Association for Information Systems

AIS Electronic Library (AISeL)

Wirtschaftsinformatik 2021 Proceedings | Track 13: Social Media and Digital Work

A No-Code Platform for Tie Prediction Analysis in Social Media Networks

Sebastian Schötteler
Nuremberg Institute of Technology, Faculty of Computer Science, Nuremberg, Germany;FAU Erlangen-Nuremberg, Schöller Endowed Chair for Information Systems (Digitalization in Business and Society), Nuremberg, Germany

Sven Laumer
FAU Erlangen-Nuremberg, Schöller Endowed Chair for Information Systems (Digitalization in Business and Society), Nuremberg, Germany

Heidi Schuhbauer
Nuremberg Institute of Technology, Faculty of Computer Science, Nuremberg, Germany

Niklas Scheidthauer
Nuremberg Institute of Technology, Faculty of Computer Science, Nuremberg, Germany

Philipp Seeberger
Nuremberg Institute of Technology, Faculty of Computer Science, Nuremberg, Germany

See next page for additional authors

Presenter Information

Sebastian Schötteler, Sven Laumer, Heidi Schuhbauer, Niklas Scheidthauer, Philipp Seeberger, and Benedikt Miethsam

A No-Code Platform for Tie Prediction Analysis in Social Media Networks

Sebastian Schötteler[1,2], Sven Laumer[2], Heidi Schuhbauer[1], Niklas Scheidthauer[1], Philipp Seeberger[1], and Benedikt Miethsam[1]

[1] Nuremberg Institute of Technology, Faculty of Computer Science, Nuremberg, Germany
{sebastian.schoetteler,heidi.schuhbauer,seebergerph64032, scheidthauerni63750, miethsambe65044}@th-nuernberg.de
[2] FAU Erlangen-Nuremberg, Schöller Endowed Chair for Information Systems (Digitalization in Business and Society), Nuremberg, Germany
{sven.laumer}@fau.de

Abstract. Conventional methods for tie prediction analysis in social media networks are often code-intensive and encompass complex steps. Against this backdrop, we used design science research to develop a no-code tie prediction analysis platform. Our evaluation indicates that the platform significantly reduces tie prediction analysis complexity and, depending on the network size, also total prediction time. Moreover, it maintains a prediction accuracy similar to that of conventional, code-intensive methods. Thus, our artifact substantially facilitates tie prediction analysis for social media network researchers and practitioners.

Keywords: Social Media Networks, Tie Formation Concepts, Tie Prediction Algorithms, Tie Prediction Analysis Platform, Social Media Analytics

1 Introduction

Humans, as social beings, naturally strive for social embeddedness [1]. Thus, social networks, such as friendship and communication networks, have been commonplace since the emergence of mankind [1]. What is new, however, is the proliferation of public and enterprise social media, such as Twitter and Yammer, and their novel potentials to achieve social embeddedness [2]. Nowadays, social networks based on social media (more briefly "social media networks") have become ubiquitous [3].

Social media networks are a subset of social networks and have their own distinct characteristics [2]. They confront researchers and managers with novel questions and challenges. In this context, a fundamental challenge is the social media tie prediction problem, which is concerned with inferring future or missing ties among actors – such as friends, employees, or students – from given social media network snapshots [4–6]. Addressing this problem provides major benefits, as would an eventual solution to it.

For instance, one application domain that benefits from research on social media tie prediction is the development of contact recommender systems for employees [6, 7]. Such systems can guide community emergence [8, 9] and prevent information overload

Internationale Konferenz Wirtschaftsinformatik 2021,
March 2021, Essen, Germany

[10]. Another domain is the detection of anomalies, such as crucial missing ties [7, 11], which can indicate conflicts that harm communication and knowledge transfer or collaboration [12, 13]. Further domains are the detection of influential communities [6, 7, 9] and the modeling of network evolution processes [6, 7, 14].

Driven by the tie prediction problem relevance induced by the various application domains, several studies have contributed to its solution [5–7, 9, 15, 16]. However, a review of these studies indicates that social media tie prediction research is complex and prone to errors, and it often relies on code-intensive methods [16–21]. This situation is particularly caused by the large number of prediction approaches [4, 5, 22] as well as the interdependencies [23] and large sizes [9] of network data samples.

Against this backdrop, our goal is to develop a no-code analysis platform – that is, a platform that enables analysis without programing – to facilitate common tie prediction analysis tasks in social media networks. This would support researchers (and perhaps practitioners) in the generation of new insights into the tie prediction problem. More specifically, our platform should facilitate three common tasks. First, it should simplify comparing the accuracy of various fundamental tie prediction approaches (Section 2) with reference to a given social media network [5–7, 9, 16, 21, 24]. Second, it should facilitate evaluating the accuracy of specific fundamental tie prediction approaches across various (types of) social media networks [15, 21, 24]. Lastly, it should simplify combining different tie prediction approaches into strong overarching predictors [5, 9, 15, 18, 19]. Thus, the following research question is addressed [25]:

How can we implement a no-code platform for tie prediction analysis in social media networks that facilitates the three aforementioned common tie prediction tasks?

To answer the research question, we followed the methodology by Peffers et al. [26], best practices from Gregor and Hevner [25], and guidelines from Hevner et al. [27]. First, we defined the problem, motivations for solving it, and objectives of a suitable solution. Then we designed and developed the platform, alias “artifact”, based on the problem statement and solution objectives. Finally, we demonstrated and evaluated the platform and communicated the findings.

In line with Gregor and Hevner [25], the paper is structured as follows. The next section provides a concise literature review on tie prediction analysis in social media networks. It moreover provides a literature review on tie formation concepts and tie prediction algorithms in social media networks and on further tie prediction artifacts relevant for the development of our platform. This is followed by a detailed description of the applied methodology, after which the developed no-code analysis platform is demonstrated. The next section describes our evaluation of the platform, followed by a discussion of the main findings. The last section concludes and summarizes the paper.

2 Related Work

In line with recommendations by Gregor and Hevner [25], we built our literature review on descriptive (Ω) and prescriptive (Λ) knowledge. Descriptive knowledge is the

"what" knowledge that sharpens the understanding of the problem, including its context and relevance, and how our proposed artifact may solve this problem [25]. Prescriptive knowledge is the "how" knowledge of relevant developed artifacts, such as concepts, algorithms, metrics, architectures, and system instantiations, which contribute to the development of our intended artifact [25]. Against this backdrop, this section contains four subsections. The first subsection contains a description of the relevant descriptive knowledge (Ω). The second subsection addresses fundamental tie formation concepts [23] in social media networks (Λ_1), meaning concepts that explain common drivers of tie formation in social media networks, which in turn can be used to predict ties. The third subsection deals with fundamental tie prediction algorithms [4] in social media networks (Λ_2), meaning algorithms that were explicitly developed for tie prediction purposes and which have been applied to social media networks. The fourth subsection describes further relevant artifacts for the development of our artifact (Λ_3).

To derive the aforementioned descriptive (Ω) and the prescriptive knowledge (Λ), we conducted a systematic literature search. More specifically, we used the search string "TITLE-ABS-KEY(("link" OR "tie") AND ("prediction" OR "formation") AND ("social media" OR "social network"))" to query the basket of journals and conferences named by [28] with the help of the Scopus database. We then studied the identified literature sources relevant to our topic, including their references and citations, to collect descriptive (Ω) and prescriptive knowledge (Λ) and further relevant sources. The findings derived from these literature sources constituted the knowledge base for the development of our platform (Section 4).

2.1 Tie Prediction Analysis in Social Media Networks (Ω)

Building on the definition provided by Wasserman and Faust [1], a social network, and thus also a social media network [2], consists of a finite set of actors (e.g., friends, employees, or students) and the ties (e.g., friendship, collaboration, or communication ties) defined on them. A fundamental challenge in such social media networks is the social media tie prediction problem, which is concerned with inferring future or missing ties among actors [4–6]. Several studies have contributed to its solution, leading to benefits for theory (e.g., novel insights into social media network evolution) and practice (e.g., techniques to enhance contact recommender systems) [5–7, 9, 15, 16]. However, social media tie prediction research is complex and prone to errors, and often relies on code-intensive methods, which may mitigate scientific progress [16–21].

To solve this problem, our goal is to develop a no-code platform with the objective to facilitate common tie prediction analysis tasks in social media networks by reducing complexity (e.g., coding complexity) while at the same time achieving a tie prediction velocity and accuracy comparable to those of conventional analysis methods [16–21]. More specifically, our platform should facilitate three common tasks derived from our literature review. First, the platform should simplify comparing the accuracy of various fundamental tie prediction approaches with reference to a given social media network [5–7, 9, 16, 21, 24]. Second, it should facilitate evaluating the accuracy of specific fundamental tie prediction approaches across various (types of) social media networks [15, 21, 24]. Third, it should simplify combining different tie prediction approaches

into strong overarching predictors [5, 9, 15, 18, 19]. This would support researchers (and perhaps practitioners) in the generation of new insights into the social media tie prediction problem. We could not identify any no-code platform for tie prediction analysis in social media networks in the existing literature, which facilitates these tasks.

2.2 Tie Formation Concepts in Social Media Networks (Λ_1)

Homophily: This concept implies that actors tend to form ties with other similar actors [29]. Homophily is often described with the idiom "birds of a feather flock together" [29]. Several types of homophilous tendencies can be observed in various types of social media networks [15, 30], indicating that this concept is relevant in such networks. For instance, past research has determined that tie formation in social media networks is driven by homophily in status [31], gender [32], function [30], topic [15], location [30], and hierarchy [32].

Dyadic Social Balance: This concept was first described by Heider [33] and explains (inter alia) tie reciprocation tendencies between two actors who are embedded in a dyad. Heider argued that unreciprocated ties in a dyad may lead to cognitive dissonance or social imbalance, which in turn may lead to uncertainty and instability [34]. Thus, actors embedded in such dyads naturally strive for reciprocity to establish cognitive consistence or social balance [33]. Research has shown that reciprocity drives tie formation in various types of social media networks [5, 30, 35, 36], suggesting that dyadic social balance is a relevant concept in social media networks.

Triadic Social Balance: The aforementioned social balance concept can be extended to the triadic level. Heider argues that, analogously to the dyadic level, actors embedded in triads tend to strive for cognitive consistence, which in turn influences triadic formation tendencies [33]. For instance, to avoid cognitive dissonance in a triad, a focal actor's alters (i.e., actors connected with him) may form a common tie, thus leading to a transitive triadic closure [37]. Colloquially, this social balance tendency is often expressed as "a friend of my friend should be my friend" [23]. Studies have determined that transitive triadic closure drives tie formation in various types of social media networks [8, 30, 38], indicating that triadic social balance is a relevant concept in social media networks. It seems to be slightly more relevant in public than in enterprise social media networks [14].

Rich-get-Richer: The rich-get-richer concept implies that the tendency of an actor to form a tie with a potential alter is proportional to the alter's tie "richness" (i.e., number of his pre-existing ties). Hence, rich actors become even more central over time [39]. This concept is also often referred to as cumulative advantage [39] or preferential attachment [14]. Ultimately, it may lead to scale-free networks that follow a power-law distribution [40]. Research has shown that tie formation in social media networks may often – yet not always [30] – be explained by this concept [14, 39, 41, 42].

In summary, tie formation in social media networks can be conceptualized using the homophily, dyadic social balance, triadic social balance, and the rich-get-richer concepts. As shown in past studies, the accuracy of tie prediction artifacts in social media networks can be enhanced when these formational concepts are considered [5, 8, 15]. Therefore, our intended artifact considers these concepts. The above list may not be exhaustive; for example, structural hole closing and transaction memory may also be relevant [23]. However, based on our systematic literature search, we are confident that the list reflects the most fundamental concepts [23] for developing our artifact.

2.3 Tie Prediction Algorithms for Social Media Networks (Λ_2)

Common Neighbors: This is a fundamental algorithm for predicting ties in various types of social networks [5, 9, 16, 22]. It is described in equation (1):

$$\textit{Common neighbors}\ (x,y) = |\Gamma(x) \cap \Gamma(y)| \tag{1}$$

The rationale behind this algorithm is that the more structurally similar two focal actors (x and y) are, as expressed by the number of their common neighbors (i.e., $|\Gamma(x) \cap \Gamma(y)|$, where $\Gamma(x)$ is the neighborhood of x, or in other words, his alters), the greater the odds of the actors establishing a common tie. Several social media network studies have applied the common neighbors algorithm for tie prediction purposes. Depending on the underlying social media network, the algorithm has achieved low [5, 9], medium [6, 7, 16, 21], or high [15, 21, 24] prediction accuracy.

Jaccard Coefficient: This is another fundamental algorithm for predicting ties in various social network types [5, 9, 16, 22, 43]. It can be defined as shown in (2).

$$\textit{Jaccard coefficient}\ (x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2}$$

It operates similarly to the common neighbors algorithm, but additionally divides the number of common neighbors across two focal actors by the number of actors that are neighbors of at least one of the focal actors (i.e., $|\Gamma(x) \cup \Gamma(y)|$). This enables addressing structural dissimilarities between the two focal actors more precisely, resulting in a more accurate representation of structural similarity. Several studies have applied this algorithm to predict ties in social media networks and have demonstrated low [9, 24], medium [5, 15, 16, 21], or high [6, 21, 24] prediction accuracy.

Adamic-Adar Index: This algorithm was originally used to predict ties between personal homepage networks [44]. Over time, it has evolved into a fundamental tie prediction algorithm that is applicable in various social network types [5, 9, 16, 22]. The algorithm is described in equation (3).

$$\textit{Adamic-Adar index}\ (x,y) = \sum\nolimits_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(|\Gamma(z)|)} \tag{3}$$

As in the common neighbors algorithm, the Adamic-Adar index algorithm counts the number of common neighbors between two focal actors. However, unlike the common neighbors algorithm, it assigns less connected common neighbors a greater weight (note that $\log(|\Gamma(z)|)$ resembles the logarithmic degree centrality of a common neighbor $z \in \Gamma(x) \cap \Gamma(y)$). The reasoning for the weighting approach is that two focal actors with ties to common neighbors having low connectivity are relatively likely to share unique characteristics, which in turn can increase the chance of a common tie emerging [44]. This algorithm has been used in various studies to predict ties in social media networks. Depending on the underlying social media network, the accuracy of prediction was low [5], medium [16, 21], or high [6, 21, 24].

Preferential Attachment: This is another fundamental tie prediction algorithm used in various social network types [5, 9, 16, 22]. It can be described as in (4).

$$\textit{Preferential attachment}\ (x,y) = |\Gamma(x)| \cdot |\Gamma(y)| \tag{4}$$

Building on the rich-get-richer concept (Section 2.2), research has shown that the probability of a tie occurring between two focal actors is correlated with the product of their degree centrality (i.e., $|\Gamma(x)| \cdot |\Gamma(y)|$) [4]. The preferential attachment algorithm has been used to predict ties in various types of social media networks, and has achieved low [6, 9, 16], medium [5, 21], or high [21] prediction accuracy.

The aforementioned list is not exhaustive. Nevertheless, after reviewing the various literature sources derived from our systematic literature search, we are confident that the list encompasses the most fundamental tie prediction algorithms [22] that should be considered in our artifact.

2.4 Further Relevant Tie Prediction Artifacts (Λ_3)

The literature review indicated that researchers often combine different tie prediction concepts and algorithms with the help of classification machine learning algorithms, such as the decision tree [5, 9, 15, 18, 19], random forest [18, 19], k-nearest-neighbors [18, 19], and logistic regression algorithm [5], to achieve a higher prediction accuracy. Moreover, studies often rely on the AUC (area under curve) metric to evaluate their achieved prediction accuracy, expressed by a value between 0 (lowest accuracy) and 1 (highest accuracy) [9, 11, 16, 22]. These artifacts were also considered for our platform.

Lastly, we could not identify any no-code platform for tie prediction analysis in social media networks in the existing literature, which addressed the aforementioned descriptive (Ω) and prescriptive knowledge (Λ). This lack was underpinned by the fact that, as far as evident, the aforementioned social media tie prediction studies derived from our systematic literature search have all relied on more complex methods (e.g., code-intensive methods [16–21]). The most similar derived artifact is the tie prediction architecture by Schall [24]. However, it lacks relevant complexity reduction features, such as an analysis preparation component, which allows translating the complete

architecture into an integrated no-code platform. Nonetheless, it supplied a sound basis for our artifact and was therefore incorporated.

3 Methodology

The goal of this research was to develop a no-code platform to facilitate tie prediction analysis in social media networks (Section 1). This objective is congruent with design science research, which aims to create and evaluate artifacts, such as constructs, models, methods, and – as in this paper – system instantiations, to solve relevant problems [27]. Thus, we adopted the design science research guidelines, best practices, and methodology proposed by Hevner et al. [27], Gregor and Hevner [25], and Peffers et al. [26], respectively. Figure 1 illustrates the applied design science research approach.

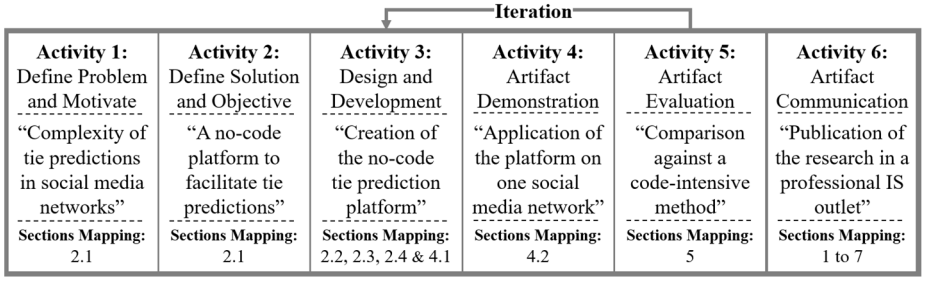

Figure 1. Design science research approach (adapted from [26])

In the first activity, we collected descriptive knowledge (Ω) [25] to specify the research problem, namely the high complexity of conventional tie prediction analysis methods in social media networks; we also used this knowledge to motivate why solving this problem would be valuable (Section 2.1) [26]. Next, we further used this knowledge [25] to infer a solution to the problem and the main objective of that solution. The solution was a platform with the objective to facilitate tie prediction analysis via a no-code paradigm (Section 2.1) [26]. In the third activity, we collected prescriptive knowledge (Λ) [25] (Sections 2.2 to 2.4) and used it to translate the solution into an artifact, namely the proposed platform. This activity involved defining the artifact's requirements and architecture as well as actually creating the artifact (Section 4.1) [26]. Next, we demonstrated how the artifact could be used to solve the specified research problem. We did so by conducting an experimental [27] tie prediction analysis in an empirically derived social media network (Section 4.2) [26]. In the fifth activity, we evaluated the artifact's complexity, velocity, and accuracy (as three common performance metrics [27] also relevant for our artifact (Section 2.1)) compared with a conventional, code-intensive method [17]. We used two social media networks for this task and repeatedly returned to activity 3 until the evaluation results were satisfactory (Section 5) [26], thus incorporating the iterative build-and-evaluate paradigm of design science research [27]. The last activity was communicating the research findings to

diffuse the created body of knowledge and the artifact. The communication process explains the specified problem and its relevance, and indicates how – and how well – the artifact solves the problem [26]. This paper represents the fulfilment of this step. Moreover, we published the created artifact into a public software repository to make it publicly available for use and refinement[1].

Following the logic of Merwe et al. [45], we mapped each activity to the corresponding sections proposed in the design science research publication schema by Gregor and Hevner [25]. This paper's structure has thus been defined, and more detailed insights regarding each activity appears in the relevant sections (Figure 1).

4 Artifact

This section has two subsections. The first describes the development procedure of the artifact, starting from a requirements definition and progressing to the architectural conception and ending with the actual artifact creation [26]. The second subsection demonstrates the developed artifact [26], namely our no-code platform to facilitate social media tie prediction analysis by reducing complexity (e.g., coding complexity).

4.1 Development

Requirements: Based on the research question (Section 1) and the review of the relevant descriptive (Ω) and prescriptive (Λ) knowledge (Section 2), we defined ten requirements for our artifact. These are as follows: First, the artifact should enable the use of (1) the derived tie formation concepts (Section 2.2) and (2) prediction algorithms (Section 2.3) as tie prediction approaches. Moreover, it should enable (3) comparison of various approaches with reference to a given social media network and (4) evaluation of a specific approach across several social media networks (Section 2.1). In addition, the artifact should enable (5) combining different approaches into a strong predictor (Section 2.1) via the determined (6) machine learning algorithms (decision tree, random forest, k-nearest-neighbors, and logistic regression algorithm) (Section 2.4). Lastly, the artifact should meet all the above requirements while also (7) following a no-code paradigm, thus rendering programing obsolete (Section 2.1). This no-code paradigm should (8) substantially reduce analysis complexity compared to conventional tie prediction analyses, while achieving a tie prediction (9) velocity and (10) accuracy comparable to those of conventional analyses (Section 2.1).

Architecture: To conceptualize the architecture of our artifact, we relied on the literature review findings and adapted the modular tie prediction architecture proposed by Schall [24] (Section 2.4). Our artifact architecture is presented in Figure 2.

[1] Public repository link: https://github.com/social-media-analytics-research/tie-prediction-tool

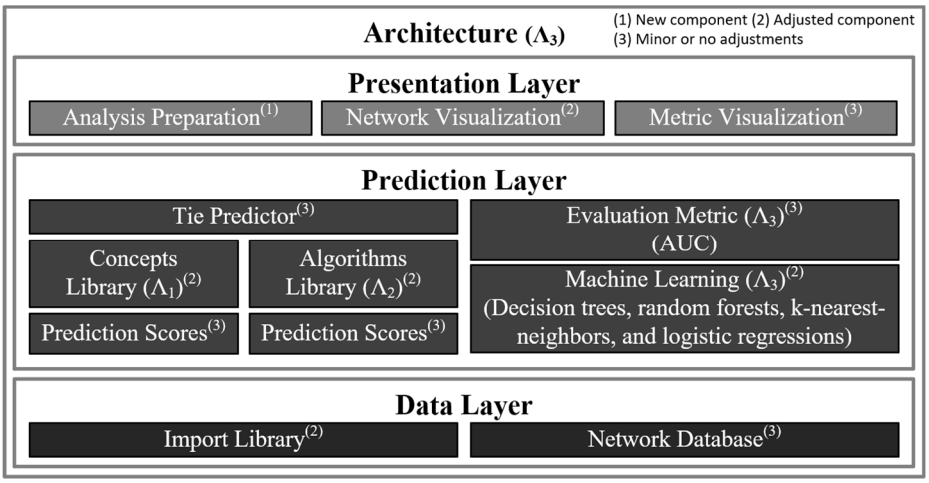

Figure 2. Artifact architecture (adapted from [24])

The data layer contains an import library that enables migrating social media networks into the network database, which is also situated in this layer. The stored networks can then be loaded from the tie predictor in the prediction layer.

The tie predictor is the main component of the prediction layer. In the first step, the tie predictor loads a network N selected from the database and extracts from it a random tie subset of specified size, while maintaining the same actor set as in N, resulting in a test network N_{TE}. This process is then repeated using the network N_{TE} to derive a training network N_{TR} [24], such that $N_{TR} \subseteq N_{TE} \subseteq N$ [6, 16, 46, 47]. Subsequently, the tie predictor uses one of the prediction approaches stored in the concepts (Section 2.2) or algorithms library (Section 2.3) to calculate the associated prediction scores for each pair of network actors (x,y) in N_{TR} and N_{TE}. This step can be augmented by combining different tie prediction approaches into a strong predictor via the implemented machine learning algorithms. To meet the defined requirements, we decided to implement the machine learning algorithms derived from the literature review, namely decision tree, random forest, k-nearest-neighbors, and logistic regression algorithms (Section 2.4). After calculating these prediction scores, the tie predictor uses the scores derived for the network N_{TR} together with N_{TR} itself to train its prediction accuracy. Lastly, the tie predictor evaluates its accuracy via the prediction scores derived for the network N_{TE}, N_{TE} itself, and the implemented evaluation metric. For this, we relied on the AUC metric determined in the literature review, which quantifies the achieved prediction accuracy as a value between 0 (lowest accuracy) and 1 (highest accuracy) (Section 2.4).

The presentation layer contains an analysis preparation component. It is used to prepare the tie prediction analysis by selecting the network N from the database, the sample sizes of the networks N_{TR} and N_{TE}, the tie prediction approach(es), and further tie prediction parameters, if available (Figure 3). This component is further used to start the core prediction process. After that process is complete, the network visualization component can be used to derive the (correctly and incorrectly) predicted ties in the network N_{TE} (Figure 4). Moreover, the metric visualization component can be used to

determine the AUC metric values for the networks N_{TR} and N_{TE}, including an associated AUC graph that visualizes the achieved accuracy (Figure 4).

We made various adaptations to the baseline architecture to reduce complexity and to implement our findings from the literature review. First and potentially most importantly, we added an analysis preparation component to unify all necessary preparation steps into an integrated no-code dashboard. Next, we adjusted the network visualization component to enable displaying the predicted ties. Furthermore, we adjusted the concept, algorithm, and machine learning components as per our literature review findings. Lastly, we extended the import library so that standard network data formats such as GEFX, GML, GraphML, and CSV [20] could be imported easily. The residual components underwent only minor or no adjustments.

Creation: In the data layer, we used PostgreSQL for the network database [48] and Python [49] together with NetworkX [20] to develop the import library. In the prediction layer, we used Python and NetworkX [20] for the tie predictor, the concepts, and algorithms library, and for each prediction score and the AUC evaluation metric component. Next, we used the Scikit-learn [50] Python package for the machine learning component. We opted for Scikit-learn because it is a well-established package both in academia and practice [50] and covers the common machine learning algorithms used for social media tie prediction identified by our literature review (Section 2.4). Moreover, it is simpler to configure compared to alternatives, such as PyTorch or TensorFlow [50], thus facilitating a further development of our platform by other users, if required. Although the two aforementioned packages contain more advanced machine learning techniques, future implementations of our platform can efficiently overcome this limitation with the help of specialized Scikit-learn wrappers, such as Skorch (a wrapper that augments Scikit-learn by PyTorch machine learning techniques) [51]. In the presentation layer, we used D3.js [52] and AngularJS [53] for the analysis preparation and each visualization component.

4.2 Demonstration

For the artifact demonstration, we empirically collected an enterprise social media network, N^D, from a geographically dispersed organization. The network contained 17 employees, and displayed 160 present and 112 absent directed collaboration ties [9]. Moreover, N^D contained employee attributes such as business function and location, which enabled homophily-based tie prediction analysis. We used the common neighbors algorithm and the homophily concept as prediction approaches. The artifact enables assigning global relevance weights to the various homophily attributes of the embedded employees. As N^D was geographically dispersed, we decided to focus solely on locational homophily. The analysis preparation is illustrated in Figure 3. The predicted ties and achieved accuracies (i.e., AUC values) are presented in Figure 4.

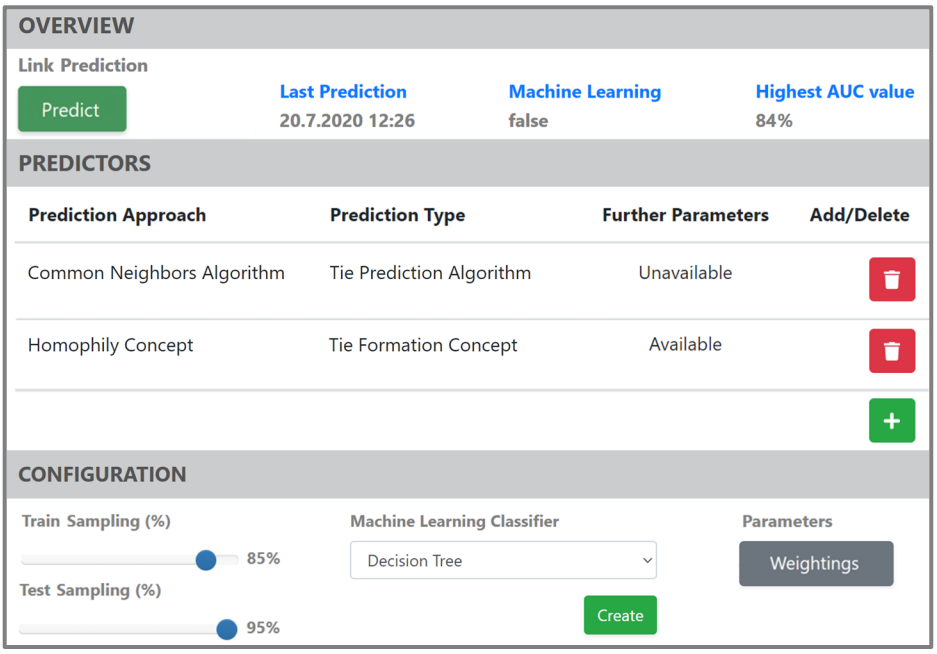

Figure 3. Analysis preparation dashboard: aggregated representation

We used the preparation dashboard to configure our prediction approach. First, we imported, named, and described our network sample N^D. Next, we adjusted the prediction settings by selecting the prediction approaches, homophily weightings (via the button "Weightings"), and a network split of N^D_{TR} = 85% and N^D_{TE} = 95%. Then, we started the actual prediction process (via the button "Predict"). Figure 3 illustrates the analysis preparation dashboard after completion of the tie prediction process.

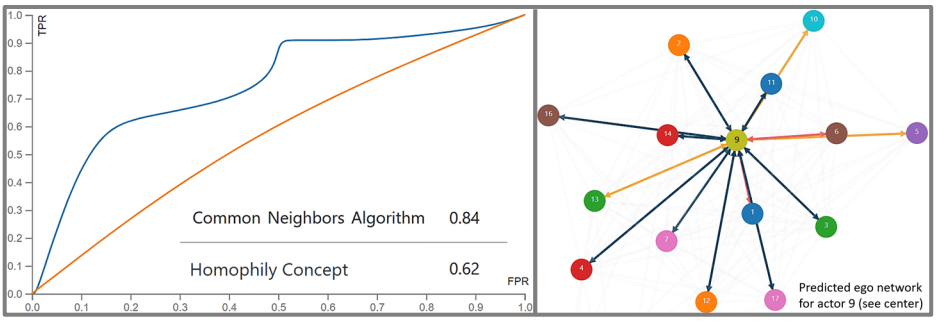

Figure 4. Metric (left) and network visualization dashboard (right): aggregated representation

The metric visualization dashboard (left panel) presents the achieved N^D_{TE} AUC values for the chosen prediction approaches; these were 0.84 for the common neighbors algorithm and 0.62 for the locational homophily concept. Both approaches achieved a

higher-than-random (i.e., $> 0{,}5$) accuracy. However, as reflected by the higher AUC value, the first approach achieved more accurate predictions. The dashboard augmented the values with an AUC graph. Lastly, the network visualization dashboard (right panel) allowed for selecting an actor's (e.g., actor 9's) egocentric network to examine his correctly (orange) and incorrectly (red) predicted ties. Blue ties were already present in N_{TR}^{D}. Both prediction approaches considered N^{D} to be undirected.

5 Evaluation

To evaluate our artifact ("ART"), we compared its performance against an alternative tie prediction method ("ALT"). For this, we built upon the alternative method from Bojanowski and Chroł [17]. For this, we used RStudio [54], a statistical programming tool often used for tie prediction analysis [17], augmented by preprocessing, tie prediction, and machine learning features. The methods were compared by assessing four performance metrics, collected along two prediction approaches and networks. The result was four evaluation cases with four value comparisons in each case (Table 1).

Table 1: Evaluation results

Approach 1: Adamic-Adar index algorithm					$N_{TR}^{E(n)} = 80\% \mid N_{TE}^{E(n)} = 95\%$			
Approach 2: Logistic regression machine learning combining the Jaccard coefficient and preferential attachment algorithm								
Network	*UCI network* (N^{E1})				*Twitter network* (N^{E2})			
Approach	Approach 1		Approach 2		Approach 1		Approach 2	
Method	ART	ALT	ART	ALT	ART	ALT	ART	ALT
Complexity (min:sec)	03:10	12:45	03:31	07:30	00:27	14:25	00:44	09:35
Velocity (min:sec)	03:46	00:11	05:56	00:27	00:13	00:06	00:24	00:10
Total time (min:sec)	06:56	12:56	09:27	07:57	00:40	14:31	01:08	09:45
Accuracy (AUC)	0.80	0.77	0.91	0.91	0.84	0.83	0.81	0.80

As mentioned earlier, we compared the methods by measuring and juxtaposing their achieved prediction performances in two different prediction approaches. For the first approach ("Approach 1"), we used the Adamic-Adar index algorithm. For the second approach ("Approach 2"), we used logistic regression machine learning to combine the Jaccard coefficient and the preferential attachment algorithm into a strong predictor.

Each approach was applied to two networks. The first network ("UCI network (N^{E1})") comprised 1899 students from a Facebook-like internal social media platform at the University of California, Irvine (UCI). It encompassed 20,296 present and 3,584,006 absent directed communication ties [55]. The second network ("Twitter network (N^{E2})") was derived from Twitter via Gephi [56] and encompassed 745 Twitter users as well as 1204 present and 553,076 absent directed communication ties.

We defined four metrics to quantify the achieved performance of each method in each approach-network combination or evaluation case. Complexity represented the effort, in minutes, to prepare the analysis. Velocity displayed the minutes of the actual

prediction calculation process. Total time indicated the cumulated complexity and velocity time. Accuracy represented the achieved AUC value. Each performance metric was recorded for each method in each evaluation case by a researcher with basic experience in both methods.

Better performance metric values in each evaluation case are underlined twice. The artifact achieved better (lower) complexity scores than the alternative method. However, the alternative method achieved better (lower) velocity scores, particularly in the second approach ("Approach 2"). Nevertheless, the artifact achieved better (lower) total prediction time scores in three of the four evaluation cases. Lastly, the methods achieved similar accuracy scores, indicating that the prediction approaches have been correctly implemented into the artifact; small variances may be attributed to the randomness factor in the tie extraction step (Section 4.1). These results are discussed in Section 6.

6 Discussion

The goal was to reduce tie prediction complexity while maintaining a velocity and accuracy similar to those attained by conventional methods. For this purpose, we developed a no-code tie prediction platform. Our evaluation results indicate that our artifact significantly reduces tie prediction complexity while maintaining similar accuracy. However, the alternative method (the chosen conventional method) achieved better velocity scores, particularly in the machine learning approach. However, for the chosen social media networks, the artifact's lower total prediction time largely overcompensated for its slower velocity. However, it is to be expected that this compensation effect diminishes proportionally to the network size. Nevertheless, various studies (e.g., [21, 57]) have used sizes for which our artifact may perform similarly, or better than, the alternative method regarding the total time. Thus, while future studies should enhance our artifact's prediction velocity, it is already suitable for tie prediction analysis – even for larger networks, if velocity is not the main priority.

Due to the iterative nature of the design science research methodology, the evaluation step was repeated several times. Here, we determined two further differences between the methods. First, the artifact displayed a substantially lower learning curve. While the researcher required several hours to perform a first tie prediction analysis in the alternative, he only required a few minutes in the artifact. Second, the artifact's no-code paradigm guarded the researcher against common pitfalls and errors that occurred in the alternative, such as mistakes in the network splitting process. Thus, the artifact promotes valid and correct prediction procedures. On the other hand, the no-code paradigm may sometimes restrict researchers. For instance, although the artifact allows random undersampling, augmenting it through additional sampling techniques could improve its accuracy [58]. Another useful addition would be to enable time-variant tie prediction analysis [6]. In general, future studies could augment the artifact by various additional features, thus aligning its flexibility with that of conventional methods.

Congruent with our evaluation results, we conclude that our developed artifact answers the research question proposed in Section 1. Furthermore, the artifact meets

the requirements in Section 4.1, apart from (9), which is only partially met due to the velocity concerns. Thus, with reference to the tie prediction problem stated in the introduction, our artifact reduces tie prediction complexity and facilitates the generation of novel insights. For instance, researchers could use the artifact to more easily explore formation mechanisms through which communities in social media networks evolve [6, 7, 9, 14], and could use the findings in order to develop guidelines to foster fruitful community evolution. Moreover, practitioners could use the artifact to more easily infer tie formation antecedents in order to develop their own recommender systems [7, 46].

7 Conclusion

Readers should consider the artifact's limitations, namely the improvable velocity score and the room for further tie prediction features. As an outlook, two optimization measures could mitigate these limitations. Firstly, to substantially enhance platform velocity, realizing artifact compatibility with alternative Python compilers, such as the PyPy just-in-time compiler, seems promising [59]. Secondly, to enable even more sophisticated tie prediction analyses, exploiting the features of the Scikit-learn package more strongly or even implementing a wrapper that augments this package, such as Skorch (Section 4.1), should prove valuable [51]. Another limitation is that the artifact was created upon literature mainly derived from the social media network domain. Although probable, it is not clear whether the artifact is directly applicable to other social network types. Lastly, future studies may strengthen the platform evaluation using additional prediction approaches, social media networks, metrics, or users [27].

Despite its limitations, our artifact extends the current body of knowledge by three novel contributions. First, conventional methods for tie prediction analysis in social media networks often rely on code-intensive approaches that encompass a set of complex steps [16–21]. In this context, our artifact provides the advantage of a no-code paradigm to reduce this complexity and to facilitate tie prediction analyses while maintaining an accuracy and, depending on the network sample size, total prediction time similar to that of conventional methods. Second, another advantage of the no-code paradigm is that it reduces the risk of mistakes and leads to a low learning curve, thus allowing to achieve valid results rapidly. Lastly, the artifact contains tailored features, such as an overview of the predicted ties and homophily type-based prediction analysis calibrations, enabling uncomplex but fine-grained analyses.

From a theoretical perspective, we contribute to research by augmenting Schall's tie prediction architecture [24] with an analysis preparation component. Our evaluation results imply that this architectural extension has various positive outcomes for resultant system instantiations (e.g., reduction of complexity, lower learning curves, and less user mistakes).

References

1. Wasserman, S., Faust, K.: Social Network Analysis – Methods and Applications. Cambridge University Press, Cambridge, USA (1994).

2. Labianca, G., Kane, G., Alavi, M., Borgatti, S.: What's Different about Social Media Networks? A Framework and Research Agenda. MISQ. 38, 274–304 (2013).
3. Meske, C., Junglas, I., Schneider, J., Jaakonmaeki, R.: How Social is Your Social Network? Toward A Measurement Model. In: Proceedings of the Fortieth International Conference on Information Systems (ICIS '19). pp. 1–9. Association for Information Systems, Munich, Germany (2019).
4. Liben-Nowell, D., Kleinberg, J.: The Link Prediction Problem for Social Networks. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM '03). pp. 556–559. Association for Computing Machinery, New York, United States (2004).
5. Cheng, J., Romero, D., Meeder, B., Kleinberg, J.: Predicting Reciprocity in Social Networks. In: Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT '11) and 2011 IEEE Third International Conference on Social Computing (SocialCom '11). pp. 49–56. IEEE Computer Society, Boston, USA (2011).
6. Yin, D., Hong, L., Davison, B.: Structural Link Analysis and Prediction in Microblogs. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11). pp. 1–6. Association for Computing Machinery, Glasgow, United Kingdom (2011).
7. Tsugawa, S., Kito, K.: Retweets as a Predictor of Relationships among Users on Social Media. PLoS ONE. 12, 1–19 (2017).
8. Brzozowski, M., Romero, D.: Who Should I Follow? Recommending People in Directed Social Networks. In: Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM '11). pp. 458–461. Association for the Advancement of Artificial Intelligence, Barcelona, Spain (2011).
9. Valverde-Rebaza, J., de Andrade Lopes, A.: Exploiting Behaviors of Communities of Twitter Users for Link Prediction. Soc. Netw. Anal. Min. 3, 1063–1074 (2013).
10. Chen, X., Wei, S.: Enterprise Social Media Use and Overload: A Curvilinear Relationship. Journal of Information Technology. 34, 22–38 (2019).
11. Luo, P., Li, Y., Wu, C., Chen, K.: Detecting the Missing Links in Social Networks Based on Utility Analysis. J. Comput. Sci. 16, 51–58 (2016).
12. Oostervink, N., Agterberg, M., Huysman, M.: Knowledge Sharing on Enterprise Social Media: Practices to Cope with Institutional Complexity: Knowledge Sharing on Enterprise Social Media. J. Comput-Mediat. Comm. 21, 156–176 (2016).
13. Azaizah, N., Reychav, I., Raban, D., Simon, T., McHaney, R.: Impact of ESN Implementation on Communication and Knowledge-Sharing in a Multi-National Organization. International Journal of Information Management. 43, 284–294 (2018).
14. Wiesneth, K.: Evolution, Structure and Users' Attachment Behavior in Enterprise Social Networks. In: Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS '16). pp. 2038–2047. IEEE Computer Society, Koloa, USA (2016).
15. Aiello, L., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., Menczer, F.: Friendship Prediction and Homophily in Social Media. ACM Trans. Web. 6, 1–33 (2012).
16. Martinčić-Ipšić, S., Močibob, E., Perc, M.: Link Prediction on Twitter. PLOS ONE. 12, 1–21 (2017).
17. Bojanowski, M., Chroł, B.: Proximity-Based Methods for Link Prediction in Graphs with R Package `linkprediction'. Kozminski University, Warsaw, Poland (2019).
18. Fire, M., Katz, G., Rokach, L., Elovici, Y.: Links Reconstruction Attack. In: Altshuler, Y., Elovici, Y., Cremers, A., Aharony, N., and Pentland, A. (eds.) Security and Privacy in Social Networks. pp. 181–196. Springer, New York, USA (2013).

19. Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., Elovici, Y.: Link Prediction in Social Networks Using Computationally Efficient Topological Features. In: Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT '11) and 2011 IEEE Third International Conference on Social Computing (SocialCom '11). pp. 73–80. IEEE Computer Society, Boston, USA (2011).
20. Hagberg, A., Schult, D., Swart, P.: NetworkX: Network Analysis in Python, https://networkx.github.io/documentation/latest/, last accessed 2020/07/12.
21. Gao, F., Musial, K., Cooper, C., Tsoka, S.: Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics. Sci. Program. 1–13 (2015).
22. Zhou, T., Lü, L., Zhang, Y.-C.: Predicting Missing Links Via Local Information. Eur. Phys. J. B. 71, 623–630 (2009).
23. Contractor, N., Wasserman, S., Faust, K.: Testing Multi-theoretical Multilevel Hypotheses About Organizational Networks: An Analytic Framework and Empirical Example. Acad. Manag. Rev. 31, 681–703 (2006).
24. Schall, D.: Link Prediction in Directed Social Networks. Soc. Netw. Anal. Min. 4, 157 (2014).
25. Gregor, S., Hevner, A.: Positioning and Presenting Design Science Research for Maximum Impact. MISQ. 37, 337–356 (2013).
26. Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. JMIS. 45–77 (2007).
27. Hevner, A., March, S., Park, J., Ram, S.: Design Science in Information Systems Research. MISQ. 28, 75–105 (2004).
28. Wang, P., Xu, B., Wu, Y., Zhou, X.: Link Prediction in Social Networks: The State-of-the-Art. Sci. China Inf. Sci. 58, 1–38 (2015).
29. McPherson, M., Smith-Lovin, L., Cook, J.: Birds of a Feather: Homophily in Social Networks. Annu. Rev. Sociol. 27, 415–444 (2001).
30. Kim, Y., Kane, G.: Online Tie Formation in Enterprise Social Media. APJIS. 29, 382–406 (2019).
31. Šćepanović, S., Mishkovski, I., Gonçalves, B., Nguyen, T., Hui, P.: Semantic Homophily in Online Communication: Evidence from Twitter. OSNEM. 2, 1–18 (2017).
32. Di Tommaso, G., Gatti, M., Iannotta, M., Mehra, A., Stilo, G., Velardi, P.: Gender, Rank, and Social Networks on an Enterprise Social Media Platform. Soc. Networks. 62, 58–67 (2020).
33. Heider, F.: Attitudes and Cognitive Organization. J. Psychol. 21, 107–112 (1946).
34. Festinger, L., Hutte, H.A.: An Experimental Investigation of the Effect of Unstable Interpersonal Relations in a Group. J. Abnorm. Psychol. 49, 513–522 (1954).
35. Rode, H.: Analyzing Motivational Determinants of Knowledge-sharing in Enterprise Social Media Platforms. Acad. Manag. Proc. 2015, 6 (2015).
36. Quercia, D., Capra, L., Crowcroft, J.: The Social World of Twitter: Topics, Geography, and Emotions. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM '12). pp. 298–305. Association for the Advancement of Artificial Intelligence, Dublin, Ireland (2012).
37. Granovetter, M.: The Strength of Weak Ties. Am. J. Sociol. 78, 1360–1380 (1973).
38. Sadri, A.M., Hasan, S., Ukkusuri, S., Lopez, J.: Analysis of Social Interaction Network Properties and Growth on Twitter. Soc. Netw. Anal. Min. 8, 2–13 (2018).
39. Su, J., Sharma, A., Goel, S.: The Effect of Recommendations on Network Structure. In: Proceedings of the 25th International Conference on World Wide Web (WWW '16). pp. 1157–1167. Association for Computing Machinery, Montréal, Canada (2016).

40. Johnson, S., Faraj, S., Kudaravalli, S.: Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment. MISQ. 38, 795–808 (2014).
41. Kumar, A., Kushwah, S., Manjhvar, A.: A Review on Link Prediction in Social Network. Int. J. Grid Distrib. Comput. 43–50 (2016).
42. Overbey, L., Greco, B., Paribello, C., Jackson, T.: Structure and Prominence in Twitter Networks Centered on Contentious Politics. Soc. Netw. Anal. Min. 3, 1351–1378 (2013).
43. Jaccard, P.: Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. Bull. Soc. Vaud. Sci. Nat. 37, 241–272 (1901).
44. Adamic, L., Adar, E.: Friends and Neighbors on the Web. Soc. Networks. 25, 211–230 (2003).
45. Van der Merwe, A., Gerber, A., Smuts, H.: Mapping a Design Science Research Cycle to the Postgraduate Research Report. In: Liebenberg, J. and Gruner, S. (eds.) ICT Education. pp. 293–308. Springer International Publishing, Cham, Germany (2017).
46. Yin, Z., Gupta, M., Weninger, T., Han, J.: A Unified Framework for Link Recommendation Using Random Walks. In: Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM '10). pp. 152–159. IEEE Computer Society, Odense, Denmark (2010).
47. Clauset, A., Moore, C., Newman, M.: Hierarchical Structure and the Prediction of Missing Links in Networks. Nature. 453, 98–101 (2008).
48. Ahmed, I., Fayyaz, A., Shahzad, A.: PostgreSQL Developer's Guide. Packt Publishing Ltd, Birmingham, United Kingdom (2015).
49. Langtangen, H.: Python Scripting for Computational Science. Springer Science & Business Media, Heidelberg, Germany (2009).
50. Hackeling, G.: Mastering Machine Learning with scikit-learn. Packt Publishing Ltd, Birmingham, United Kingdom (2017).
51. Tietz, M., Nouri, D., Bossan, B.: Skorch 0.9.0 Documentation: A scikit-learn Compatible Neural Network Library That Wraps PyTorch, https://skorch.readthedocs.io/en/stable/, last accessed 2020/11/07.
52. Bostock, M., Ogievetsky, V., Heer, J.: D3 Data-Driven Documents. IEEE Trans. Visual. Comput. Graphics. 17, 2301–2309 (2011).
53. Körner, C.: Data Visualization with D3 and AngularJS. Packt Publishing Ltd, Birmingham, United Kingdom (2015).
54. Gandrud, C.: Reproducible Research with R and R Studio. CRC Press, Boca Raton, USA (2018).
55. Panzarasa, P., Opsahl, T., Carley, K.: Patterns and Dynamics of Users' Behavior and Interaction: Network Analysis of an Online Community. J. Am. Soc. Inf. Sci. 911–932 (2009).
56. Hammer, L.: Guide: Analyzing Twitter Networks with Gephi 0.9.1, https://lucahammer.com/2016/09/06/guide-analyzing-twitter-networks-with-gephi-0-9-1/, last accessed 2020/07/18.
57. Divakaran, A., Mohan, A.: Temporal Link Prediction: A Survey. New Gener. Comput. 38, 213–258 (2020).
58. Blagus, R., Lusa, L.: Smote for High-Dimensional Class-Imbalanced Data. BMC Bioinformatics. 14, 106 (2013).
59. Bolz, C., Cuni, A., Fijalkowski, M., Rigo, A.: Tracing the Meta-Level: Pypy's Tracing Jit Compiler. In: Proceedings of the 4th Workshop on the Implementation, Compilation, Optimization of Object-Oriented Languages and Programming Systems (ICOOOLPS '09). pp. 18–25. Association for Computing Machinery, New York, USA (2009).