{a}_2) \\ \wedge \neg(\mathsf{area}(\mathsf{c}) = \mathsf{a}_1 \wedge \mathsf{area}(\mathsf{c}) = \mathsf{a}_2), \\ \text{where } \mathsf{a}_1, \mathsf{a}_2 \in \mathcal{A} \text{ are pairwise distinct and } \mathsf{c} \in \mathcal{C}. \quad (4)$$

**Expert Model.** A model produced by a domain-specific technique is integrated by explicitly setting values to the company characteristics that it describes. Note that such an expert-provided model can be partial or completely specified; it may record information about limited number of firms (depending on the coverage of $\mathcal{C}$ it provides) or their characteristics (e.g., Table 1). Specifying explicit values for firm attributes found within an expert-provided model is performed similarly to the micro data specification. For our example, the analyst partially specified 4 transaction flows.

Finally, the conjunction of macro and micro data, context knowledge, domain insights, domain bounds, and expert-provided models collectively describe the input to an SAT/SMT solver. The formula is a conjunction of a finite collection of literals, thus in CNF form. Note that in practical settings, this programmatic construction can lead to formulae with large numbers of symbols, since it is often of interest to apply the specified constraints to all companies in the considered domain.

### 4.2 Transaction Network through Satisfiability

For our network generation and validation purposes, we essentially ask for an assignment of attributes (Table 1) to firms ($\mathcal{C}$) that respects all the constraints specified. This process is referred to as satisfiability testing. The process decides whether or not a satisfying assignment for the unknown components within the problem exists, i.e., an assignment of the variables that renders the specified formula true. Notice how the values of $\mathsf{f}(\mathsf{c}, \mathsf{t})$ symbols make up the firm-to-firm transaction network, by reflecting the transaction flow that company $\mathsf{c}$ incurred on company $\mathsf{t}$. Given a problem specification, the computation of the firm-to-firm transaction network can be achieved by employing a SAT/SMT solver, from which a satisfiable assignment is requested.

Certainly, if there exists no satisfiable solution to the conjuction of the formula components (as discussed in Sec. 4.1), a firm-to-firm transaction network cannot be computed. Assuming that the constraint specification is correct, the logical conclusion is that the expert-provided model is faulty. In this regard, the generation of a satisfiable solution lends itself to model validation. Recall our running example; the expert's econometric model estimated that the transaction flow of company 1 to company 2 is 10 units. However this is not valid, as it does not respect the other constraints (namely, domain knowledge about startups as well as locations in Vienna). A value of 11 for that transaction flow for instance, is valid. This is something that is non-trivial to observe and verify, especially when the model is complex and firms have complex interplexed relations. The corrected network is illustrated in Fig. 2. Given an unsatisfiable formula, the subset of clauses whose conjunction is still unsatisfiable (its *unsatisfiable core*) can often be produced. Practical methods exist for computing and expressing the unsatisfiable core as a resolution graph proving the unsatisfiability of the original problem [11]. We identify the application of such methods as a promising avenue of future work, where facilities for model debugging can be provided to computational economists.

If the formula is satisfiable, the firm-to-firm transaction network produced is a valid solution to the problem. However, note that there may exist many other satisfying solutions. This represents the biggest threat to validity for the approach advocated, and raises issues of model quality. Quality in this setting, refers to how close to reality are the values that were produced by the solving procedure. Although they certainly satisfy the constraints, they may be far from actual financial behaviour. This problem amounts to the "vagueness" of constraints specified – the more expert knowledge is introduced to the problem, the more "quality" the model has. Nevertheless, acknowledging this natural shortcoming, a feature that this approach provides is guesstimating missing parts of an expert-provided partial model. In principle, it may range from completely unknown to almost fully specified, with some information missing. In the former case, results may be untrustworthy, while in the latter the approach advocated can aid in filling gaps in the model which otherwise would be left unspecified.

Generation of a firm-to-firm transaction network in practice, involves programmatically building formulae as outlined in Sec. 4.1 and interacting with a SAT/SMT solver. Given a problem instance, the process entails the following steps: (i) the appropriate CNF formula representation is encoded (ii) a solver is invoked upon it, and (iii) the firm-to-firm transaction model is computed from the satisfiability assignment of the solver. We note that any SMT-LIB compliant solver can be used; the satisfiable assignment obtained is then used to derive the network.

## 5 Evaluation

For evaluating the proposed approach, we developed tool support and a proof-of-concept implementation based on the Z3 solver [10]. The technique we advocate for information guesstimation and validation is based on the satisfiability of SAT/SMT, a highly computationally expensive operation. To this end, our evaluation goals target the feasibility of our approach in terms of realizability and performance. Concretely, we aim to (i) demonstrate the feasibility of information guesstimation and validation based on real-world firm data (consisting of macro, micro, context information and domain bounds); and (ii) assess the performance for the validation of expert-provided models with respect to degrees

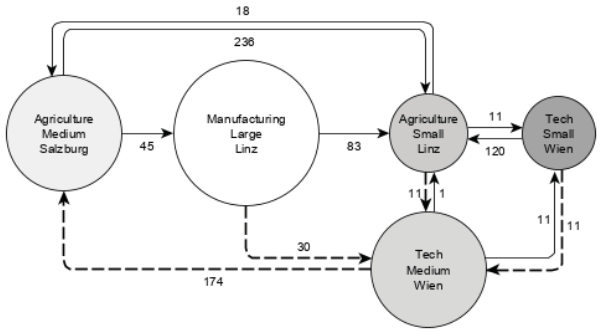


Figure 2: Guesstimated transaction network of the (corrected) example. Attributes (type, size, location) are reported in nodes, while weights on edges capture the monetary amount. The transactions of the original expert model are dashed, while the guesstimated ones have solid lines.

of model under-specification. We present our evaluation setup in Sec. 5.1 and the experimental results obtained in Sec. 5.2. We conclude with a discussion in Sec. 5.3.

### 5.1 Experiment Setup

We generate input configurations (termed *problem instances*) from available public financial information, based on the methodology outlined in Sec. 3 and using the information sources outlined in the following. *Macro Data* are publicly available IO tables [16] that capture financial activity between economy sectors in Austria for the year of 2017. Each cell in the IO table is encoded as per Formula 1. *Micro Data* are obtained from an anonymized subset of the *Sabina* dataset [2] and describe 400 firms in Austria. They include the firms' location, financial sector, personnel employed and balance sheet amounts, which provide explicit assignments to firm characteristics. *Domain Insights* can be arbitrary since they are expert-provided; to this end, our experimental setup specifies that each firm transacts with at least 10 others in Austria. For our experimental setup, we synthesize random *Expert Models* capturing explicit company transactions. To ensure uniformity, the synthesized models are valid. Finally, *Domain Bounds* encode ground truths; firms may have non-zero sums of flows to others, and they always have a single associated sector, location and size.

Problem instances are then obtained by step-wise increasing the cardinality of the sets of considered companies by incremental steps of 10, from 100 to 400. As a result, each problem instance increases in the number of constraints with the considered firms (micro data) and of the specified expert model. However, note that specified macro data, domain context and domain bounds are equivalent for all problem instances. As such, we can assess the performance of our technique by controlling the firms size. To ensure uniformity, companies are sourced incrementally from the same set of 400; a set of companies of greater cardinality includes the same companies of a set of smaller cardinality. Finally, recall that validation and guesstimation are results of the same process, so every instance represents a single problem.

Recall that an evaluation objective is to assess the validation performance with respect to the degrees of guesstimation. To this end, we additionally generate partially specified models for each problem instance. A partially specified expert model includes an explicit assignment of financial flows of some percentage of firms, assumed to be sourced from an expert-provided model. Since we are not concerned with economic aspects and actual values have no influence to the process, we use random values. This occurs for every problem instance considered. Our prototypical implementation employs the constraint generation procedure described in Sec. 4 and is deployed on a laptop computer featuring an Intel i5 2.3GHz processor and 15G RAM. We evaluate performance in practice by drawing from the problem instance dataset previously described; for each number of companies considered, we invoke the procedure for partially-specified expert models of 99%, 50% and 10%.

### 5.2 Experiment Results

In Fig. 3, the number of companies considered over guesstimation/validation calculation time is illustrated – each data point is a single problem instance (i.e. a macro-micro-domain-facts-model configuration). The size of the resulting SAT/SMT encoding in symbols (as per Sec. 4) corresponding to the underlying satisfiability problem is represented by the shading of points. The number of symbols within the formulae range from $500k$ to $6.5M$. The trend shows the solving time increasing with the formula size.

Observe the difference in performance for company sets of the same cardinality (i.e., within a vertical line in Fig. 3). Within the lower end, programmatic formula construction and startup overheads yield quite similar performance. As cardinality increases, the hardness of the problem rises and a divergence starts to appear as the problem size increases beyond 150 firms. For instance, for 400 firms, low specification (10%) of a concrete model (i.e., increased amount of the guesstimation workload component) leads to a model being generated over 64 minutes. Conversely, for the same firm size, partial specification of 40% and 90% leads to networks obtained in 50 and 44 minutes, respectively.

### 5.3 Discussion

Using the technique advocated, guesstimation and validation in a financial setting are feasible over real firm information sources. We especially note that the validation component of the technique guarantees that the network generated respects the constraints specified in the initial steps of the methodology, due to its satisfiability foundations.

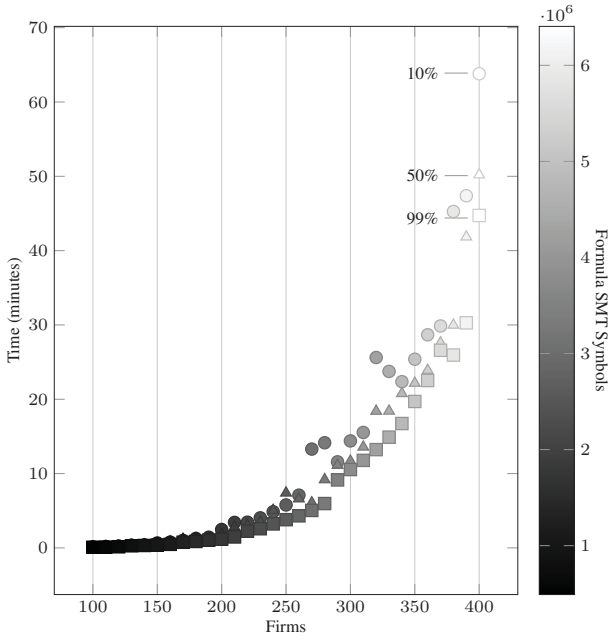


Figure 3: Validation/guesstimation time over increasing firm set sizes, with respect to different degrees of partially specified models (□ : 90%, △ : 50%:, ○ : 10%). Shading of points indicates the symbols of the resulting SMT encoding.

Regarding threats to validity, we first and foremost note that the quality of the inferred model depends on the quality of the constraints specified. Naturally, in situations where the problem is underspecified, the solving process may assign transaction flows that are not representative. However, we believe that our approach has merit for validation of expert-provided, domain-specific models obtained through other computational finance means (e.g., SARA [9]), and may aid the design process by spotting faults that could have gone unnoticed. To this end, we view our approach as complementary to domain-specific ones. To illustrate the combined validation-guesstimation facilities, we compared naive guesstimation (i.e., an underspecified model) with models with increasing percentage of specification.

Regarding the applicability of the proposed approach, scaling up to realistic numbers of companies – e.g. all in a given country – poses challenges for the computational demands of the satisfiability solving involved. For reference, firms in Austria contained in [2] are in the range of tens of thousands. To tackle the problem for large sets of companies (e.g., in Austria), other techniques need to be investigated, like clustering companies with a similar economic profile – we regard this as an important avenue of future work. Such techniques have the potential to constrain the problem significantly and even render the computation possible online, to be integrated e.g., in an online analytics process and data pipeline. Finally, due to the satisfiability solving [12] that underlies the validation process, different problem instances perform differently; hardness of SMT satisfiability is naturally beyond our scope. However, we note that optimizations of the formulation of the constraints may yield benefits in terms of performance.

## 6 Related Work

Charemza [8] introduces a guesstimation-based approach to define the parameters of a large model used with the intention of forecasting. The paper describes an iterative methodology called *Repetitive Stochastic Guesstimation* (RSG), which refines the estimated data starting from an initial "guess" and actual measurements. Therefore, the evolution of the model is strongly influenced by the choice of the initial values: as downside, this could "tempt" the researcher to change assumptions until a desired result is obtained, but this is also true for traditional econometric models. Later, it was shown that the RSG algorithm converged [3]. This approach fueled the use of Guesstimation in creating models to empower economic forecasting or decision dupport. As a matter of fact, RSG is part of the algorithm pipeline to determine the inter-country econometric model of Russia, Ukraine and Belarus. [6]. In this case, the RGS is used to provide an initial estimate of the parameters of the model. Once the initial estimate has been tested for stability, its parameters are passed to a "Long-run Adjustment Model" (LAM) [7]. It consists of blocks of trade equations that take into account the principal macroeconomic characteristics of the investigated economies (such as investment, consumption, consumer prices, wages, employment, etc.) and the market interactions between countries. The solution of the equations represents the macro-economic variables of the model. The similarities with our approach include the presence of a guesstimation technique in the modeling pipeline and the use of system constraints to obtain the final model. However, LAM aims at estimating macroeconomic parameters over country-wide aggregated data and guided high-order trade relations between countries. Moreover, LAM is tailored for east-european countries, while our technique can be applied to any bounded region. Estimation techniques are also used to fill the missing gaps in other information sources. In the paper by Rueda-Cantuche et al. [15] the authors survey existing estimation approaches to complement "Use Tables" (i.e. an ingredient to the compilation of IO Tables) in the central Europe financial context in absence of "superior data". As in our evaluation, different scenarios with increasing uncertainty (i.e. missing data) are prepared, and different estimation approaches are evaluated against them. Differently from our technique, the paper suggests estimating the missing values leveraging the construction process of such tables, rather than relying on stochastic approaches. Moreover, it assumes that some (complete) data

about the previous years is available. To estimate the parameters of an economic model, simulation is another popular approach. Assenza et al. [4] propose an economic Agent Based Model (ABM) whose agents are households, firms and banks. The behavior of each category of agent is encoded in the model. The agents will operate freely and independently during each simulation step, and their individual actions will impact the system as a whole. There are several inputs to modify during each simulation step, such as the cardinality of each set of agents. After the simulation, it is possible to evaluate, a posteriori, how the system evolved (by comparing the starting point and the initial parameters with the conditions of the system at the end point). Poledna [14] et al. extended this approach specifically for the Austrian economy, creating a set of agents encoding a behaviour that imitates their real-world counterparts as close as possible. ABMs are considered useful for forecasting, and can remain valid also for large models. Differently from our approach, however, in both cases the goal is to get an understanding of the phenomena as a whole rather than focusing on the single agent's transactions. Finally, network reconstruction is used to investigate systemic economic risk by considering firm-bank and interbank relations [13].

## 7 Conclusions and Future Work

Within the context of supporting economic analytic processes, we presented a technique to guesstimate firm-to-firm financial networks by fusing information from different sources at different levels of detail and completeness. The theoretical foundations of our technique lie within satisfiability modulo theories, thus obtained transaction networks respect constraints inherent in domain knowledge and input data by construction. We outlined a taxonomy of typical data sources in the domain, and we programmatically constructed formal predicates describing them, to support the systematic guesstimation of missing information. We finally demonstrated how our approach could be applied to the validation of existing partial models and we framed our work among current guesstimation approaches within finance.

Regarding future work, considering the perspective of practitioners aiming to use our technique, integration within a visual analytics process is highly desirable. Domain specific languages, interfaces and tooling integration would go a long way in supporting complex data pipelines and analytics processes. Acknowledging its interdisciplinarity, such user interfaces would be essential for validation of the approach advocated with financial experts. Finally, we focused on static networks – it would be of major interest to evaluate inclusion of temporal data as a distinct information source and their respective encoding.

## Acknowledgments

Research partially supported by the TU Wien Research Cluster SmartCT and FFG Austria project Nr. 857136.

## References

[1] Database - eurostat. https://ec.europa.eu/eurostat/web/esa-supply-use-input-tables/data/database. Accessed: 2019-06-26.

[2] Wirtschaftsuniversitt wien: Sabina - info - datenbanken. https://www.wu.ac.at/bibliothek/recherche/datenbanken/info/sabina/. Accessed: 2019-06-10.

[3] A. Agapie. Convergence of the guesstimation algorithm. *Communications in Statistics-Theory and Methods*, 38(5):711–718, 2009.

[4] T. Assenza, D. D. Gatti, and J. Grazzini. Emergent dynamics of a macroeconomic agent based model with capital and credit. *Journal of Economic Dynamics and Control*, 50:5–28, 2015.

[5] C. Barrett and C. Tinelli. Satisfiability modulo theories. In *Handbook of Model Checking*. Springer, 2018.

[6] W. Charemza, Y. Kharin, S. Makarova, V. Malugin, V. Majkowska, Y. Raskina, Y. Vymyatnina, and A. Huryn. Intercountry econometric model of the economies of belarus, russia and ukraine. 2007.

[7] W. Charemza, S. Makarova, and V. Parkhomenko. Lam modelling of east european economies: Methodology, eu accession and privatisation. EcoMod Network, 2002.

[8] W. W. Charemza. Guesstimation. *Journal of Forecasting*, 21(6):417–433, 2002.

[9] Y. Chevaleyre, P. E. Dunne, U. Endriss, J. Lang, M. Lemaitre, N. Maudet, J. Padget, S. Phelps, J. A. Rodriguez-Aguilar, and P. Sousa. Issues in multiagent resource allocation. *Informatica*, 30(1), 2006.

[10] L. De Moura and N. Bjørner. Z3: An efficient smt solver. In *Intl. Conf. on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.

[11] N. Dershowitz, Z. Hanna, and A. Nadel. A scalable algorithm for minimal unsatisfiable core extraction. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 36–41. Springer, 2006.

[12] E. Nudelman, K. Leyton-Brown, H. H. Hoos, A. Devkar, and Y. Shoham. Understanding random sat: Beyond the clauses-to-variables ratio. In *Intl. Conf. on Principles and Practice of Constraint Programming*, pages 438–452. Springer, 2004.

[13] S. Poledna, A. Hinteregger, and S. Thurner. Identifying systemically important companies by using the credit network of an entire nation. *Entropy*, 20(10):792, 2018.

[14] S. Poledna, M. Miess, and S. Thurner. Economic forecasting with an agent-based model. 2017. Preprint.

[15] J. M. Rueda-Cantuche, A. F. Amores, J. Beutel, and I. Remond-Tiedrez. Assessment of european use tables at basic prices and valuation matrices in the absence of official data. *Economic Systems Research*, 30(2):252–270, 2018.

[16] M. P. Timmer, E. Dietzenbacher, B. Los, R. Stehrer, and G. J. de Vries. An illustrated user guide to the world inputoutput database: the case of global automotive production. *Review of International Economics*, 23(3):575–605, 2015.

[17] L. Weinstein and J. A. Adam. *Guesstimation: Solving the world's problems on the back of a cocktail napkin*. Princeton University Press, 2009.