Family Link Detection in Uncertain Settings with MV-Datalog+-

Marta Bernardini[1], Matthias Lanzinger[2], Rosario Laurendi[1] and Stefano Sferrazza[2,3]

[1]*Banca d'Italia, Italy. The views and opinions expressed in this paper do not necessarily reflect the official policy or position of Banca d'Italia.*
[2]*University of Oxford, United Kingdom*
[3]*TU Wien, Austria*

Abstract

We report on a novel work-in-progress system for inferring family links in company ownership graphs. Our system combines fuzzy observations of possible family links from anagraphic data with MV-Datalog$^{\pm}$, a new framework for many-valued logical reasoning in Datalog$^{\pm}$. With this approach we achieve levels of flexibility and explainability benefits that is difficult to match by current Machine Learning methods. We describe the key challenges, main components of our approach, and experimentally illustrate its effectiveness at this early stage in its development.

1. Introduction

Company ownership graphs play a central role for central banks, financial authorities and national statistical offices [1, 2, 3, 4] and represent a valuable tool for solving a wide range of problems in the context of banking supervision, credit-worthiness evaluation, anti-money laundering, insurance fraud detection, economic and statistical research. One particular such ownership graph that we consider in this paper is the *Enterprise Knowledge Graph* of Italian companies [5], in which people and companies are the nodes, while the (labelled) edges represent the fraction of company shares owned by a person or a company.

One of the main challenges for detecting regulatory issues in the Enterprise Knowledge Graph is the obfuscation of actual company ownership. Such obfuscation can be highly problematic. For example, according to European Central Bank regulations, company C_y is not eligible as a guarantor for C_x if it is too "close" to it in terms of ownership [6]. It is therefore important to identify potential *close links* in company ownership graphs to effectively detect problematic constructs. One particularly common type of potential close link is through family ties. Oftentimes, company ownership is distributed over family members to obfuscate the actual ownership structure leading to issues as in the example illustrated in Figure 1.

Consequently, if we detect a family link between a person P_a and a person P_b that respectively own a company C_x and C_y, we can infer that even if C_x and C_y do not strictly fulfil the definition of close link, we should prevent both companies from being the guarantor of the other. However, company ownership graphs often are derived from data where family relationships are not explicitly available and must be reconstructed through other methods. The lack of public data makes the task of finding family links highly challenging since civil registers can not share personal data due to privacy regulations. Note that this lack of ground truth data also severely limits the applicability of popular Machine Learning techniques for related problems like Knowledge Graph embedding.

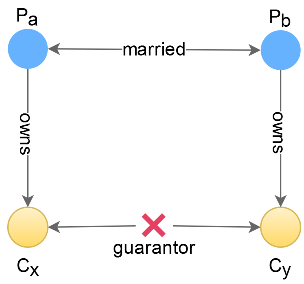

Figure 1: A typical case of hidden close link where C_x cannot act as guarantor for C_y due to the family relationship between P_a and P_b

In this paper, we present novel in-progress work on combining similarity measures that indicate potential family links with new methods for many-valued reasoning over uncertain data in KGs to obtain a unified picture of potential *family links* in the Enterprise Knowledge Graph of Italian companies.

In a first step, we infer connections between persons based on the information pertaining to company ownership and the available anagraphic data, such as addresses and names of company owners. For example, given per-

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK
matthias.lanzinger@cs.ox.ac.uk (M. Lanzinger); ste.sferrazza@gmail.com (S. Sferrazza)
0000-0002-7601-3727 (M. Lanzinger); 0000-0001-9569-4428 (S. Sferrazza)

sons P_a and P_b with the same surname, both having shares in the same companies and living in the same region, we may reasonably infer a family link between them. This hypothesis is strengthened by the knowledge of the Italian economic system, where a large number of companies have families as their major shareholders. We extend such inference beyond exact matches of data, e.g., if P_a and P_b in the scenario above lived in neighbouring regions rather than the same region the inference could still be argued, although with less certainty.

At the same time, family links observe several common logical properties, such as transitivity or symmetry, that can be used to detect further family links through logical reasoning. However, since the similarity metrics produce uncertain observations, the use of classical logical reasoning (which is limited to statements that are either fully true or false) is severely limited in this setting. Rather than reducing our uncertain observations to true/false via thresholds or similar techniques, we propose the use of MV-Datalog$^{\pm}$ [7], a recent extension of Datalog$^{\pm}$ [8] with *many-valued* semantics. Many-valued here refers to the use of *degrees of truth*, which we use to model uncertainty, rather than the classical paradigm of every fact being either absolutely true or false. MV-Datalog$^{\pm}$ allows us to combine the uncertain observations with logical reasoning and certain knowledge, such as a company ownership KG, expert knowledge, or known relationships, in a single framework.

In the following, we discuss how these two components can interact to provide a powerful reasoning system for detecting family links in the Enterprise Knowledge Graph for Italian companies. While the presentation focuses on this specific setting, we believe that our methods can be widely applicable and that uncertain reasoning systems open up a variety of new opportunities for the use of KGs. Details of the problem setting and the derived observations of possible links between persons are presented in Sections 2 and 3, respectively. We discuss the interaction with MV-Datalog$^{\pm}$ for unifying and extending uncertain observations in Section 4. Results of preliminary experiments are discussed in Section 5. We conclude with an outlook of planned next steps in Section 6.

2. The Enterprise Knowledge Graph of Italian Companies

A Knowledge Graph (KG) is a semi-structured data model [9] composed of three components: (i) a *ground extensional component* (or simply extensional component), that is, a set of relational constructs for schema and data, which can be effectively modelled as a *property graph*; (ii) an *intensional component*, that is, a set of *inference rules* over the constructs of the ground extensional component; (iii) a *derived extensional component* that can be produced as the result of the application of the inference rules over the ground extensional component (with the so-called "reasoning" process).

Our target is the Enterprise Knowledge Graph of Italian companies, containing the most updated data at our disposal. We focus on *non-listed companies*. For each of them, the graph contains several features including legal name, registered office address, incorporation date, legal form, shareholders. A shareholder can be either a company or a person, with the standard anagraphic information. Shares can be associated with different legal rights (e.g., ownership, bare ownership, etc.). We focus on all forms of ownership and include in the KG companies having at least one shareholder. The graph counts 11.97M nodes, representing the shareholders, and 14.18M edges denoting share ownership.

3. Extracting Uncertain Family Links

In this section, we describe our technique for extracting uncertain family links from the Italian companies KG. Considering the complex analysis that we aim to perform, taking into account all the person nodes in the graph would lead to the intractability of the uncertain reasoning inference. Furthermore, most of the nodes would not be relevant for the scope of our analysis. Thus, spending computational time over those would not be wise. The first objective is to reduce the search space in such a way that it is likely a family connection exists each time two person nodes are compared.

Our approach is based on the following assumption: people owning a similar amount of shares across the same companies tend to have family connections. According to these remarks, we consider the following three-step process in order to extract family links having a degree of uncertainty. The first two steps aim at reducing the search space through partitioning techniques. In the last step, each pair of nodes in the reduced space is compared to evaluate three different similarity scores. The higher the scores, the more likely the family connection.

(i) The first step involves the computation of weakly connected components (WCCs) by exploiting a more complex concept of the traditional ownership that is called *integrated ownership*. This is a notion of accumulated ownership from a company C_x to a company C_y: it accounts for the ownership that C_x retains of C_y along direct and indirect connections. In terms of the flow of dividends, integrated ownership can be seen as the cumulative flow from C_x to C_y, justified by direct and indirect shareholding [5].

The graph consists of about 21 million connections based on the integrated ownership. This amount of relations

takes into account both direct and indirect links. The number of connections is significantly cut by removing the ones characterized by an integrated ownership value $O < 0.001$. At this point, some of the nodes are no longer linked to any other node in the graph due to the removal of related edges.
This diminished graph presents about 1,3 million WCCs, in average composed by a number of nodes between 3 and 10, see Figure 2[1], while the largest WCC has more than 1,5 million nodes. All the WCCs composed by a single node have not been further considered. According to the WCC properties, a node is only reachable from the nodes within the same component. This feature fully satisfies our previous assumption, since two persons could have a family link only if at least one connection (direct or indirect) through their companies exists.

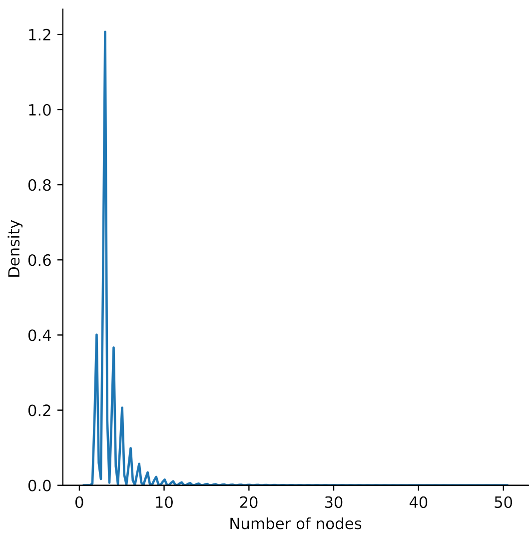

Figure 2: Density distribution of nodes in the WCCs

(ii) In step 2, new partitions are obtained by clustering the persons having identical surnames in the same WCC. The largest WCC contains almost 60.000 clusters.
Figure 3 illustrates an example of the resulting clusters. The bubble colour identifies the WCC. Note that, the same surname could result in more than one cluster, where each cluster belongs to a distinct WCC. For instance, there are 1500 *Rossi* clusters, each one contained in a different WCC. Note that this mechanism allows for finding parent-child or sibling relationships only. However, this technique could be extended to cover other common cases, for example employing address-based clustering.

[1]For the sake of visualization the communities with the higher number of nodes and density next to 0 have been omitted

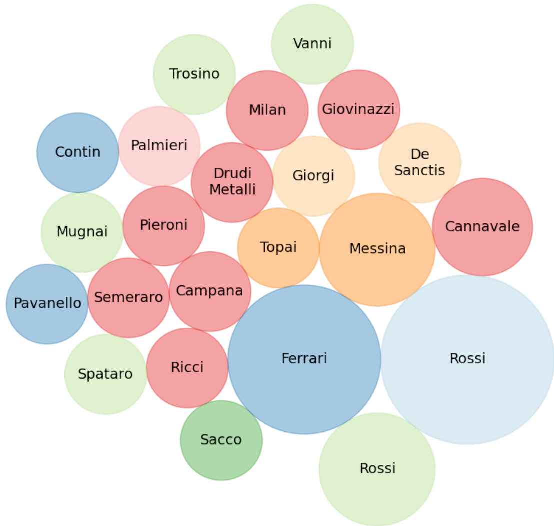

Figure 3: An insight of the clustering result

(iii) In step 3, all the pairs of nodes possibly having a family connection, i.e. all the nodes within the same cluster, are compared to evaluate the similarity scores. The final set of candidates is composed of approximately 41 million pairs.

The similarity score based on the amount of shares across the same companies is computed as follows:

$$own_{P_a P_b} = \frac{\sum\limits_{C_i \in (C_a \cap C_b)} O_{P_{a_i}} + O_{P_{b_i}}}{\sum\limits_{C_i \in C_a} O_{P_{a_i}} + \sum\limits_{C_i \in C_b} O_{P_{b_i}}} \tag{1}$$

Where C_a and C_b represent, respectively, the set of companies where a person P_a and person P_b have a percentage of integrated ownership; $O_{P_{x_i}}$ is the share of a person P_x over a company C_i. In detail, $own_{P_a P_b}$ represents the ratio between the sum of shares owned by P_a and P_b over a set of common companies and the sum of shares owned over all their companies.

Let us consider an example where three persons P_1, P_2 and P_3 have the following shares over the companies C_x, C_y and C_z:

$$\begin{array}{ll} O_{P_{a_x}} = 0.3 & O_{P_{a_y}} = 0.2 \\ O_{P_{b_x}} = 0.4 & \\ O_{P_{c_x}} = 0.3 & O_{P_{c_z}} = 0.8 \end{array}$$

The resulting own scores are:

$$\begin{aligned} own_{P_a P_b} &= 0.77 \\ own_{P_a P_c} &= 0.37 \\ own_{P_b P_c} &= 0.46 \end{aligned}$$

We can notice that the score increases when there is a concentration of interests over the same companies. Indeed, while P_a and P_b ownership revolves around C_x, P_c has a relevant share of another company, C_z. This brings to lower scores associated with P_b,P_c and, in particular, P_a,P_c.

The second score takes into account the string similarity between the living addresses. Given two persons P_a and P_b having address $address_{P_a}$ and $address_{P_b}$, the similarity score is given by the Levenshtein distance [10] as follows:

$$addr_{P_a P_b} = levenshtein(address_{P_a}, address_{P_b}) \tag{2}$$

Finally the third score evaluates the proximity of the address postal code, derived from the living address attribute, as:

$$postal_{P_a P_b} = \frac{k}{n} \tag{3}$$

Where k represents the number of the first common digits shared by the two postal codes and n is the postal code length. For instance, the postal codes 00100 and 00198 have the first three digits in common ($k = 3$, $n = 5$) and the resulting similarity is 0.6.

The result of the process is summarized in the following table:

id_{P_1}	id_{P_2}	...	$own_{P_1 P_2}$	$addr_{P_1 P_2}$	$postal_{P_1 P_2}$
a	b	...	0.77	0.6	0.7
a	c	...	0.37	0.5	0.8
b	c	...	0.46	0.1	0

Table 1
Similarity scores table

4. Uncertain Reasoning with MV-Datalog$^\pm$

Datalog and its more recent extension Datalog$^\pm$ are popular formal languages for rule-based reasoning. They combine high expressiveness with several important properties for use in practice. In particular, they allow for efficient reasoning and explainable answers despite being expressive enough to capture a number of popular formalisms, such as the OWL 2 QL and OWL 2 EL profiles of the OWL 2 for semantic web reasoning, or popular description logics like DL-Lite$_\mathcal{R}$ or $\mathcal{E}$L [11]. At the same time, rule-based languages are designed to be easily understandable and writable by humans, allowing domain experts to introduce their knowledge in the reasoning process. Finally, extensive research in the field has also identified methods and fragments that allow for highly efficient reasoning in these languages (e.g., [12, 13]).

This confluence of desirable properties has led to a recent surge in the use of Datalog and its derived languages in complex Data Analysis tasks [14]. Here we make use of a very recent such extension named MV-Datalog$^\pm$ [7], which extends Datalog$^\pm$ to many-valued (or *fuzzy*) semantics. That is, instead of the classical dichotomy between true and false facts, one considers *degrees of truth*, expressed as rational numbers in the interval $[0, 1]$ (cf., [15]). Many-valued logical semantics are interesting in practice since degrees of truth can be used to naturally express *vagueness* or *uncertainty* that often occurs in real-world data analysis tasks. Moreover, important benefits of rule-based reasoning, such as transparency and explainability, are preserved when moving to many-valued semantics. For space reasons, we can not recall the semantics of MV-Datalog$^\pm$ in detail here. Intuitively, rules are interpreted as in Łukasiewicz logic and we find the model such that no atom can be less true and every rule is satisfied to some given truth degree. Such *minimal K-fuzzy models* are in fact unique and can be computed in polynomial time in data complexity when there is exitential quantification in the head of rules. With existential quantification, current results [7] are restricted to cases where the oblivious chase is finite. However, in practice, utilization of termination strategies for chase sequences – e.g., as used for guarded or warded Datalog$^\pm$ [12] – can allow for good approximation of minimal K-fuzzy models in settings where the plain oblivious chase may not terminate.

In the application presented in this paper, we use MV-Datalog$^\pm$ to reason over the uncertain observations described in Section 3. Here we highlight some key ways in which natural constructs in MV-Datalog$^\pm$ implement important behaviour in our family link inference system. Note that the following examples have been simplified for the sake of presentation and that a full system can consider much more complex constructs.

Combining Multiple Sources of Uncertainty As discussed in the previous section, we start from some different vague observations on similarities between persons. Ultimately we want to combine these observations into a single judgement of how certain we are wrt the eventual family link between two people. By using MV-Datalog$^\pm$ we express how to combine these different sources of vague observations simply through the logical relationships of the observations. Note that in this case these connections are direct representations of expert knowledge but can also be derived from other sources, e.g., laws and regulations. For example, from our hypothesis, we believe that two people who own the same amount of shares in the same company (the *own* score interpreted as a fuzzy predicate from Section 3), and who live in the same region are related. Using MV-Datalog$^{\pm}$ we can express this directly through the simple rule for the vague observations *own* and *postal* as described above.

$$\mathit{related}(P_a, P_b) \leftarrow \mathit{own}(P_a, P_b),\ \mathit{postal}(P_a, P_b).$$

Importantly, if the region and ownership structure of two people is exactly the same, i.e., the respective *own* and *postal* observations are completely true (truth degree 1), then the relatedness of the respective persons will be inferred as completely true. If one or both of the observations are uncertain, the truth degree of $\mathit{related}(P_a, P_b)$ is derived through more complex means in a way that is consistent also with all the other rules.

In a similar fashion, we infer parent/child relationships for persons living at the same address where one person is much younger than the other. Again rule-based reasoning makes it simple to specify such a rule, but also to express that the child relationship is a special case of $\mathit{related}$:

$$\begin{array}{ll} \mathit{child}(P_a, P_b) & \leftarrow \mathit{age\_gap}(P_a, P_b),\ \mathit{addr}(P_a, P_b). \\ \mathit{related}(P_a, P_b) & \leftarrow \mathit{child}(P_a, P_b). \end{array}$$

This also illustrates how the vague observation *addr* can imply relatedness in a significantly different way than *own* or *postal*. The final solution model for the $\mathit{related}$ relation is then derived from all of these logical relationships and interactions globally through the program to provide a unified answer that is consistent with the input observations and the specified logical rules. Furthermore, it is possible to specify *weights* for rules to express their importance relative to one another and can further affect the solution of MV-Datalog$^{\pm}$ inference.

Fuzzy Closure over Logical Properties Family relations follow some logical axioms such as symmetry or transitivity. Naturally, if we detect that persons P_a and P_b are related, as well as P_a and P_c, we can deduce that also P_b and P_c are related. In rule-based reasoning, all such implied connections can be easily deduced by expressing the respective properties in logical terms. With uncertain information such situations become more complex, say we deduce $\mathit{related}(P_a, P_b)$ with truth 0.7 and, independently, $\mathit{related}(P_b, P_a)$ with truth 0.9, then it can be unclear how to consolidate this when assuming that $\mathit{related}$ is symmetric. Enforcing such properties via ad-hoc computation is often highly challenging especially when trying to reconcile multiple properties at the same time (such as symmetry *and* transitivity). MV-Datalog$^{\pm}$ offers natural semantics for such situations in which it will produce the equilibrium where every rule is as satisfied as possible. This allows us to express these important properties in simply their natural logical form as we would in any Datalog program for certain knowledge:

$$\begin{array}{ll} \mathit{related}(P_a, P_b) \leftarrow & \mathit{related}(P_b, P_a). \\ \mathit{related}(P_a, P_c) \leftarrow & \mathit{related}(P_a, P_b),\ \mathit{related}(P_b, P_c). \end{array}$$

Explainability and Transparency In a regulatory context, it is critical to know why a system infers family links. Many popular Machine Learning methods struggle to provide such explainability and transparency at the required level. By inferring family links through logical reasoning, our system is fully transparent and can provide full explanations, consisting of the individual steps of deduction that ultimately led to a link being inferred.

Consider a link between *A* and *B* that has been inferred with truth degree (i.e., level of certainty) 0.78 as a simple example with explanation (in formal terms):

$$\begin{array}{ll} \mathit{child}(A, C) & \leftarrow \mathit{age\_gap}(A, C),\ \mathit{addr}(A, C) \\ \mathit{related}(A, C) & \leftarrow \mathit{child}(A, C) \\ \mathit{related}(C, B) & \leftarrow \mathit{own}(C, B),\ \mathit{postal}(C, B) \\ \mathit{related}(A, B) & \leftarrow \mathit{related}(A, C),\ \mathit{related}(C, B) \end{array}$$

where the observation $\mathit{postal}(C, B)$ has truth degree 0.8, and $\mathit{addr}(A, C)$ is 0.98 true, and all other input observations are certain (truth degree 1). A natural reading of the explanation is that we deduce (with certainty 0.98) that *A* is a child of *C*, making them related with the same certainty. In this simple example, *C* and *B* have the exact same company ownership and live in very similar regions, hence we deduce they are related with truth degree 0.8. By transitivity of relatedness, this then also implies that *A* and *B* must be related and by the semantics of MV-Datalog$^{\pm}$ [2] the resulting fact $\mathit{related}(A, B)$ has truth 0.78. Note that the truth degree is generally not derived by simple forward propagation but requires more complex considerations since different paths of rule application can lead to the same consequences (say *A* and *B* also had some common ownership). For a formal definition of the semantics of MV-Datalog$^{\pm}$ we refer to Lanzinger et al. [7].

Advanced Data Cleaning MV-Datalog$^{\pm}$ also allows for *constraints*, that is rules with *falsum* $\perp$ in their head which express that the body of the rule shall not be satisfied. In a many-valued context, such constraints can be a valuable tool for data cleaning according to the logical properties of the domain. For example, if we are worried about data inaccuracies leading to two persons being inferred as children of each other, we can express that this is impossible through the following constraint:

$$\perp \leftarrow \mathit{child}(P_a, P_b), \mathit{child}(P_b, P_a)$$

[2] In this simple example this works out to the Łukasiewicz t-norm $\mathit{Truth}(\alpha \odot \beta) = \max\{0, \mathit{Truth}(\alpha) + \mathit{Truth}(\beta) - 1\}$ of the two body atoms α and β.

Constraints semantically behave like all other rules in our many-valued setting in that the solution model is the model that globally satisfies all rules as much as possible, which implies violating all constraints as little as possible. Again weighting also allows us to specify the relative importance of each constraint.

Interaction with Classical Knowledge Bases As a final point, since MV-Datalog$^{\pm}$ is an extension of plain Datalog, it is also well suited for tasks where we want to combine reasoning over vague inputs with classical knowledge bases. It would, for example, be natural to extend our system with various kinds of social graphs. Interaction with such classical (non-vague) knowledge bases requires no special interfaces since MV-Datalog$^{\pm}$ behaves exactly like plain Datalog on rules where the inputs are not vague (i.e., have truth 0 or 1).

5. Preliminary Experiments

Here we briefly report on early experiments with a proof-of-concept system built following the ideas described above in Sections 3 and 4. We compute fuzzy relations *own*, *addr*, and *postal* following the methods and clustering described in Section 3. Additionally, we consider a (not fuzzy) relation *age_gap* to specify whether an age gap between two persons is large enough for a parent/child relationship. Using these relations as input we use an early implementation of a MV-Datalog$^{\pm}$ reasoner, built on top of the state-of-the-art KGMS *Vadalog* [16] as well as the mathematical optimization solver Gurobi [17], to infer a fuzzy $related$ relation (including explanations for every inferred fact) from rules that implement the main ideas shown in Section 4, including transitive and symmetric closure and multi-layered combination of different fuzzy observations[3].

Our preliminary experiments followed two main goals. Understanding whether the reasoning in MV-Datalog$^{\pm}$ is feasible in acceptable time for our setting and how the inferred relationships match the intuition and intention of our rules.

With respect to the performance, we ran tests with a small and a medium-sized dataset, with each aforementioned relation containing 1000 or 10000 tuples, respectively. On a standard consumer laptop (8GB Memory, 1.4 GHz Quad-Core Intel Core i5), inference of the $related$ relation (as well as all corresponding explanations) takes roughly 10 seconds for the small instance and about 160 seconds for the medium-sized test dataset. Note that due to the effective partitioning of the search space through the combined use of WCCs and surname clusters as outlined in Section 3, these instance sizes are already representative of many real instances[4]. We can conclude that the computational effort required of our system is low enough for our setting and we expect this to still hold when further complexity is introduced (see Section 6). We also expect significant further improvements of these runtimes with further development of the MV-Datalog$^{\pm}$ reasoner, which itself is in an early stage of development.

With respect to the resulting $related$ relation. We infer almost 1000 tuples for the relation in the small instance (this includes the tuples resulting from transitive and symmetric closure), with varying degrees of truth. Similarly, we infer a $related$ relation containing about 9000 tuples for the medium-sized instance. Manual inspection of results and respective explanations confirm that the tuples and their inferred degree of truth match the intentions of domain experts that proposed the rules, confirming the effectiveness of our proposed approach.

6. Conclusion and Outlook

In this paper, we considered the family link detection problem in the setting of the Italian companies knowledge graph, described in Section 2. We present a novel proof-of-concept system based on uncertain predictions of family links, paired with advanced fuzzy reasoning. We use graph-theoretic techniques to reduce the search space significantly and then compute a number of fuzzy relations that may indicate family links between persons. We discuss how fuzzy reasoning in MV-Datalog$^{\pm}$ provides a powerful framework for reasoning on such relations as an alternative to popular Machine Learning methods, especially in settings with little to no training data. Early experiments suggest that this approach can produce effective systems and in particular so in regulatory applications where transparency and explainability are highly valuable.

The presented system design provides high flexibility in further development. Further data sources can be introduced naturally as further input relations with no additional effort. Adapting and adding rules is a straightforward task that can be performed at any time by domain experts.

One particular promising next step is the introduction of geocoding in addition to the computation of the score similarity $addr_{P_a,P_b}$. Geocoding overcomes certain limits of traditional methods based on string similarity. Indeed, it is able to reflect the real proximity of addresses according to their geographical coordinates. This addition can make our inference based on location much more robust, especially when we have two geographically close addresses with a total diverse nomenclature and in

[3] Integration of additional classical knowledge bases was not part of the tested program

[4] Reasoning needs to only be performed on single clusters in WCCs independently of other clusters and WCCs.

the case where addresses present severe transliteration errors.

Further planned extensions include the integration with additional knowledge graphs containing certain knowledge and in particular social graphs. The complex integration of various sources of certain knowledge with fuzzy observations will allow us to model family links in a more detailed way and produce more accurate inferred *related* relations.

Finally, the current work focuses on comparisons beyond people with the same surnames. Future work will aim to identify possible links with different surnames while retaining the computational benefits of the clustering as far as possible.

Acknowledgments

Stefano Sferrazza was supported by the Austrian Science Fund (FWF):P30930. Matthias Lanzinger acknowledges support by the Royal Society "RAISON DATA" project (Reference No. RP\R1\201074).

References

[1] J. B. Glattfelder, Ownership networks and corporate control: mapping economic power in a globalized world, Ph.D. thesis, ETH Zurich, 2010.

[2] A. Romei, S. Ruggieri, F. Turini, The layered structure of company share networks, in: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2015, pp. 1–10.

[3] A. Chapelle, A. Szafarz, Controlling firms through the majority voting rule, Physica A: Statistical Mechanics and its Applications 355 (2005) 509–529.

[4] F. Barca, M. Becht, The control of corporate europe. oxford university press, european corporate governance network (2001).

[5] P. Atzeni, L. Bellomarini, M. Iezzi, E. Sallinger, A. Vlad, Weaving enterprise knowledge graphs: The case of company ownership graphs, in: EDBT, OpenProceedings.org, 2020, pp. 555–566.

[6] GUIDELINE (EU) 2011/14 OF THE ECB, https://www.ecb.europa.eu/ecb/legal/pdf/l_33120111214en000100951.pdf, 2014. [Online; accessed 15-Feb-2019].

[7] M. Lanzinger, S. Sferrazza, G. Gottlob, Mv-datalog+-: Effective rule-based reasoning with uncertain observations, 2022. arXiv:2202.01718.

[8] A. Calì, G. Gottlob, T. Lukasiewicz, A. Pieris, Datalog+/-: A family of languages for ontology querying, in: O. de Moor, G. Gottlob, T. Furche, A. J. Sellers (Eds.), Datalog Reloaded - First International Workshop, Datalog 2010, Oxford, UK, March 16-19, 2010. Revised Selected Papers, volume 6702 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 351–368. URL: https://doi.org/10.1007/978-3-642-24206-9_20. doi:10.1007/978-3-642-24206-9_20.

[9] L. Bellomarini, D. Fakhoury, G. Gottlob, E. Sallinger, Knowledge graphs and enterprise AI: the promise of an enabling technology, in: ICDE, 2019.

[10] V. I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, volume 10, Soviet Union, 1966, pp. 707–710.

[11] A. Calì, G. Gottlob, T. Lukasiewicz, A general datalog-based framework for tractable query answering over ontologies, J. Web Semant. 14 (2012) 57–83. URL: https://doi.org/10.1016/j.websem.2012.03.001. doi:10.1016/j.websem.2012.03.001.

[12] G. Berger, G. Gottlob, A. Pieris, E. Sallinger, The space-efficient core of vadalog, in: D. Suciu, S. Skritek, C. Koch (Eds.), Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019, ACM, 2019, pp. 270–284. URL: https://doi.org/10.1145/3294052.3319688. doi:10.1145/3294052.3319688.

[13] A. Calì, G. Gottlob, M. Kifer, Taming the infinite chase: Query answering under expressive relational constraints, J. Artif. Intell. Res. 48 (2013) 115–174. URL: https://doi.org/10.1613/jair.3873. doi:10.1613/jair.3873.

[14] L. Bellomarini, G. Gottlob, A. Pieris, E. Sallinger, Swift logic for big data and knowledge graphs, in: C. Sierra (Ed.), Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, ijcai.org, 2017, pp. 2–10. URL: https://doi.org/10.24963/ijcai.2017/1. doi:10.24963/ijcai.2017/1.

[15] P. Hájek, Metamathematics of Fuzzy Logic, volume 4 of *Trends in Logic*, Kluwer, 1998. URL: https://doi.org/10.1007/978-94-011-5300-3. doi:10.1007/978-94-011-5300-3.

[16] L. Bellomarini, E. Sallinger, G. Gottlob, The Vadalog System: Datalog-based Reasoning for Knowledge Graphs, Proc. VLDB Endow. 11 (2018) 975–987. URL: http://www.vldb.org/pvldb/vol11/p975-bellomarini.pdf. doi:10.14778/3213880.3213888.

[17] Gurobi Optimization, LLC, Gurobi Optimizer Reference Manual, 2021. URL: https://www.gurobi.com.