

---

# The Dynamics of Innovation

Non-linear Modelling of Patent Citations

Doctoral Dissertation submitted to the  
Faculty of Informatics of the Università della Svizzera italiana  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy

presented by  
Edoardo Filippi-Mazzola

under the supervision of  
Ernst C. Wit

November 2024

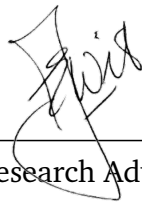


---

Dissertation Committee

**Fabio Crestani**      Università della Svizzera italiana, Lugano, Switzerland  
**Igor Pivkin**        Università della Svizzera italiana, Lugano, Switzerland  
**Viviana Amati**     Università degli Studi di Milano Bicocca, Milan, Italy  
**Christoph Stadtfeld** Eidgenössische Technische Hochschule, Zurich, Switzerland

Dissertation accepted on 14 November 2024



---

Research Advisor

**Ernst C. Wit**

---

PhD Program Director

**Walter Binder, Stefan Wolf**

---

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.



---

Edoardo Filippi-Mazzola  
Lugano, 14 November 2024

# Abstract

This dissertation explores the development and application of advanced modeling techniques to analyze network dynamics, with a specific focus on patent citation networks. This study primarily centers on Relational Event Models (REMs), which have proven effective in modeling sequential interactions within networks but face significant limitations when applied to large datasets and complex non-linear relationships. To address these challenges, this work develops extensions of the REM, including the Stochastic Gradient Relational Event Additive Model (STREAM) and the Deep Relational Event Additive Model (DREAM), both of which incorporate non-linear modeling techniques and deep learning approaches. The research presented here offers several key contributions. First, a method for computing textual similarity scores using embeddings from patent abstracts is proposed, demonstrating efficiency and effectiveness in capturing complex relationships within the citation profiles of patent data. Second, the dissertation provides a comprehensive review of REMs, highlighting their evolution and identifying areas for further development. The introduction of STREAM addresses the computational challenges of applying REMs to large-scale networks. In an application to patent citations, it reveals non-linear patterns in patent citation rates, particularly during periods of heightened technological innovation. Finally, DREAM leverages neural networks to model non-linear effects in large dynamic networks, offering a scalable and robust solution for analyzing large relational datasets with complex drivers. To demonstrate the high flexibility of these models, an application to the European interbank market is presented.

# Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Ernst Wit, for his exceptional guidance and mentorship throughout my academic journey, from my earliest research endeavors to the completion of this Ph.D. thesis. His steadfast support, profound expertise, and thoughtful insights have not only shaped the research presented here but have also played a pivotal role in my growth as an independent scholar. His commitment to fostering a stimulating and encouraging environment has been crucial in helping me navigate the complexities of our field, sparking a genuine curiosity that will continue to guide my future research. I am sincerely thankful for the opportunity to have worked under his supervision.

Special thanks are due to Chiara Rodella, whose support, patience, and encouragement have been a constant source of strength throughout this journey. Her belief in me, even during the most challenging moments, has kept me motivated and focused. Her kindness, love, and understanding have made all the difference, and I feel incredibly lucky to have her by my side.

I would also like to express my heartfelt gratitude to my parents for their continuous support and belief in my abilities. Their encouragement and confidence in me have been invaluable, and I owe much of what I have achieved.

To my dear friend Niccolò, whose friendship and constant encouragement have been a source of motivation from kindergarten days to Volume 79 of *Social Networks*. His humor, insight, and unwavering support have made even the most challenging times more manageable.

To my colleagues and friends, I am deeply thankful for the moments we have shared and for the collaborative spirit that enriched my experience. Your insights, discussions, and camaraderie have been instrumental in the completion of this research project.

Lastly, I extend my sincere appreciation to everyone who has indirectly contributed to this research project. This includes the members of my dissertation committee for dedicating their time and expertise to reviewing my work and for providing valuable feedback.

# Contents

<b>Contents</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Analyzing the Patent Citation Network . . . . .	3
1.2 Dynamic Network Modelling . . . . .	4
1.3 Background and methods . . . . .	6
1.3.1 Introduction to Relational Event Modelling . . . . .	6
1.3.2 Importance of Non-linear Modelling . . . . .	8
1.4 Overview of the Dissertation . . . . .	9
1.4.1 Drivers of decrease of patent similarities from 1976 to 2021 . . . . .	9
1.4.2 Relational Event Modelling . . . . .	10
1.4.3 A Stochastic Gradient Relational Event Additive Model for Modelling US Patent Citations from 1976 until 2022 . . . . .	10
1.4.4 Modelling Non-linear Effects with Neural Networks in Re- lational Event Models . . . . .	11
1.4.5 Analyzing Non-linear Network Effects in the European In- terbank Market . . . . .	11
<b>2 Drivers of the Decrease of Patent Similarities from 1976 to 2021</b>	<b>12</b>
2.1 Introduction . . . . .	12
2.2 Materials and methods . . . . .	15
2.2.1 USPTO patent data . . . . .	15
2.2.2 The unreliability of institutional classification schemes . . . . .	16
2.2.3 Patent similarity based on pre-trained SBERT . . . . .	17
2.2.4 Modeling similarity scores through Generalized Additive Models . . . . .	20
2.3 Results . . . . .	22
2.4 Discussion . . . . .	25
2.5 Conclusion . . . . .	26

<b>3</b>	<b>Relational Event Modelling</b>	<b>28</b>
3.1	Introduction	28
3.2	Historical Context of Relational Event Models	30
3.3	Relational Data and Study Design	32
3.3.1	Network Structure	32
3.3.2	Sampling and Recording	32
3.4	Specifications of Relational Event Models	33
3.4.1	Types of Relational Event Models	33
3.4.2	Network Covariates	38
3.5	Estimation and Computation	41
3.5.1	Partial Likelihood Estimation	41
3.5.2	Baseline Hazard Estimation	44
3.5.3	Model Comparison and Diagnostic Tools	44
3.5.4	Bayesian Estimation of the Relational Event Model	45
3.5.5	Tools for Analyzing Relational Event Models	45
3.6	Applications of Relational Event Models	46
3.6.1	Communication	46
3.6.2	Ecology	47
3.6.3	Health and Healthcare	47
3.6.4	Political Science	48
3.6.5	Sociology	48
3.7	Open Issues and Challenges	49
3.7.1	Procedures for Assessing Goodness of Fit	49
3.7.2	Relational Big Data	49
3.7.3	Current Developments of Relational Event Models	50
3.8	Concluding Remarks	51
<b>4</b>	<b>A Stochastic Gradient Relational Event Additive Model for Modelling US Patent Citations from 1976 until 2022</b>	<b>53</b>
4.1	Introduction	53
4.2	Patent citations as event history data	55
4.3	Stochastic Gradient Relational Event Additive Model (STREAM)	58
4.3.1	Relational event model	58
4.3.2	Case-control sampling of the risk-set and logit approximation	59
4.3.3	Basis expansion of covariates	60
4.3.4	Recovering baseline hazard	62
4.3.5	Parameter estimation using stochastic gradient descent	62
4.4	Modeling patent citations	64
4.4.1	Potential drivers of patent citations	65



4.4.2	Implementation . . . . .	68
4.4.3	Interpretation of results on USPTO patent citation data 1976-2022 . . . . .	69
4.4.4	Estimated baseline hazard . . . . .	73
4.5	Conclusions . . . . .	75
<b>5</b>	<b>Modelling Non-linear Effects with Neural Networks in Relational Event Models</b>	<b>77</b>
5.1	Introduction . . . . .	77
5.2	Background and Methods . . . . .	79
5.2.1	Relational Event Model . . . . .	79
5.2.2	Nested case control sampling . . . . .	81
5.3	Deep Relational Event Additive Model . . . . .	82
5.3.1	Non-linear modelling with Neural Networks . . . . .	82
5.3.2	Uncertainty estimation with Gaussian Process Regression . . . . .	84
5.4	Simulation study . . . . .	86
5.4.1	True function recovery . . . . .	86
5.4.2	Accuracy comparison with GAM . . . . .	87
5.4.3	Time efficiency comparison with GAM . . . . .	88
5.5	US patent citation network . . . . .	89
5.6	Conclusions . . . . .	92
<b>6</b>	<b>Analyzing Non-linear Network Effects in the European Interbank Mar- ket</b>	<b>94</b>
6.1	Introduction . . . . .	94
6.2	The European Interbank Market . . . . .	96
6.3	Background and methods . . . . .	97
6.3.1	The Relational Event Model . . . . .	97
6.3.2	Nested case control sampling . . . . .	99
6.3.3	Modeling non-linear effects with neural networks . . . . .	100
6.3.4	Network statistics . . . . .	102
6.4	Results . . . . .	105
6.5	Conclusions . . . . .	107
<b>7</b>	<b>Conclusions</b>	<b>108</b>
7.1	Summary of Key Findings . . . . .	108
7.2	Open Challenges and Limitations . . . . .	110

---

<b>A</b>	<b>Supplementary materials for Chapter 4: “A Stochastic Gradient Relational Event Additive Mode for modelling US patent citations from 1976 until 2022”</b>	<b>112</b>
A.1	B-spline recursive formulation . . . . .	112
A.2	Model selection . . . . .	113
<b>B</b>	<b>Supplementary materials for Chapter 5: “Modelling Non-linear Effects with Neural Networks in Relational Event Models”</b>	<b>115</b>
B.1	Oscillatory activation functions . . . . .	115
B.2	ADAM . . . . .	116
B.3	Simulation study model selection . . . . .	117
B.4	Complex polynomials for simulation study . . . . .	118
B.5	US patent citation network model selection . . . . .	118
B.6	US patent citation network fitted effects . . . . .	119
	<b>Bibliography</b>	<b>122</b>

# Chapter 1

## Introduction

The idea that knowledge is a cumulative process has been widely acknowledged, tracing back to the earliest considerations of intellectual progress [Machlup, 1980]. At the dawn of the twentieth century, knowledge dynamics began to draw significant attention within the philosophy of science, becoming a central theme for many scholars [Kuhn, 1962; Merton, 1973]. Early notions of knowledge accumulation, which often depicted it as a straightforward and linear process, were rigorously challenged by philosophers like Karl Popper. Popper argued that human knowledge is inherently speculative, evolving within specific socio-historical contexts and driven by the need to solve particular problems [Popper, 1965]. He proposed a gradual and non-linear knowledge development model, where progress is not about the mere accumulation of facts but about the iterative refinement of theories and ideas. In stark contrast, Thomas Kuhn introduced the notion of “paradigm shifts”, suggesting that science advances not through a steady accumulation of knowledge but through periodic revolutionary changes that disrupt the status quo and lead to new frameworks of understanding [Kuhn, 1962]. These shifts represent profound transformations in scientific thinking, where entire fields can be redefined by a single breakthrough.

In recent decades, the exploration of knowledge dynamics has been significantly advanced by the proliferation of vast, reliable datasets, particularly those maintained by patent offices and scientific literature databases worldwide [Hall et al., 2001]. The sheer volume and accessibility of this information have opened new avenues for analyzing how knowledge evolves and spreads across different fields. Researchers have increasingly turned to citation networks as a powerful tool to study these dynamics, encompassing both the citations found in patents and those within the broader scientific literature [Trajtenberg and Jaffe, 2002]. Within this context, citation systems are often conceptualized as knowledge net-

works [Breschi and Lissoni, 2004], where each citation serves as an observable link between ideas, connecting past innovations with new ones. The practice of requiring citations to reference “prior art” or previous research adds a layer of validity to this interpretation, providing a structured way to trace the lineage of ideas and innovations [Albert et al., 1991].

The primary conceptual argument driving this line of research is that citation networks can be viewed as observable trails of otherwise unobservable knowledge flows. These flows represent the transmission of ideas across time and space, linking disparate concepts and fostering breakthrough innovations that propel entire industries forward [Arts and Veugelers, 2014]. Moreover, understanding these networks offers valuable insights into the mechanisms of knowledge diffusion, revealing how ideas are transmitted, transformed, and integrated into the collective knowledge base. As citation networks grow and evolve, they reflect the dynamic interplay between innovation and tradition, highlighting the ongoing conversation between past discoveries and future possibilities. This perspective not only enriches our understanding of how knowledge advances but also underscores the critical role of interconnectedness in driving scientific and technological progress. As our understanding of knowledge dynamics deepens, it becomes increasingly clear that knowledge drives innovation. Innovation, in turn, is a fundamental economic indicator that enables companies and countries to assess their growth across various sectors. This raises the question of where innovation originates. One approach to addressing this question is to analyze patent data. As legal documents, these are designed to protect inventors’ rights to benefit from their inventions by temporarily preventing competitive exploitation. This line of research rests on the premise that patents serve as proxies for innovation and knowledge creation. Earlier studies [Wang et al., 2010; Park et al., 2013] utilized patent data to compute statistical indicators that help measure the productivity and value of investments at both company and country levels. These indicators suggest that innovation stems from investments in innovative activities.

The central hypothesis and common thread connecting the research papers in this dissertation is that innovation networks, as evidenced by patent citations, possess intrinsic predictive value in determining the causes of the influence and success of innovations. By examining these networks, we can gain insights into the dynamics of knowledge flow and the factors that drive technological advancement and economic growth.

However, patent citations typically come in vast datasets, comprising millions of records, which presents significant challenges for analysis. Such a volume of data can overwhelm traditional analytical methods, necessitating the devel-

opment of more sophisticated tools to extract meaningful patterns and insights. Conventional network science models, while powerful, are often not equipped to handle datasets of this magnitude, limiting their applicability to such large-scale systems. To address these challenges, the scope of this dissertation is dual. First, it aims to conduct a comprehensive analysis of the patent citation network, uncovering the intricate relationships and dynamics that underpin the innovation process. Second, it seeks to advance the current state of network science modelling by developing and refining models that are capable of handling and analyzing large datasets effectively. This double focus not only contributes to our understanding of innovation networks but also pushes the boundaries of what it is possible to model in network science, paving the way for new methodologies that can be applied to other large-scale complex systems.

## 1.1 Analyzing the Patent Citation Network

The analysis of patent data has long been recognized as a valuable research activity in empirical studies on the economics of innovation and technical change [Hausman et al., 1984]. Since the early 1980s, researchers have been exploring patent citation modeling to uncover the underlying patterns in technological innovation. For instance, Carpenter and Narin [1983] analyzed citation patterns to determine the reliance of U.S. technologies on foreign sources. This work laid the foundation for the use of patent citations to address a wide range of questions, such as whether knowledge innovation is geographically localized [Jaffe et al., 1993] or whether citation patterns correlate with the socioeconomic value of a particular patent [Trajtenberg, 1990]. Over the years, patent citation analysis has become an established and essential component of scientometrics, offering valuable insights into the dynamics of knowledge creation and dissemination [Meyer, 2000].

However, traditional patent citation analyses often do not model citation patterns in their entirety. Instead, they tend to utilize these patterns to derive ad hoc indicators that are primarily focused on economic or financial outcomes. Much of the work in this area has concentrated on correlating summary measures of citation patterns with factors such as geographic location, monetary value, or the appropriation and return on R&D investments [Levin, 1988]. These analyses have proven valuable in identifying key indicators of technical change, productivity, inventive activity, and economic growth at various levels, including individual companies, specific sectors of the economy, and entire national economies [Griliches et al., 1986].

In recent years, there has been a growing recognition of the value of analyzing patent citations not just as isolated indicators but as part of a broader network of knowledge flows. The idea is based on the intuition that the knowledge network induced by observed patterns of patent citations exhibits characteristics typical of complex adaptive systems [Fleming and Sorenson, 2001; Radicchi et al., 2012; Sorenson et al., 2006]. These systems are marked by intricate interactions among their components, leading to emergent properties that cannot be easily predicted by examining individual elements in isolation. For example, within the patent citation network, one might observe various forms of clustering, where certain patents or groups of patents become central hubs of innovation, heavily cited by subsequent patents and thereby influencing future technological developments [Kejřar et al., 2011].

In this regard, the analysis of patent citations as a network offers several advantages. By treating citations as links in a broader network, researchers can better understand how knowledge flows through different technological domains, how ideas are recombined to generate new innovations, and how certain patents play a pivotal role in driving technological change. Network analysis allows for the identification of key patents that act as bridges between different technological fields, facilitating the transfer of knowledge across domains.

Moreover, this network turn can help to uncover not only the structural properties of the patent citation network but also explain the presence of links, thus addressing the crucial question of why a particular citation was made. Patents are typically cited for a variety of reasons, reflecting their importance, relevance, and influence within a specific technological field. This is often attributed to the novelty that a patent introduces; a patent that presents a groundbreaking or highly innovative concept is more likely to be cited by subsequent patents that build upon or incorporate the new idea. However, novelty is not the only factor. Other contributing factors may include the technical proximity of a patent, which can significantly influence its citation frequency [Jaffe et al., 1993]. Additionally, the timing of a patent's issuance relative to the technological lifecycle is another critical factor, as patents issued at the forefront of an emerging technology trend or during periods of rapid technological advancement are more likely to be cited [Trajtenberg, 1990].

## 1.2 Dynamic Network Modelling

Patent citations occur at specific points in time, making the timing of these citations crucial for understanding the dynamics of innovation and its diffusion. A

key question that arises is: why is a citation observed at a particular moment, rather than earlier or later? Various factors likely influence this timing, including the age of the patent, its influence within the field, the relevance of the scientific domain at that particular time, and other contextual elements.

Marco [2007] proposed an event history analysis to study the timing of patent citations, positing that the citation hazard  $h_i(t)$  of a patent  $i$  follows a Weibull distribution and depends on a set of covariates  $x_i$  as follows:

$$h_i(t) = \lambda_i \rho (\lambda_i t)^{\rho-1},$$

where  $\log \lambda_i = x_i^\top \beta$ . Marco’s analysis revealed that the hazard rate is influenced by various factors, including the application field, grant year, patent age, and others, such as a frailty component to account for unobserved heterogeneity. The strength of this approach lies in its direct modeling of the citation process, which allows for the inclusion of control variables alongside the variables of interest. This enhances the likelihood that the detected effects are genuine, rather than artifacts of confounding variables. However, one notable limitation of this approach is that it overlooks an essential explanatory component of the citation process: the underlying network structure within which these citations occur.

In contrast, Acemoglu et al. [2016] introduced a network-based perspective on patent citations to uncover the so-called “innovation network”. Their approach aimed to explore how citation patterns serve as evidence of cross-fertilization between different technology categories and subcategories. The induced category network revealed strong self-citation patterns within each category but also highlighted significant cross-category dissemination of ideas, particularly from the chemical category to various other technological fields. The network perspective is especially powerful in describing the interconnectedness of innovation, as it enables the application of social network analysis techniques to the study of citation networks.

However, to capture the evolution of this citation network, Acemoglu et al. [2016] divided citation patterns into 5-year time windows, repeating the analysis within each window. While this method provides a temporal snapshot of network dynamics, it introduces artificial boundaries that may obscure the continuous nature of the innovation process, potentially leading to misinterpretations of the results.

A significant development in the study of patent citation networks involves the use of Exponential Random Graph Models (ERGMs) [Lusher et al., 2012] to model the structural dependencies within these networks. Chakraborty et al. [2020] highlighted that various patent characteristics influence citation forma-

tion, where both technical attributes (e.g., technological class, field of application) and social processes, such as preferential attachment, play crucial roles in shaping citation patterns. However, ERGMs typically model networks as static snapshots, thereby not accounting for the temporal dynamics inherent in citation processes. This static perspective overlooks the sequential and time-dependent nature of citations, making it challenging to address questions about time-sensitive phenomena, such as the decay of influence over time or event-driven effects that could continue to shape citation behavior.

Recent advancements in methodology have begun to bridge the gap between event-history analysis and network analysis by incorporating time-varying network structures into hazard models. For instance, [Butts \[2008\]](#) introduced Relational Event Models (REMs) that allow for the dynamic modeling of interactions within a network, accounting for the changing structure of relationships over time. Similarly, [Brandes et al. \[2009\]](#) proposed the use of dynamic network analysis in conjunction with survival analysis techniques to model the co-evolution of networks and events, offering a more subtle understanding of how networks influence and are influenced by temporal processes. Integrating these approaches presents an opportunity for a more comprehensive understanding of the dynamic nature of patent citation networks. By combining event history analysis with network analysis, future research could account for both the timing of citations and the evolving network structure in which these citations are embedded.

## 1.3 Background and methods

In this section, we provide an introductory overview of some of the key theoretical and methodological concepts that will be expanded in the research presented in this dissertation. This dissertation primarily focuses on the broad framework of Relational Event Models (REMs). By exploring how REMs work, we will then introduce other key concepts that revolve around non-linear modelling, which will be crucial for expanding the current methodological framework of REMs. Specifically, we will introduce the concept of non-linear modelling using spline approaches and deep learning methodologies.

### 1.3.1 Introduction to Relational Event Modelling

The class of statistical models known as Relational Event Models (REMs) was first introduced by [Butts \[2008\]](#) and [Perry and Wolfe \[2013\]](#). These models have become highly relevant in the field of network science, as they assume that the



dynamics of nodes interacting within a graph can be modeled as a series of subsequent events occurring over time. REMs aim to capture the evolving dynamics of these relationships as time progresses. In this regard, REMs extend traditional event history analysis by focusing exclusively on interactions between entities, allowing researchers to model how past events influence future interactions within a network.

REMs analyze data that are typically collected in the form of edgelists, where each event  $e_i$ , for  $i = 1, \dots, n$ , is recorded as a triple  $e_i = (s_i, r_i, t_i)$ , with  $s_i$  denoting the sender (the initiator of the interaction),  $r_i$  the receiver (the recipient of the interaction), and  $t_i$  representing the time at which the event occurs. Based on this setup, we define a counting process  $N_{sr}(t)$  that represents the number of interactions between a sender  $s$  and a receiver  $r$  up to time  $t$ , defined as

$$N_{sr}(t) = \#\{s \text{ interacts with } r \text{ up to time } t\}.$$

According to the Doob-Meyer decomposition theorem, this local submartingale can be decomposed into a predictable increasing process  $\Lambda_{sr}(t)$  and a martingale noise term  $M_{sr}(t)$ , i.e.

$$N_{sr}(t) = \Lambda_{sr}(t) + M_{sr}(t).$$

If it exists, it is possible to describe the tendency for  $s$  to interact with  $r$  through the stochastic intensity function (i.e., the hazard of an event),

$$\lambda_{sr}(t) = \frac{d\Lambda_{sr}}{dt}(t),$$

which describes the instantaneous propensity for the event  $(s, r)$  to occur at time  $t$ . Given the history of previous events  $\mathcal{H}_{t-}$  up to time  $t$ , the intensity function can be modeled using the proportional hazard function introduced by Cox [1972]. The intensity function is then expressed as the product of a baseline hazard  $\lambda_0(t)$  and an exponential function of covariates  $x_{sr}(t)$  and parameters  $\beta$ :

$$\lambda_{sr}(t \mid \mathcal{H}_{t-}) = \lambda_0(t) e^{\sum_{k=1}^q \beta_k x_{srk}(t)} \mathbb{1}_{\{(s,r) \in R(t \mid \mathcal{H}_{t-})\}},$$

where  $R(t \mid \mathcal{H}_{t-})$  represents the risk set of potential events at time  $t$ , i.e., the set of all potential events that could have occurred at time  $t$ . The indicator function  $\mathbb{1}$  accounts for the presence of the observed event  $(s, r)$  in the risk set. The history  $\mathcal{H}_t$  is a filtration that describes the state of the network and covariate process up until time  $t$ .

The covariates  $x_{sr}(t)$  are edge-specific risk determinants that drive the relational

process. In network science terminology, these covariates can be either endogenous or exogenous. Endogenous covariates depend on the history of past interactions within the network, capturing the internal dynamics of the system. Exogenous covariates, on the other hand, are based on external characteristics of the nodes (monadic covariates) or pairs of nodes (dyadic covariates), such as node attributes or geographic proximity. By incorporating these covariates, REMs can account for a wide range of factors that influence the likelihood of an event occurring, allowing for a detailed examination of the drivers behind interactions within the network.

### 1.3.2 Importance of Non-linear Modelling

Many statistical models assume a linear relationship between covariates and coefficients. While this assumption allows for a straightforward and simple definition of numerous statistical models, it can sometimes lead to an unrealistic interpretation of reality. In fact, many complex relationships cannot be adequately represented by linear models. For example, diminishing returns, exponential growth or decay, and threshold effects would be poorly captured in a linear context. One of the most practical tools for introducing non-linearity into a statistical model is the use of splines. Splines are piecewise polynomial functions that approximate complex, smooth curves in data. By modelling different polynomial functions in different regions of the data, splines provide the flexibility to model non-linear relationships without the need to manually specify a particular functional form.

Incorporating splines into REMs offers several advantages. First, it enables the model to capture more nuanced patterns in the timing and sequence of events, such as changes in the rate of interaction over time or varying effects of covariates across different periods. This flexibility is particularly valuable when relationships between entities are likely to be influenced by multiple interacting factors.

However, the use of splines also introduces challenges. The expansion from univariate covariates to model matrices significantly increases computational requirements. Without adequate strategies, splines could become a less appealing option when dealing with large datasets.

This challenge drove the exploration of deep learning approaches in this dissertation, particularly for enhancing the modeling of relational events in large-scale networks. According to the Universal Approximation Theorem [[Hornik, 1991](#)], a feedforward neural network with at least one hidden layer and a sufficient number of neurons can approximate any continuous function on a compact subset

of  $\mathbb{R}^n$  to an arbitrary degree of accuracy [LeCun et al., 2015; Goodfellow et al., 2016]. This theorem highlights the power of neural networks as universal function approximators, capable of non-parametrically modeling complex non-linear relationships. Moreover, deep learning models are specifically designed to handle large datasets, making them ideal candidates for integration with REMs to address issues related to dimensionality.

Thus, the Universal Approximation Theorem and the mathematical techniques used to estimate neural networks in large contexts provide a strong theoretical foundation for using neural networks to enhance the modeling of large event sequences and capture intricate patterns and dependencies within event data. As such, integrating neural networks with REMs not only extends the capabilities of these models but also opens new routes for modeling and analyzing more complex effects, leading to new interpretations of the dynamic processes that govern network interactions.

## 1.4 Overview of the Dissertation

This dissertation comprises five distinct research projects, each addressing a unique aspect of modeling the patent citation network and large networks. Each project is presented as an individual chapter within this dissertation. These chapters represent independent contributions, each of which has been published in peer-reviewed scientific journals. The publication details for each project are provided at the beginning of the respective chapters.

The organization of the dissertation allows for a comprehensive exploration of the central research problem from multiple perspectives. Each chapter not only builds upon the previous work but also develops new methods, focus on important applications, or identify a general theoretical framework. Below we briefly describe the structure of the thesis, which besides this introduction consists of five additional chapters.

### 1.4.1 Drivers of decrease of patent similarities from 1976 to 2021

The citation network of patents citing prior art arises from the legal obligation of patent applicants to properly disclose their invention. One way to study the relationship between current patents and their antecedents is by analyzing the similarity between the textual elements of patents. Many patent similarity indicators have shown a constant decrease since the mid-70s. Although several explanations have been proposed, more comprehensive analyses of this phenomenon

have been rare. In this project, we use a computationally efficient measure of patent similarity scores that leverages state-of-the-art Natural Language Processing tools, to investigate potential drivers of this apparent similarity decrease. This is achieved by modeling patent similarity scores using generalized additive models. We found that non-linear modeling specifications can distinguish between distinct, temporally varying drivers of the patent similarity levels that explain more variation in the data compared to previous methods. Moreover, the model reveals an underlying trend in similarity scores that is fundamentally different from the one presented previously.

#### 1.4.2 Relational Event Modelling

Advances in information technology have increased the availability of time-stamped relational data such as those produced by email exchanges or interaction through social media. Whereas the associated information flows could be aggregated into cross-sectional panels, the temporal ordering of the events frequently contains information that requires new models for the analysis of continuous-time interactions, subject to both endogenous and exogenous influences. The introduction of the *Relational Event Model* (REM) has been a major development that has led to further methodological improvements stimulated by new questions that REMs made possible. In this review, we track the intellectual history of the REM, define its core properties, and discuss why and how it has been considered useful in empirical research. We describe how the demands of novel applications have stimulated methodological, computational, and inferential advancements.

#### 1.4.3 A Stochastic Gradient Relational Event Additive Model for Modelling US Patent Citations from 1976 until 2022

Until 2022, the US patent citation network contained almost 10 million patents and over 100 million citations, presenting a challenge in analyzing such expansive, intricate networks. To overcome limitations in analyzing this complex citation network, we propose a stochastic gradient relational event additive model (STREAM) that models the citation relationships between patents as time events. While the structure of this model relies on the Relational Event Model, STREAM offers a more comprehensive interpretation by modeling the effect of each predictor non-linearly. Overall, our model identifies key factors driving patent citations and reveals insights in the citation process.

#### 1.4.4 Modelling Non-linear Effects with Neural Networks in Relational Event Models

Dynamic networks offer an insight of how relational systems evolve. However, modelling these networks efficiently remains a challenge, primarily due to computational constraints, especially as the number of observed events grows. This research project addresses this issue by introducing the Deep Relational Event Additive Model (DREAM) as a solution to the computational challenges presented by modelling non-linear effects in Relational Event Models (REMs). DREAM relies on Neural Additive Models to model non-linear effects, allowing each effect to be captured by an independent neural network. By strategically trading computational complexity for improved memory management and leveraging the computational capabilities of graphic processor units (GPUs), DREAM efficiently captures complex non-linear relationships within data. This approach demonstrates the capability of DREAM in modelling dynamic networks and scaling to larger networks. Comparisons with traditional REM approaches showcase DREAM superior computational efficiency. The model potential is further demonstrated by an examination of the patent citation network, which contains nearly 8 million nodes and 100 million events.

#### 1.4.5 Analyzing Non-linear Network Effects in the European Interbank Market

Diverging from patent citations, we demonstrate the vast range of applications that the Deep Relational Event Model offers by analyzing the temporal evolution of financial transactions within the European interbank market. The European interbank market has been a crucial component of the financial system where European banks engage in short-term borrowing and lending amongst themselves. This market primarily facilitates the redistribution of liquidity within the financial system, allowing banks with surplus funds to lend to those experiencing shortfalls. The sheer number of interactions has prevented until now a detailed analysis of the shape of the network dynamics. Our results reveal distinct patterns in the network's behavior before and after the 2008 financial crisis, highlighting shifts in transaction dynamics and the roles of financial institutions. The analysis uncovers trends in reciprocity, cyclic closure, and other network dynamics, offering insights into how the crisis influenced the market structure and interactions.

## Chapter 2

# Drivers of the Decrease of Patent Similarities from 1976 to 2021

The following chapter was published as:

Filippi-Mazzola, E., Bianchi, F., and Wit, E. C. (2023). Drivers of the decrease of patent similarities from 1976 to 2021. *Plos One*, 18(3), e0283247.

### 2.1 Introduction

Understanding the characteristics of ground-breaking innovations is crucial for technology-based firms striving for success [Henderson and Clark, 1990]. Patent indicators serve this purpose and support, among others, the development of product strategies [Wang and Chen, 2019; Park et al., 2013], monitoring of existing technological trends [Wang et al., 2010], the detection of promising opportunities of investments [Yoon and Kim, 2012], the assessment of the technological impact of novel applications [Verhoeven et al., 2016; Veugelers and Wang, 2019], and recognizing similar technologies [An et al., 2021; Kuhn et al., 2020; Whalen et al., 2020; An et al., 2018].

Patent indicators using institutional classifications and citation information are predominant [Gress, 2010; Verhoeven et al., 2016; Acemoglu et al., 2016; Veugelers and Wang, 2019] in patent analysis. Patent classification systems like the International Patent Classification (IPC) are usually processed for identifying technologically similar patents. However, sharing the same patent class may not fully capture technological relatedness. Despite numerous methods for analyzing technological relatedness and closeness based on such classes [Yan and Luo, 2017], their usage can be problematic when patents need to be identified, compared, or matched with similar technologies.

In contrast, patent indicators using patent descriptions and lexical contents are less common in patent analysis. A keyword-based approach using frequency and co-occurrence of contents is typically used for computing the technological similarity between pairs of patents [Younge and Kuhn, 2016]. Within the set of patent indicators, patent similarity is a fundamental measure in the evaluation of technological novelty [Wang and Chen, 2019] and infringement risks associated with others using or selling inventions without authorization [An et al., 2021]. Patent descriptions can be mined for combinations of words and unique expressions for text-based indicators for patent similarity. This transforms unstructured textual data into actionable knowledge through latent relationships between patent documents [Immordino, 2019].

With the development of new and more sophisticated deep learning techniques, Natural Language Processing (NLP) tools have been proven to provide valid alternatives to canonical technology class measurements. The idea is to use the textual elements of patents as inputs for defining vectors of similarity. In this way, it is possible to use continuous distance measures between any two patents, e.g., Euclidean distance, cosine similarity, or Mahalanobis distance to measure patent (dis)similarity. Although the idea of mapping patents into a vector space can be traced back to Jaffe [1986, 1989], only recently these methods have been applied to patent analysis. For example, Younge and Kuhn [2016] used a bag of words methodology [Turney and Pantel, 2010] to develop a machine-automated patent-to-patent similarity measure based on the technical descriptions of patent applications. Adopting the same approach, Kuhn et al. [2020] analyzed pairs of patent citations in the US between 1975 and 2014. Simple vocabulary-based approaches of textual similarity scores across citing and cited patents, may contain major drawbacks caused by the sparsity of the output matrix. Although there have been developments to address this weakness, like the automatic indexing and retrieval approach [Deerwester et al., 1990], a neural network (NN) approach, such as the one proposed by Whalen et al. [2020], is preferred as semantics and context are prioritized within the estimated positional embeddings. The introduction of language based NN models has opened up the way for more complex applications within patent similarity analyses. While early contributions have focused on patent abstract data for correctly classifying patents into their technological classes [Lee and Hsiang, 2020; Bekamiri et al., 2021], the focus is now shifting towards mapping patents into multidimensional spaces to detect patterns and gain relational insights. In this regard, Hain et al. [2022] proposed using a K-nearest-neighbors algorithm to spot closely related patents by training a Word2Vec NN model [Mikolov et al., 2013] on 48 million abstracts. Regardless of the amount of data processed, the computational cost of these approaches are

high. Instead, the current availability of generic models pre-trained on massive corpora is rapidly increasing [Liu et al., 2021]. This has enabled researchers to unlock vast complex natural language models with fewer computational resources, paving the way for a new set of tools.

In the context of textual similarity analysis in patent citations, Kuhn et al. [2020] and Whalen et al. [2020], noted a decrease in the average textual similarity per year between citing and cited patents. The aim of this manuscript is to investigate the drivers of patent similarity decline during a period of approximately forty years, from 1976 to 2021, with 1976 the year when the *US Patent Trading Office* (USPTO) started collecting the full text for all granted patents in digital databases. Previous studies of the decrease of patent similarity attribute this drop to fundamental changes that occurred in the data generation process. Kuhn [2010] claim that legal changes in the applicant’s duty of disclosure has led to a drastic increase in the number and scope of cited references. As a consequence, more citations have been included that are further afield from the citing patent. Pursuing this hypothesis, Kuhn et al. [2020] show how the skewed distribution of backward citations has become less informative for research practices, as a small minority of patent applications are now generating a large majority of patent citations in the overall citation network.

We propose to use pre-trained models to compute the embeddings. In this sense, we avoid any computational procedure by proposing instead a ready-to-use approach for computing similarity scores. We focus on patent abstracts that contain the most concise information regarding the patenting technology [Choi et al., 2022; Hain et al., 2022]. Thanks to the reduced size of the abstract corpus, we are able to compute the positional embeddings via a pre-trained SBERT model in a reasonable amount of time. We encode the entire set of roughly 10 million abstracts into fixed sized vectors and compute the vector of similarity scores across 100 million patent citations through a parallelized lazy loading scheme.

We will first describe the USPTO patent data on which we base our analysis. We then describe the SBERT embedding of the abstract data and the calculation of the patent similarity scores. The scores confirm the downward trend in the patent similarity scores. Then we propose various Generalized Additive Models (GAMs) [Hastie and Tibshirani, 1986] with the aim of detecting the drivers of patent similarity over time, in particular, whether this is a temporal endogenous process or due to exogenous patent attributes. In contrast to previous studies, our approach also aims to resolve the problem of the temporal boundary of the citation network by considering the time lag between the citing and the cited patents.



## 2.2 Materials and methods

### 2.2.1 USPTO patent data

Intellectual property history can be traced back to the 19th century when the first patenting office was established in Paris. Since then the patenting documentation has evolved and the availability of patent data has grown dramatically. One of the main challenges of analyzing patent data is retrieving the required information from the large amount available. Moreover, patents are legal documents, mostly consisting of textual elements. Unfortunately, the non-availability of standardized patent formats through the years has caused difficulties in building standardized data bases. Moreover, the juridical procedures of patenting are country-specific. This creates inconsistencies in the data from different countries, as some patenting offices will use different citation procedures. A striking example is a distinction in the citation process between the USPTO and the European Patent Office (EPO). Both the USPTO and the EPO require applicants to fulfill their duty of disclosure by citing all the required prior arts. The examiner committee of the USPTO adds citations to the application by integrating all those prior arts that are considered relevant for the patent to be correctly disclosed. On the other side, the EPO examiner committee does not include any further citations in the examination process. The committee limits its range of action by evaluating the validity of the patent combined with the disclosed prior arts. From this perspective, a combined analysis of multiple patenting offices' data would result in unreliable conclusions.

For this reason, we focus our analysis exclusively on patents that have been issued by the USPTO from January 1976 up to September 2021. Starting from 1976, the USPTO has created an online public repository storing all the issued patents, including guidelines for data quality and standardization in the textual component of submitted legal documents. Although the USPTO data are broadly available across different periods, we have noted that most common repositories contain many inaccuracies. Such issues are usually the result of heavy preprocessing procedures used to combine, correct, or fill missing values from distinct sources to integrate the range of data that the USPTO provides publicly. To retain the highest quality possible in our dataset, we avoid third-party preprocessing and download data directly from the USPTO digital repository (<https://bulkdata.uspto.gov/>). After downloading the required XML files, these were processed and combined to obtain CSV files through an open-source software tool (available at: <https://github.com/iamlemec/fastpat>).

Our dataset consists of a time-stamped citation network along with patent at-

tributes. For each granted patent we consider its backward citations, and for each patent in the dataset we include International Patent Classification (IPC) codes. In line with the network science vocabulary, we refer to citing patents as senders and to cited patents as receivers.

### 2.2.2 The unreliability of institutional classification schemes

Patent classification schemes like those illustrated in the IPC Table 2.1 are designed for examiners to ease the examination process of patent applications by rapidly searching for similar or related technologies. Studies on innovation use such instruments to analyze potential technological patterns, usually through similarity levels derived from co-class proximity measures [Yan and Luo \[2017\]](#).

Table 2.1: International Patent Classification (IPC) scheme for a generic patent classified as A01C 3/04.

<b>A</b>				
<i>Section</i>	<b>A01</b>			
	<i>Class</i>	<b>A01C</b>		
		<i>Subclass</i>	<b>A01C 3/00</b>	
			<i>Group</i>	<b>A01C 3/04</b>
				<i>Subgroup</i>

It has been argued that institutional classification schemes do not offer a reliable picture of patent similarity. [Younge and Kuhn \[2016\]](#) explain how many sources of bias may emerge when comparing patents through the technological classes they belong to. On the one hand, patent classes are not fixed – i.e., new technological classes may be created and old classes may be merged, split, and/or reassigned in a way that affects the depth of technological spaces. On the other hand, the classes may be too broad or too tight, leading to inaccurate comparisons.

We compare a random sample of 1 million citations through their *sections* and *sub-classes* as defined in the IPC classification, where sections are the broader category and sub-classes are the preferred level of analysis in empirical applications. Figure 2.1 clearly shows that any measure of patent similarity based on institutional classifications suffers from a selection bias in the hierarchy of classification layers. While technology sections tend to self-cite, which produces higher similarity scores, technology sub-classes tend to cite outside of their area.

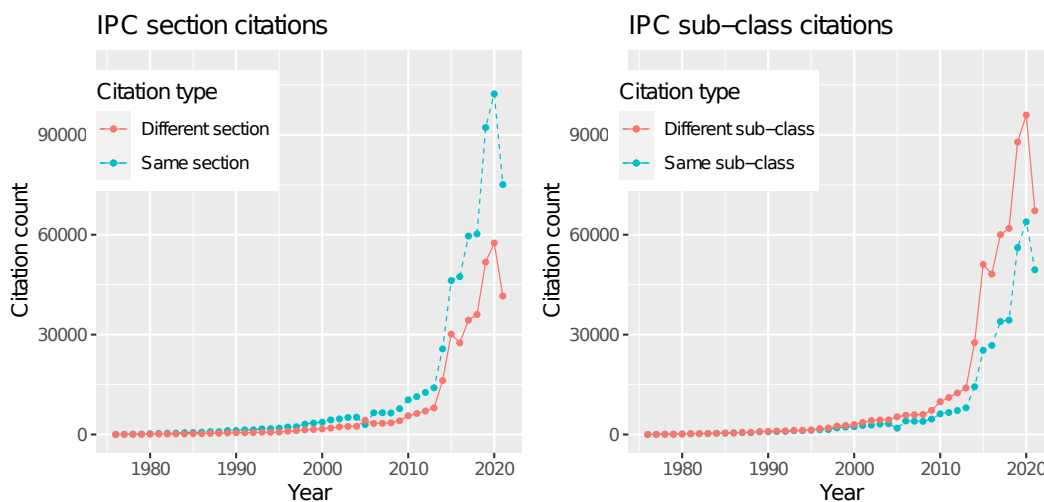


Figure 2.1: IPC citations comparison. Amount of citations within the same section (left) and sub-class (right) during the observation period.

### 2.2.3 Patent similarity based on pre-trained SBERT

NLP tools can be used to interpret text and translate it into a mathematical form for other algorithms to accomplish predefined tasks. What has been a revolution for NLP was the introduction of the Transformer architecture [Vaswani et al., 2017] in a field that was previously dominated by Recurrent Neural Networks and Long-Short-Term-Memory Networks. The great step that made the Transformers the new go-to tools for NLP is the focus on attention mechanisms [Bahdanau et al., 2014] which replaced recurrence functions with a large number of parameters. Instead of processing the sentence sequentially (or word-by-word), the attention mechanism processes the entire sequence to give weights to the input. In this sense, it decides how much each word in the input is associated to the sentence. In this way, it runs a probabilistic-like approach that prioritizes certain parts over others. Transformers then combine an encoder-decoder architecture which solely relies on the attention mechanism to forward more parts of the input sequence at once (see Rothman [2021] for an overview).

The *Bidirectional Encoder Representations from Transformer* (BERT) [Devlin et al., 2019] takes this concept and extends it by using the context coming from both sides of the current analyzed part of the input. This change is significant as often a word may change meaning while the sentence develops. Each word added augments the overall meaning of the word being analyzed. The more words that are present in total in each input sequence, the more ambiguous the word in

focus becomes. BERT accounts for the augmented meaning by reading the input bidirectionally, accounting for the effect of all other words in the input on the focus word and eliminating the left-to-right shift that biases words towards a certain meaning as the sentence progresses.

Although BERT outperforms any other benchmark that was set by previous NLP tools to encode the meanings of words into queries, it does not perform well when it comes to comparing similarities of entire sentences. A large disadvantage of the BERT network structure as presented by [Devlin et al. \[2019\]](#) is that independent sentence embeddings are not computed, which makes it difficult to derive sentence embeddings from BERT. In response, [Reimers and Gurevych \[2019\]](#) modified the standard BERT architecture for semantic textual similarity, called Sentence-BERT (or SBERT), while also reducing computing time. The main difference with the regular BERT architecture is encoding the semantic meaning of whole sentences instead of individual words. The SBERT architecture is characterized by a so-called twin network, which allows it to process two sentences simultaneously. The twins are identical down to every parameter, which allows it to think of this architecture as a single model used multiple times. At the end of the SBERT pipeline, the model contains a final pooling layer that enables the creation of a fixed-size representation for input sentences of varying lengths. With this, it is possible to encode documents into fixed-sized vectors, while taking their semantics into account.

The downside of models based on Transformer architectures is that these are among the most computationally intensive Neural Networks to train. The peer-to-peer *Hugging Face* repository solves this deficiency and allows researchers to upload trained models in an open-source fashion. With this tool, access to deep and complex neural networks is within the reach of every user. Moreover, pre-trained SBERTs have two other advantages over competing models. The first is that SBERT pre-trained models uploaded on *Hugging Face* are trained either for specific or general purposes. General purpose models are trained on billions of generic documents, which grants flexibility to use SBERT for any task. The second one is the ease of use granted by the package *Sentence Transformers*, which simplifies the procedure of creating and downloading the pre-trained weights from *Hugging Face* with a few lines of code. These two reasons, combined with the established benchmarks that SBERT has set in the field of NLP, make this the go-to model for our task. The *Sentence Transformers* package in python gives access to pre-trained models from the *Hugging Face* repository that encodes sequences into fixed-sized vectors. From this library, we downloaded a model, trained and fine-tuned on more than one billion public documents that encodes texts into vectors of size 384.

Similarly to Whalen et al. [2020] and Choi et al. [2022], we removed non-utility patents (such as plants or designs) from our data when computing embeddings. In this way, we encoded approximately 7.5 million patents into a fixed-size space through SBERT. By parallelizing a scheme of lazy loading procedures, we managed to compute the patent similarity scores for almost 100 million patent citations within minutes. Confirming results from previous studies, Figure 2.2 shows that the average similarity per year between citing papers is decreasing over time. The cosine similarity ranges between -1 and 1 by construction. We multiplied them by 100 for ease of representation, thus making the range of potential scores range between -100 and +100.

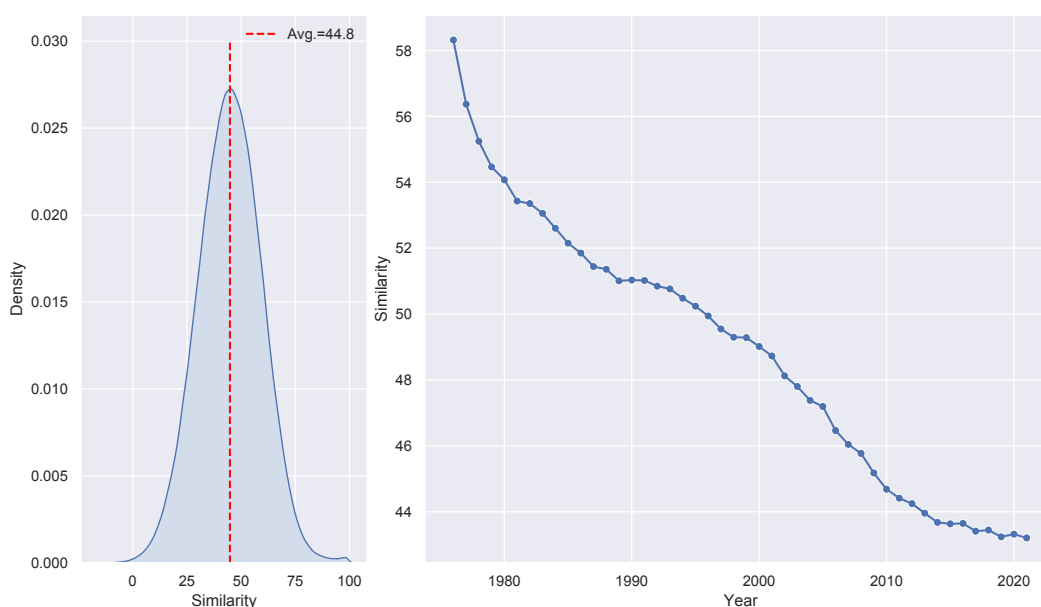


Figure 2.2: Citation textual similarity. Left: textual similarity distribution. Right: average textual similarity per year.

Embeddings were computed on a cluster node in parallel with two NVIDIA graphic processor units, models GeForce GTX 1080 Ti with 3584 CUDA cores and 100 GB of RAM. Computation time stands within one hour (more information on the resources that were used can be found on <https://intranet.ics.usi.ch/HPC>). Thanks to the usage of a pre-trained architecture, there is no training involved in the computation of the embeddings. This leaves the machine resources to take advantage of this method by taking the abstracts as input and providing only one forward passage among the neural network. Given this simplified procedure, the same results can be obtained within a reasonable time using less powerful computational resources — e.g., standard Colab notebooks.

### 2.2.4 Modeling similarity scores through Generalized Additive Models

It has been claimed that due to the changes in the generative process of patent citations, these citations have become less informative and representative [Kuhn et al., 2020]. We argue instead that with the correct application of informative statistical models, it is still possible to gain important insights on the main drivers of the decrease of patent similarity. As such, we argue that the backward citation process still plays a major role in determining the technological proximity of patents and the direction in which the network of citations is expanding.

We propose to model textual similarity scores through Generalized Additive Models (GAMs) [Hastie and Tibshirani, 1986] by extending the approach of Kuhn et al. [2020]. GAMs can be used to estimate the non-linear effects of covariates on the dependent variable. More in detail, while in linear models the predictor is a weighted sum of the  $p$  covariates,  $\sum_{j=1}^p \beta_j x_j$ , in GAMs this term is replaced by a sum of functions, e.g.  $\sum_{j=1}^p \sum_{l=1}^q \beta_j b_l(x_j)$ , where the  $b_1(\cdot), \dots, b_q(\cdot)$  are specific parametric basis functions – e.g. smoothing splines or complex polynomial splines. Essentially, GAMs are particularly useful for uncovering nonlinear drivers of some processes.

**Model 0.** Using this modeling technique, the decrease of patent similarity can be visualized by a simple GAM with the patent publication date as a unique covariate modeled by a smooth term (*Model 0*).

**Model 1.** The average decrease of similarity in backward citations is associated with an increase in the average temporal lag elapsed between citing and cited patents (see Figure 2.3). This result seems to suggest that applicants and examiners cite prior arts which are increasingly temporally distant from their application/grant time. By itself, this effect could be a source of the reduction of similarity levels as the innovation process gives reasons to believe that temporally distant technologies are less similar to newer ones. In addition, in a period of approximately forty years, the legal and technical language has seen some important changes. Although the usage of SBERT should eventually mitigate the change of language as the model would account for context and semantics, the language evolution follows the technological development present inside patents, thus increasing the reduction of a potential similarity effect.

In our model specification, the temporal component of patent citations is captured by the covariate *temporal difference*. Its effect on patent similarity is mod-

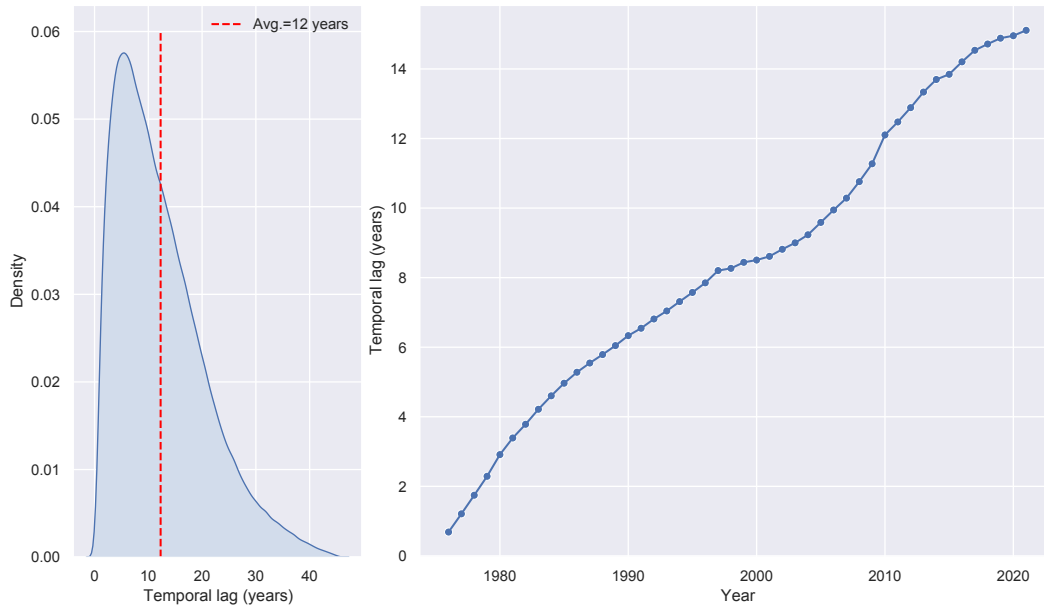


Figure 2.3: Temporal lag. Left: temporal lag distribution. Right: average temporal lag by year.

eled through a smoothing spline of the time lag (in days) between the issued dates of the cited and citing patent. Together with the sender publication date, we account for temporal effects that address the impact of two fundamental temporal components that influence the similarity levels (*Model 1*).

Model 2. [Kuhn \[2010\]](#) argued how legal changes that occurred within the early 2000s amplified the incentives to disclose, increasing the number of citations per issued patent. This idea is also discussed in [Kuhn et al. \[2020\]](#), where a negative effect for the number of backward citations has been observed. However, the inflation of the number of citations during the last period may well be a bias in which a linear effect does not properly take this into account. We address this issue by adding to the model a further covariate fitted through a smooth term: the *backward citation count*. With this explanatory variable, we correct for the increasing number of backward citations done by a given citing party. Furthermore, we consider the type of applicant who is providing the citation, discriminating between organizations (both profit and non-profit) and private owners. The reason for this is straightforward: if there is an inflation in the number of citations, these could be more present within organizations than by private applicants as the former tend to cite more on average. We added three

distinct fixed effects in the form of binary variables: *is the same organization* if the citing and the cited company of the patent coincide, *is citing party an organization* and *is cited party an organization* if either the owner of the cited is an organization (*Model 2*).

*Model 3*. To complete our analysis, we introduced effects related to the IPC for both the citing and the cited party. As we explained, the usage of technological classes for the assessment of patent similarity is disputable [Younge and Kuhn, 2016], but still very important as a common source of knowledge for applicants and examiners. Following Yan and Luo [2017], for each pair of citing/cited patents we computed the *Jaccard index* of individual components in the IPC scheme. What we obtained are five distinct distributions that summarize the technological relatedness between two patents at different levels of the hierarchical classification system.

## 2.3 Results

The naive *Model 0* is in line with previous studies, suggesting that patent similarity is decreasing with time (Figure 2.4). However, this view fails to take into account various confounding effects. In turn, we will focus our attention to the time lag between the citing patent and those patents it cites (*Model 1*), exogenous information associated with citing and cited patent owners, the increasing number of citations per patent (*Model 2*), and finally the IPC classification of the patents (*Model 3*). Empirical results are reported in Table 2.2. For each model specification, Figure 2.4 compares the estimated splines associated with each effect.

Temporal effects. *Model 1* shows that the larger the temporal lag of a citation, the lower the similarity between the citing and the cited patent. Correcting for temporal lag to study patent similarity is important because it corrects for the patent citation network temporal boundary – i.e., only similarity information for patent pairs is available for patents issued after 1976. Related studies Kuhn et al. [2020] do not apply such a correction, with the obvious consequence that their results may be biased and do not reflect the actual changes in similarity patterns. With this correction in place, the results show that patent similarity levels only started to decrease at the beginning of the 1990s after experimenting an initial increasing trend.



Table 2.2: Coefficient estimates for the fixed parametric effects in the three model specifications. Refer to Figure 2.4 for the smoothing splines terms. The asterisk symbol represents the levels of statistical significance, while values in parenthesis are the related standard errors. Model assessment criteria can be found at the bottom of the table. For our purposes, we compared three different criteria: the Akaike Information Criterion (AIC), the Generalized Cross Validation (GCV) criterion and the Deviance explained (R-squared).

	Model 0	Model 1	Model 2	Model 3
Intercept	45.16*** (0.149)	45.16*** (0.015)	45.933*** (0.062)	41.601*** (0.065)
Is the same organization?			8.758*** (0.055)	7.883*** (0.053)
Is sender an organization?			-1.225*** (0.056)	-1.068*** (0.053)
Is receiver an organization?			-1.509*** (0.045)	-1.226*** (0.043)
Jaccard index: section				2.248*** (0.056)
Jaccard index: class				1.901*** (0.064)
Jaccard index: sub-class				2.497*** (0.058)
Jaccard index: main-group				3.639*** (0.059)
Jaccard index: sub-group				4.168*** (0.073)
<b>Smoothing splines</b>				
Publication date	X	X	X	X
Temporal difference (days)		X	X	X
Sender citation count (log)			X	X
<b>Assessment criterions</b>				
AIC	8'229'483	8'202'824	8'138'156	8'058'040
GCV	220.491	214.695	201.243	185.742
R-squared	2.6%	5.17%	11.1%	18%

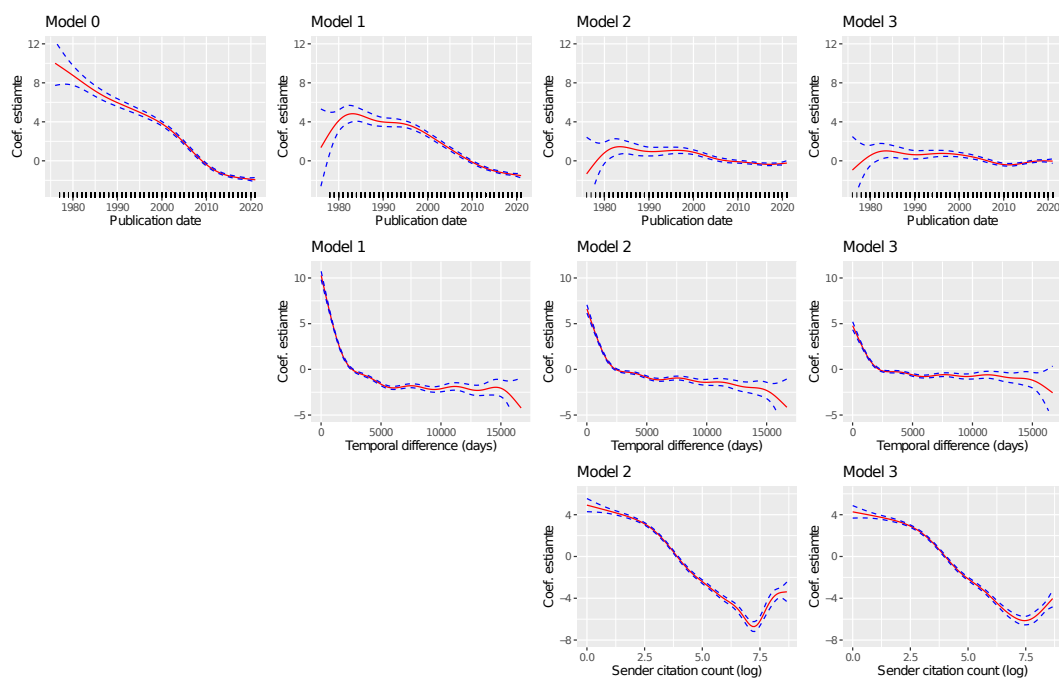


Figure 2.4: Models splines. Smoothing splines estimates for the three fitted GAMs. Columns: Effects fitted per model. Rows: Comparison of the same effects for the distinct models. Assessment criteria can be found at the bottom of Table 2.2.

Citation effects. *Model 2* reveals a downward trend in the log count of the sender citation effect. This suggests that the higher the number of patent citations, the lower the pairwise similarity between citing and cited patents. The effect due to the inflation of citations also mitigates the decline of patent similarity since the 1990s.

*Model 2* shows that citations in which patents are owned by an organization tend to have lower similarity levels. This seems to suggest that organizations tend to include in their patents a series of citations that are loosely related to theirs. On the other side, citations between patents that are owned by the same organization have an important positive impact on similarity. This may be a side-effect caused by the fact that, within organizations, the office responsible for filling patent applications may be the same. As such, words and technical descriptions may be coming from the same group of authors. Moreover, organizations involved in patent-intensive industries will likely tend to cite their previous patents.

**Class effects** The importance of including technology-related information through the Jaccard similarity of IPC is highlighted by the significant increase of deviance

explained and the reduction of the Generalized Cross-Validation (GCV) score. Empirical estimates of *Model 3* show the import effect of lower levels of the patent classification scheme. Jaccard score computed at the IPC *sub-group* level has the strongest impact on patent pairwise similarity. This should be expected given the complex structure of the IPC framework. Patents that share the lowest branch of the classification system will likely include technologies with similar features and consequently similar textual and semantic components.

Finally, when controlling for technology similarity through Jaccard indices, the spline associated with *publication date* in *Model 3* inverts its downward trend in 2011 and progressively increases up to 2021. This is an important result, as it suggests that after accounting for important confounders, the similarity between patents is not decreasing at all, but in fact may recently be slowly increasing.

## 2.4 Discussion

Patent similarity is a complex concept for which only proxy measures exist. Early approaches focused on classification-based measures, whereas more recently text-based similarity measures were introduced. In this paper, we focused our attention on a similarity measure based on SBERT, a direct evolution of the well-known BERT. Nevertheless, the field of sentence embeddings is experiencing a constant development that is raising the bar in terms of model performances. A competitive alternative to SBERT could be the newly released “Definition Sentences” (DefSent, Tsukagoshi et al. [2021]) model, where performances of this have proven to be marginally higher [Tsukagoshi et al., 2022]. However, at present, there are not enough pre-trained models on which we can provide a fair comparison. Furthermore, although training other language models on the patent corpus [Whalen et al., 2020; Hain et al., 2022] could potentially improve the accuracy of the patent similarity score, it may not change our results. Although we have argued that the SBERT-based measure used in this work is a state-of-the-art approach, the main conclusions presented in this manuscript mustn’t depend exclusively on this way of calculating patent similarity. In parallel, we have repeated our analysis with the patent similarity data used by Kuhn et al. [2020], which resulted in the same substantial results as presented in this manuscript.

When studying real-world networks, the issue of network boundaries is almost unavoidable. In the study of patent citation networks, two important boundaries we encountered in our analysis are the fact that the citing patents are US-based patents from after 1976. These temporal and spatial boundaries mean, first, that our concrete conclusions are firmly restricted to the modern US reality. More gen-

eral conclusions would be extrapolations with more or less empirical support. It would be interesting to repeat the study in alternative jurisdictions. Secondly, particularly the temporal boundary has an important effect on the main conclusion presented in this manuscript. The sharp decline of raw patent similarity since 1976 is only reversed to an effective increase in similarity if we accept the hypothesis that patent citation lag is an effective way to account for the temporal boundary in the patent citation network. A longer observation period would make allow us to test this assumption more carefully.

## 2.5 Conclusion

Text-based similarity measures are among the most widely used indicators of patent relatedness. In this paper, we propose an efficient way to compute textual similarity scores using patent abstracts instead of entire technical descriptions. The measure we used shows a similar pattern concerning those of related studies, namely a decrease in text-based similarity starting in 1976 and continuing until recent times. The disadvantage of previous techniques is that they typically involve computationally intensive procedures that do not allow replication. In contrast, the approach of this paper avoids computational bottlenecks by making use of a pre-trained neural network.

Although the changes in the legal framework have had consequences on the citation process, there are other components responsible for the apparent decrease in patent similarity. Simplistic model formulations have obscured the true effect of various factors on the trend in patent similarities. Our empirical analysis focuses on the large body of patent similarities and uses a Generalized Additive Models (GAM) [Hastie and Tibshirani, 1986] to uncover the non-linear relationship between several drivers on the one hand and patent similarities on the other. Explanatory variables in the model specification include both patent-related attributes and network-based measures of technological novelty. Using several model specifications, our analyses show that the observed downward trend of patent similarity scores in the last forty years is the result of several distinct endogenous effects.

The main contribution of this work concerns the analysis of possible factors that are generating the observed downward trend in patent similarity. Using multiple GAMs, we modeled a combination of fixed and non-linear effects. What emerged from the empirical analysis is that the trend in patent similarity is affected by a series of phenomena. The most important one is the effect of time lag between citing and cited patents. This can be seen from the transition between *Model 0*

and *Model 1*, in which we account for the time difference between the publication dates of sender and receiver. With this effect in place, the curve changes its shape and shows an interesting increase up to the mid-80s. Finally, the introduction of citation and class effects further corrects the curve by considering two other well-known phenomena: the tendency to increase the number of citations per patent application and the increase in patent class assignments. Thanks to these adjustments, we conclude that the levels of similarity have not been constantly decreasing since 1976, but instead, they show a more oscillating behavior.

# Chapter 3

## Relational Event Modelling

The following chapter was published as:

Bianchi, F., Filippi-Mazzola, E., Wit, E. C., and Lomi, A., (2024). Relational Event Modeling. *Annual Review of Statistics and Its Application*, 11(1), 297–319.

### 3.1 Introduction

Statistical models for social and other networks are receiving increased attention not only in specialized field journals such as *Network Science* or *Social Networks*, but also in prominent interdisciplinary science journals such as *Science* [Borgatti et al., 2009; Butts, 2009], *PNAS* [Stadtfeld et al., 2019], and *Science Advances* [Elmer and Stadtfeld, 2020].

Attention to statistical models for networks is on the rise also in well-established generalist statistics journals such as, for example, the *Journal of the American Statistical Association* [Hunter et al., 2008], the *Journal of Applied Statistics* [Snijders et al., 2010a], *Statistical Science* [Schweinberger et al., 2020], and the *Journals of the Royal Statistical Society (series A, B, and C)* [Fienberg, 2012; Krivitsky and Handcock, 2014; Gile and Handcock, 2017; Vinciotti and Wit, 2017; Koskinen and Snijders, 2023]. The *Annual Review of Statistics and Its Applications* itself has recently demonstrated considerable interest in models for social networks by publishing comprehensive and up-to-date reviews on two popular classes of statistical models: *Exponential Random Graph Models* (ERGMs) [Amati et al., 2018] and *Stochastic Actor-Oriented Models* (SAOMs) [Snijders, 2017].

Since the publication of these reviews, the increasing availability of time-stamped resulting from innovation in data production, collection, storage, and retrieval technologies has shown that network data samples collected at fixed time intervals are likely to miss fundamental differences in the time scales over which

relational processes unfold [Golder et al., 2007]. Computer-mediated communication [Lerner and Lomi, 2023], sociometric badges [Wu et al., 2008; Stehlé et al., 2011; Elmer and Stadtfeld, 2020], electronic trading platforms [Zappa and Vu, 2021], on-line interaction logs [Tonellato et al., 2023], and video recordings [Pallotti et al., 2022], are just some of the new data-generating technologies capable of producing large quantities of relational event data connecting sender and receiver units.

During the same period, studies based on event-oriented designs have become also increasingly common. While the empirical opportunities offered by relational event data have long been acknowledged by students of social networks [Freeman et al., 1987; Marsden, 1990; Borgatti et al., 2009], statistical models affording a degree of temporal resolution consistent with the frequency of observed social interaction [Butts, 2009] have become available only during the last fifteen years [Butts, 2008].

Time scales vary considerably based on interaction settings. High-frequency transactions in financial markets [Bianchi and Lomi, 2022] occur within seconds or fractions of seconds, while communication in emergencies [Butts, 2008; Renshaw et al., 2023] may take several minutes. Email exchange [Perry and Wolfe, 2013] can extend over hours, whereas interaction generated by more complex forms of social coordination among corporate actors may become observable only over days or even weeks [Amati et al., 2019]. In these various cases, aggregating time-stamped relational event data into network ties defined over conventional, or convenient, periods, is unlikely to afford high-fidelity representations of the underlying interaction processes [Tuma and Hannan, 1984].

Perhaps the main motivation that inspired the development of the *Relational Event Model* (REM) proposed by Butts [2008], was to provide a general analytical framework where *sequential ordering* and *timing* replace *concurrency* and *temporal aggregation* of network edges as the “dominant concepts of phenomenal concern” [Butts, 2008, p.192] in the analysis of social interaction. This involves deriving, specifying, and estimating statistical models capable of assimilating and analyzing complex relational data without altering — through time aggregation — the natural time structure, and sequential ordering of observed social interaction. Conversation [Gibson, 2005], communication [Pilny et al., 2017], market exchange [Lomi and Bianchi, 2021], and other, more complex, forms of social coordination [Lerner and Lomi, 2020a] can be understood only concerning the timing and sequential order of the relational events of interest [Abbott, 1992], which contain important information that is typically lost when time-stamped events are aggregated into binary network “ties” [Pallotti et al., 2022].

Since its introduction, the REM has been significantly refined and adapted to

an ever-increasing diversity and sophistication of emerging empirical problems [Butts et al., 2023]. This review piece provides an opportunity to position relational event modeling in the broader context of statistical models for network science and assess the current state of the field, incorporating a broad review of contemporary methodological, computational, and inferential developments in this class of statistical models for directed social interaction.

The paper is organized into eight main sections. Section 3.2 traces the intellectual origins and historical context of REMs. Section 3.3 identifies the observation plans and empirical research design elements typically associated with event-oriented studies of network dynamics. Section 3.4 surveys the available classes of REMs developed, at least in part, in response to specific problems with no satisfactory modeling solutions. Section 3.5 is dedicated to issues of empirical model specification and estimation. Section 3.6 examines the broad area of applications where REMs seem to have found fertile ground for development. Section 3.7 reviews the main challenges and open issues that orient current research. Section 3.8 concludes with a summary and a discussion of the main promises of relational event modeling.

## 3.2 Historical Context of Relational Event Models

Contemporary statistical models for social and other kind of networks rely heavily on the formalism of graph theory [Butts, 2009] — an inheritance left, in part, by earlier network models developed in sociometry [Moreno, 1934; Jennings, 1948], and within the structural tradition of social and cultural anthropology [White, 1963; Lévi-Strauss, 1971; Hage, 1979; Barnes and Harary, 1983; Hage and Harary, 1984]. Concepts such as those of “degree”, “path distance”, “reciprocity”, and “transitive closure” that are central in contemporary statistical models for networks [Snijders, 2001; Amati et al., 2018] are firmly rooted in the mathematical representation of networks as graphs [Barabási, 2013].

The graph-theoretic formalism that inspired early Markov random graph models [Holland and Leinhardt, 1977, 1981; Frank and Strauss, 1986] may be considered to be the intellectual antecedent of contemporary statistical models for social networks [Wasserman and Pattison, 1996; Snijders, 2001; Snijders et al., 2006]. It is only in relatively recent times that the limitations in representing networks of social relations as graphs have started to become apparent. REMs entail an alternative understanding of social relations as emergent from sequences of relational events connecting a sender behavioral unit to one or more receivers [Butts, 2008, 2009; Perry and Wolfe, 2013; Lerner and Lomi, 2023].



REMs provide a framework for analyzing and making inferences about the relational processes and dynamics in complex social systems. They are designed to capture the patterns of dependence in the occurrence and timing of relational events, such as communications, transactions, or social interactions, within a network.

Formally, REMs are rooted in event history models [see, e.g., [Keiding, 2014](#), for a review], expressing the hazards of an event to occur, as a function of the history of previous events, as well as potentially additional nodal and relational attributes [[Butts, 2008](#)]. REMs allow to study how events unfold over time, how they are influenced by various attributes and how they, in turn, affect the structure and evolution of the network.

The exact hazards of a REM may follow different functional forms and specifications. In many instances, the hazard is assumed to be piecewise constant, resulting in waiting times that are conditionally exponentially distributed. However, the precise definition of hazards will depend on the research questions and the empirical context, and these choices have implications for computational complexity, model fit, and the interpretation of the model [[Butts and Marcum, 2017](#); [Schaefer and Marcum, 2017](#); [Stadtfeld et al., 2017](#)].

Two other modeling frameworks are appropriate for the analysis of longitudinal networks, the temporal ERGM and the SAOM. The formulation of the ERGM [[Robins et al., 2007](#)] may be understood in terms of the conditional probability of a network edge, given the relationships among all other network edges [[Wasserman and Pattison, 1996](#); [Snijders et al., 2006](#)]. In the ERGM the edges connecting pairs of senders and receivers are assumed to be interdependent, and such interdependencies can be captured by local configurations of network ties such as triangles or star-shaped structures [[Robins et al., 2007](#)]. More specifically, the existence of a tie in a social network is conditionally independent of ties that are far distant in space [[Frank and Strauss, 1986](#)]. Temporal extensions of the ERGM [[Hanneke et al., 2010](#); [Krivitsky and Handcock, 2014](#)] have been introduced to model network dynamics over time.

The SAOM [[Snijders, 1996, 2001, 2005, 2017](#)], is a probability model evaluating change in network ties (relational states) observed in adjacent time periods. While the SAOM takes into account relational states observed at multiple time points, it operates under the assumption that ties change continuously over time through a series of micro-steps occurring between observations. Each micro-step involves at most one alteration [[Holland and Leinhardt, 1977](#)] in the network with social actors choosing what ties to alter based on a multinomial choice probability model [[McFadden, 1973](#)] accounting for the local configurations in which potential ties are embedded.

### 3.3 Relational Data and Study Design

The relational event modeling framework has been introduced to analyze the dynamics underlying social networks, with a focus on understanding the complex interactions and dependencies between behavioral units over time. REMs are flexible because they integrate temporal, structural, and attributional aspects of the data within a single framework. This flexibility sustains the specification of very general, yet detailed models of network evolution [Brandes et al., 2009].

#### 3.3.1 Network Structure

The building block of the REM is the relational event, defined as an interaction initiated by a sender unit and directed toward one or more targets [Butts, 2008]. Examples include hospitals sharing patients [Vu et al., 2017; Amati et al., 2019], credit institutions exchanging financial assets [Lomi and Bianchi, 2021; Zappa and Vu, 2021; Bianchi and Lomi, 2022], students telephoning each other [Stadtfeld and Block, 2017], people interacting face-to-face [Elmer and Stadtfeld, 2020] and in social groups [Hoffman et al., 2020], or employees exchanging emails [Quintane et al., 2013; Perry and Wolfe, 2013].

REMs can also model relations between units belonging to different classes in the context of more general bipartite processes. Examples include non-native species invading new spatial niches [Juozaitienė et al., 2022], editors modifying Wikipedia pages [Lerner and Lomi, 2017], computer developers fixing software problems [Tonellato et al., 2023], and political actors supporting or rejecting proposed legislation [Brandenberger, 2018b; Haunss and Hollway, 2022].

Although relational events are typically defined between a single sender and a single receiver, generalizations are possible where a sender may reach multiple receivers simultaneously. This situation is common in technology-mediated communication [Lerner and Lomi, 2023], citation networks [Lerner and Hâncean, 2023; Filippi-Mazzola and Wit, 2024b], and social gatherings attended by many participants [Lerner et al., 2021].

#### 3.3.2 Sampling and Recording

Event network studies typically look at social units interacting within a bounded environment. Examples of geographical areas consist of student houses [Stadtfeld and Block, 2017] or hospitals within geographical regions [Vu et al., 2017; Zachrison et al., 2022], whereas examples of temporal boundaries involve interaction between emergency teams during different phases of a crisis [Butts, 2008],

or hospitals transferring patients in specific days of the week [Amati et al., 2019]. Relational events taking place outside the main observation domain are usually not recorded, so the social networks reconstructed from streams of relational events are often incomplete. Unlike studies of statistically independent observations, network boundary specification might affect the validity, robustness, and replicability of the results [Laumann et al., 1989].

### 3.4 Specifications of Relational Event Models

Table 3.1: Notation

Notation	Meaning
$(t, s, r)$	A relational event in which sender $s$ interacts with receiver $r$ at time $t$
$\lambda_{sr}(t)$	Rate/hazard at which sender $s$ contacts receiver $r$ at time $t$
$\mathcal{H}_t$	Filtration of the process, containing all information about relational events until time $t$
$L, L_p$	Likelihood and partial likelihood
$\mathcal{R}(t)$	Risk set at time $t$ , consisting of all relational events that can happen at that time
$\tilde{\mathcal{R}}(t)$	Sampled risk set at time $t$ , consisting of the event $(s, r)$ at time $t$ and a set of sampled non-events
$x_{sr}(t)$	Dyadic covariate(s) with corresponding fixed effect(s) $\beta$
$z_{sr}(t)$	Dyadic covariate(s) with corresponding random effect(s) $\gamma$
$a$	Alter, i.e., an individual different from sender or receiver

The units of analysis in the REM are the edges connecting individual pairs of senders and receivers. Those edges are typically stored in tuples  $(t, s, r)$ , where  $s$  is the sender,  $r$  is the receiver, and  $t$  is the time of the relational event connecting  $s$  to  $r$ . At its core, the REM is defined as a point process for directed pairwise interactions that, in turn, are modeled through their rate function  $\lambda$ . The model assumes that  $\lambda$  may depend upon sender, receiver, past event history, and/or exogenous covariates.

#### 3.4.1 Types of Relational Event Models

We consider a fixed time interval  $[0, T]$ , with  $0 < T < \infty$ , in which *events* occur. Events are defined as time-stamped interactions between senders and re-

ceivers. Both the set of senders  $\mathbf{S}$  and receivers  $\mathbf{R}$  are assumed to be finite. For *one-mode networks* the set of senders and receivers overlap,  $\mathbf{S} = \mathbf{R}$ , whereas for *two-mode* or *bipartite networks* they are distinct. The relational event process is a marked point process [Daley and Vere-Jones, 2003; Cressie, 2015] based on more general approaches [Borgan et al., 1995; Aalen et al., 2008] for event history sequences  $\{(t_i, s_i, r_i) : i \geq 1, s_i \in \mathbf{S}, r_i \in \mathbf{R}\}$ , and defined on a probability space  $(\Omega, \mathcal{F}, P)$  adapted to the filtration  $\mathcal{H}_t$ , consisting of the history of process [Andersen et al., 1993]. In principle, the marks  $(s, r)$  can be individuals or sets of senders and receivers.

Associated with this marked point process, we define a multivariate counting process  $N$ , whose components  $N_{sr}$  record the number of directed interactions between  $s$  and  $r$ ,

$$N_{sr}(t) = \sum_{i \geq 1} \mathbb{1}_{\{t_i \leq t; s_i = s; r_i = r\}}.$$

According to the Doob–Meyer decomposition theorem [Meyer, 1962], there exists a predictable process  $\Lambda_{sr}$  and a residual martingale process  $M_{sr}$ , such that the counting process can be written as

$$N_{sr}(t) = \Lambda_{sr}(t) + M_{sr}(t).$$

The aim of the REM is to describe the structure of the predictable cumulative hazard process  $\Lambda_{sr}$ . By assuming that the counting process is an inhomogeneous Poisson process, we can write the cumulative hazard as

$$\Lambda_{sr}(t) = \int_0^t \lambda_{sr}(\tau) d\tau,$$

where  $\lambda_{sr}$  is the hazard function of the relational event  $(s, r)$ . The general REM is defined as

$$\lambda_{sr}(t) = \mathbb{1}_{\{(s,r) \in \mathcal{R}(t)\}} \lambda_0(t) \psi\{\beta(t), x_{sr}(t)\}, \quad (3.4.1)$$

where  $\lambda_{sr}(t)$  is only non-zero if the event  $(s, r)$  is contained in the *risk set*  $\mathcal{R}(t)$  of possible events at time  $t$ ,  $\lambda_0(t)$  is the baseline hazard function unrelated to  $(s, r)$ ,  $\psi$  is a *relative risk function* as defined in Thomas [1981],  $x_{sr}(t)$  is the  $\mathcal{H}_t$  measurable set of endogenous and exogenous (possibly) time-varying variables, and  $\beta(t)$  are the effect sizes. The function  $\psi$  is positive and, for identifiability, normalized by setting  $\psi(\beta, 0) = 1$ .

The relational event literature has focused on the exponential risk function  $\psi(\beta, x) = \exp(\beta^\top x)$ . Under the conditional independence framework [Besag, 1974, 1975], network statistics are, by construction, dependent through space and time. Ac-

cordingly, the propensity of pairs of senders and receivers to mutually connect depends on which relational events they have sent or received in the past, their order in time, if not their exact timing, and on exogenous factors like nodal or network attributes.

### The Original Relational Event Model

Butts [2008] introduced the REM as a piece-wise homogeneous Poisson process, whereby the rate function was specified as

$$\lambda_{sr}(t) = \lambda_0 \exp\{\beta^\top x_{sr}(t)\} \mathbb{1}_{\{(s,r) \in \mathcal{R}(t)\}}, \quad (3.4.2)$$

where  $x_{sr}(t)$  is the collection of sufficient network statistics associated with the parameter vector  $\beta \in \mathbb{R}^p$ . The covariates in  $x_{sr}(t)$  can depend on nodal characteristics, such as the age of  $r$  or the age difference between the pair  $(s, r)$ , and also on the history of the process. For instance, the covariates  $x_{sr}(t)$  can include a count of the number of past events between  $(s, r)$  before time  $t$ .

In the original formulation, the baseline rate of the interaction process  $\lambda_0$  is assumed constant. This means that the waiting times between events is exponentially distributed, and once an event takes place, the context of the action is changed, and all the possible relational event waiting times are restarted.

### Weighted and Signed Networks

One of the first extensions of the original REM involved geopolitical conflicts and cooperation [Brandes et al., 2009]. The authors highlighted that each interaction event between countries, international organizations, or ethnic groups could be given either a positive or negative weight according to the degree of cooperation or hostility between countries.

In this extended relational event definition  $(t, s, r, w)$ , the interaction weight  $w$  is modeled alongside the event  $(t, s, r)$  as  $p\{(t, s, r, w) \mid \mathcal{H}_{t-}\} = p\{(t, s, r) \mid \mathcal{H}_{t-}\} \times p\{w \mid (t, s, r), \mathcal{H}_{t-}\}$ . The second term models the interaction weights depending on which countries are interacting and how they interacted in the past, whereas the first term is the usual REM likelihood, which is also allowed to depend on past interaction weights included in  $\mathcal{H}_{t-}$ .

Subsequent versions of the REM have accommodated weighted network statistics [Amati et al., 2019; Bianchi and Lomi, 2022] associating a decay function  $f(t, t_i, \alpha)$  to each past event  $(s_i, r_i)$  where  $t$  is the current time, and  $\alpha$  a decay parameter. In its simplest formulation, the temporal relevance of an event is

assumed to decrease according to a power law,  $f(t, t_i, \alpha) = (t - t_i)^{-\alpha}$ , though other specifications may be used instead [Lerner and Lomi, 2020b].

### Multicast and Polyadic Interaction

Polyadic (or “multicast”) interaction occurs when a single event links an individual sender to multiple receivers simultaneously. The possibility of one-to-many interaction was explicitly recognized by DuBois et al. [2013b] in their model of teacher-students interaction in a classroom context. In the same year, Perry and Wolfe [2013] provided a more general formulation specification of the intensity function for a sender  $s \in \mathbf{S}$  addressing a set of receivers  $R = \{r_1, \dots, r_m\} \subset \mathbf{R}$ , taking

$$\lambda_{sR}(t) = \lambda_{0s}(t; |R|) \exp \left\{ \beta_0^\top \sum_{r \in \mathcal{R}_s(t)} x_{sr}(t) \right\} \prod_{r \in R} \mathbb{1}_{\{r \in \mathcal{R}_s(t)\}}. \quad (3.4.3)$$

The event rate function involves two types of stratification, discussed in greater detail in Sections 3.4.1 and 3.4.2, which explains the sender-specific definition of the risk set  $\mathcal{R}_s(t)$  and baseline hazard, which also depend on the size of the receiver set. The network statistics are defined as the sums of the individual dyadic statistics  $x_{sR}(t) = \sum_{r \in R} x_{sr}(t)$ .

Lerner and Lomi [2023] introduced the *Relational Hyperevent Model* (RHEM) by generalizing the definition of multicast network statistic. A hyperedge covariate  $x_{sR}$  is a function of the sender and the *entire set* of receivers that cannot, necessarily, be decomposed into the sum of dyadic covariates. An example of hyperedge covariate is *inertia*, the tendency toward repeating past relational events, defined as

$$x_{sR}^{\text{inertia}}(t) = \sum_{t_i < t} f(t, t_i, \alpha) \mathbb{1}_{\{s_i=s, R_i=R\}}.$$

In email exchange, for instance, the existence of a mailing list may result in a moderator communicating with exactly the same group of individuals. Instead, *unordered repetition* captures interaction within a stable set of actors with turn-taking among the senders [Gibson, 2005]. Replacing dyadic covariates with their hyperedge counterparts not only provides a richer collection of network effects but also improves model fit [Lerner and Lomi, 2023].

Alternative approaches for modeling polyadic interactions have been proposed. Kim et al. [2018] introduced the *Hyperedge Event Model*, assuming that dyadic intensities stochastically determine the sender of the next event and its receiver set. Mulder and Hoff [2021] introduced a latent variable model whereby all

potential receivers are assigned to the receiver set on the basis of a suitability score depending on the sender and receiver-specific characteristics.

### Separable Intensity Functions

Stadtfeld et al. [2017] developed the *Dynamic Network Actor Model* (DyNAM), an actor-oriented model for relational event data built upon the same paradigm introduced by Snijders [1996, 2017]. The distinctive feature of the DyNAM is that it explains social interaction in terms of “individuals’ preferences, available interaction opportunities, and individuals’ perception of the social network they are embedded in” [Stadtfeld and Block, 2017, p.318]. Accordingly, the inferential framework of the DyNAM consists of modeling a composite process made of the sender’s decision to initiate a relational event at a certain point in time, and the sender’s decision to address a specific receiver.

The DyNAM decomposes event rates into two, sender-centred, components, i.e.,

$$\lambda_{sr}(t) = \underbrace{\lambda_s(t)}_{\text{select sender } s} \times \underbrace{p(r|s)}_{\substack{\text{sender } s \\ \text{addresses receiver } r}} \quad (3.4.4)$$

Similar to Snijders [2005], senders’ event rates  $\lambda_s$  are modeled through an exponential link function evaluating nodal attributes at individual and network levels,

$$\lambda_s(t) = \lambda_0 \exp(\theta^\top x_s) \mathbb{1}_{\{s \in \mathcal{R}(t)\}}. \quad (3.4.5)$$

Following McFadden [1973], receiver choice is modeled via a multinomial distribution, i.e.,

$$p(r|s, \mathcal{H}_t) = \frac{\exp\{\beta^\top x_{sr}(t)\}}{\sum_{r' \in \mathcal{R}_s(t)} \exp\{\beta^\top x_{sr'}(t)\}}. \quad (3.4.6)$$

The timing distribution does not explicitly depend on the receiver characteristics and this typically sacrifices model fit over an actor-oriented interpretation of the model parameters.

The DyNAM has recently been extended with the incorporation of the DyNAM-i [Hoffman et al., 2020], which explains sequences of joining and leaving events in the context of group-based interactions. This extension captures the specific nature of group conversations and interactions that typically occur in cliques; and the need to align network modeling strategies with the increasing use of sensor technologies, such as Bluetooth or RFID badges, to detect collective interaction. Vu et al. [2017] exploited the separability of intensity functions by decomposing the stream of relational events into event times and event destinations. The



separable sender intensity and receiver choice model adds more flexibility to the relational event framework allowing for the separation between senders and receiver effects, and not only between event weights and dyads as in Brandes et al. [2009].

### 3.4.2 Network Covariates

Given the general expression of the REM hazard in Eq. (3.4.1), the drivers of the relational process  $x_{st}(t)$  describe known,  $\mathcal{H}_{t-}$  measurable, quantities quantifying endogenous or exogenous statistics of the sender, the receiver or both.

#### Endogenous VS Exogenous Covariates

In statistical models for networks [e.g., Snijders et al., 2010b], covariates are endogenous to the extent that they depend on past interaction. Covariates are exogenous when they depend on characteristic of single nodes (monadic covariates) or pairs of nodes (dyadic covariates). One example of endogenous covariate is reciprocity, while gender and geographical distance are exogenous covariates, representing monadic and dyadic characteristics, respectively. A basic selection of endogenous and exogenous covariates is displayed in Table 3.2. Other ad-hoc specifications, such as those based on exchange sequences in conversational analysis, are explained in Butts [2008].

An additional consideration refers to the *hierarchy principle*, whereby lower-order interaction terms should always be included in the presence of higher-order interaction terms [Pattison and Robins, 2002; Wang et al., 2013]. In the REM, for example, failing to account for heterogeneity of the senders and receivers may result in incorrect detection of triadic effects [Juozaitienė and Wit, 2022a].

#### Heterogeneity

There are two fundamental types of heterogeneity in event networks. Endogenous heterogeneity, or emergence, refers to the inherent stochasticity of the process itself combined with the dependence of future interactions on current ones. An example is *virality*, whereby, for instance, a paper gets cited because it was cited many times before. In the REM, endogenous heterogeneity can be captured by endogenous covariates.

The second type of heterogeneity, extrinsic variation, is either observed, such as the prestige of the institutions of the authors of a paper, or latent, such as the quality of the work it represents. Latent extrinsic heterogeneity in the REM



Table 3.2: Basic menu of network covariates defined in the REM

Network Covariate	Mechanism	Formula
Out-degree		$\sum_{a \neq s} N_{sa}(t^-)$
Out-intensity		$\frac{1}{\text{out-degree}_s(t)} \sum_{a \neq s} \sum_{i: s_i=s, r_i=a} f(t, t_i, \alpha)$
In-degree		$\sum_{a \neq s} N_{as}(t^-)$
In-intensity		$\frac{1}{\text{in-degree}_s(t)} \sum_{a \neq s} \sum_{i: s_i=a, r_i=s} f(t, t_i, \alpha)$
Repetition		$N_{sr}(t^-)$
Reciprocation		$N_{rs}(t^-)$
Transitive closure		$\sum_{a \neq s, r} N_{sa}(t^-) N_{ra}(t^-)$
Cyclic closure		$\sum_{a \neq s, r} N_{as}(t^-) N_{ra}(t^-)$
Sending balance		$\sum_{a \neq s, r} N_{sa}(t^-) N_{ra}(t^-)$
Receiving balance		$\sum_{a \neq s, r} N_{as}(t^-) N_{ra}(t^-)$
Node attribute		$x_{sr}(t)$
Node matching		$\text{dist} \{x_s(t) = x_r(t)\}$

Notes. (Out/in)-(degree/intensity) statistics can refer to both senders and receivers. Node  $a$  is a third (alter) trading counterpart of the sender-receiver pair  $(s, r)$ . The term  $N_{sr}(t^-)$  is the number of relational events flowing from sender  $s$  to receiver  $r$  right before time  $t$ , while  $f(t, t_i, \alpha) = (t - t_i)^{-\alpha}$  is the decay function accounting for the temporal relevance of previous relational events. In depicting network statistics, solid line arrows ( $\rightarrow$ ) refer to past relational events, while dashed arrows ( $--\rightarrow$ ) indicate current relational events.

can be expressed by means of random effects. Juozaitienė and Wit [2022a] and Uzaheta et al. [2023] proposed mixed effect extensions of the REM, i.e.,

$$\lambda_{sr}(t) = \mathbb{1}_{\{(s,r) \in \mathcal{R}(t)\}} \lambda_0(t) \exp \{ \beta^\top x_{sr}(t) + \gamma^\top z_{sr}(t) \},$$

with dyadic covariates  $z_{rs}(t)$  and  $\gamma \sim N(0, \Sigma)$  the random effects. Estimation of the random effects variance can be done via Expectation Maximization [Dempster et al., 1977] or Laplace approximations of the likelihood [Pinheiro and Bates, 2006], Hamiltonian Monte Carlo Methods [Uzaheta et al., 2023], or in certain cases via a penalized zero order spline approach [Wood, 2017].

DuBois et al. [2013a] suggested a model that accounted for sender and receiver heterogeneity by means of a stochastic block structure. More recently, in an analysis of a communication network Juozaitienė and Wit [2022a] showed that incorporating random effects for both the sender and receiver enhances the model fit compared to model specifications that solely rely on endogenous degree-based statistics. Therefore, the inherent differences between individuals in the network drives part of the heterogeneity in the interactions.

### Stratification

Conceptually, stratification can be introduced either to model event streams in multiplex networks or to account for heterogeneity by specifying different baseline intensity functions for individual sets of dyads. Perry and Wolfe [2013]; Bianchi and Lomi [2022], for example, use sender-based stratification, effectively allowing each sender to have its own individual baseline hazard, i.e.,

$$\lambda_{sr}(t) = \lambda_{0s}(t) \exp \{ \beta_0^\top x_{sr}(t) \} \mathbb{1}_{\{(s,r) \in \mathcal{R}(t)\}}.$$

Receiver-based stratification is defined in a similar fashion, and usually occurs when there is heterogeneity in those nodes that are repeatedly targeted as receivers. In citation networks, for example, groundbreaking articles or patents have very distinct, individual citation profiles, which makes a receiver-based baseline hazards an attractive option, given that they not have to be estimated individually [Filippi-Mazzola and Wit, 2024b].

Juozaitienė and Wit [2022b] proposed a stratified version of the REM, in which distinct baseline hazards are associated with distinct families of temporal network effects, such as reciprocity in its direct and generalized forms. Subsequent baseline hazard estimation reveals the tendency of some endogenous covariates to have very specific temporal effect-profiles.

## 3.5 Estimation and Computation

The fundamental information about a sequence of relational events  $\{(t_i, s_i, r_i) : i = 1, \dots, n\}$  is contained in its likelihood function. For a REM, this function can be expressed as the product of the conditional generalized exponential event time densities and their associated multinomial relational event probabilities, i.e.,

$$L(\beta) = \prod_{i=1}^n \sum_{(s,r) \in \mathcal{R}(t_i)} \lambda_{sr}(t_i) \exp \left\{ - \sum_{(s,r) \in \mathcal{R}(t_i)} \int_{t_{i-1}}^{t_i} \lambda_{sr}(\tau) d\tau \right\} \frac{\lambda_{s_i r_i}(t_i)}{\sum_{(s,r) \in \mathcal{R}(t_i)} \lambda_{sr}(t_i)}, \quad (3.5.1)$$

where  $\lambda$  is a function of  $\beta$  and  $\mathcal{R}(t)$  is the risk set, i.e., the set of all possible events that could have occurred at time  $t$ .

Estimating the parameters of REMs by maximizing the full likelihood poses several challenges. The likelihood function is indeed a complex object that involves explicit integration across the unknown hazard function and sums over large risk sets. In this section, we will explore computational alternatives proposed to overcome the complexity of the full likelihood approach.

### 3.5.1 Partial Likelihood Estimation

The proportional hazard model [Cox and Oakes, 1984] offers an attractive alternative to fully parametric models due to its absence of distributional assumptions regarding activity rates, which are then treated as nuisance parameters. It offers an effective simplification of the full REM likelihood through the application of the partial likelihood  $L_p$  [Cox, 1975] to counting processes [Andersen and Gill, 1982] on network edges, which only involves multinomial event probabilities, i.e.,

$$L_p(\beta) = \prod_{i=1}^n \left( \frac{\exp \{ \beta^\top \mathbf{x}_{s_i r_i}(t_i) \}}{\sum_{(s,r) \in \mathcal{R}(t_i)} \exp \{ \beta^\top \mathbf{x}_{sr}(t_i) \}} \right). \quad (3.5.2)$$

This expression eliminates the unknown baseline hazard, resulting in a more adaptive representation of the underlying network dynamics, while being able to estimate the parameters in a straightforward way by maximizing  $L_p(\beta)$ . An example is given in Figure 3.1.

As Butts [2008] noted, the partial likelihood corresponds to the full likelihood when only the event orderings are known, but not the exact timings. However, the partial likelihood approach faces a limitation in large networks, as the risk set in its denominator tends to expand quadratically with the number of nodes.

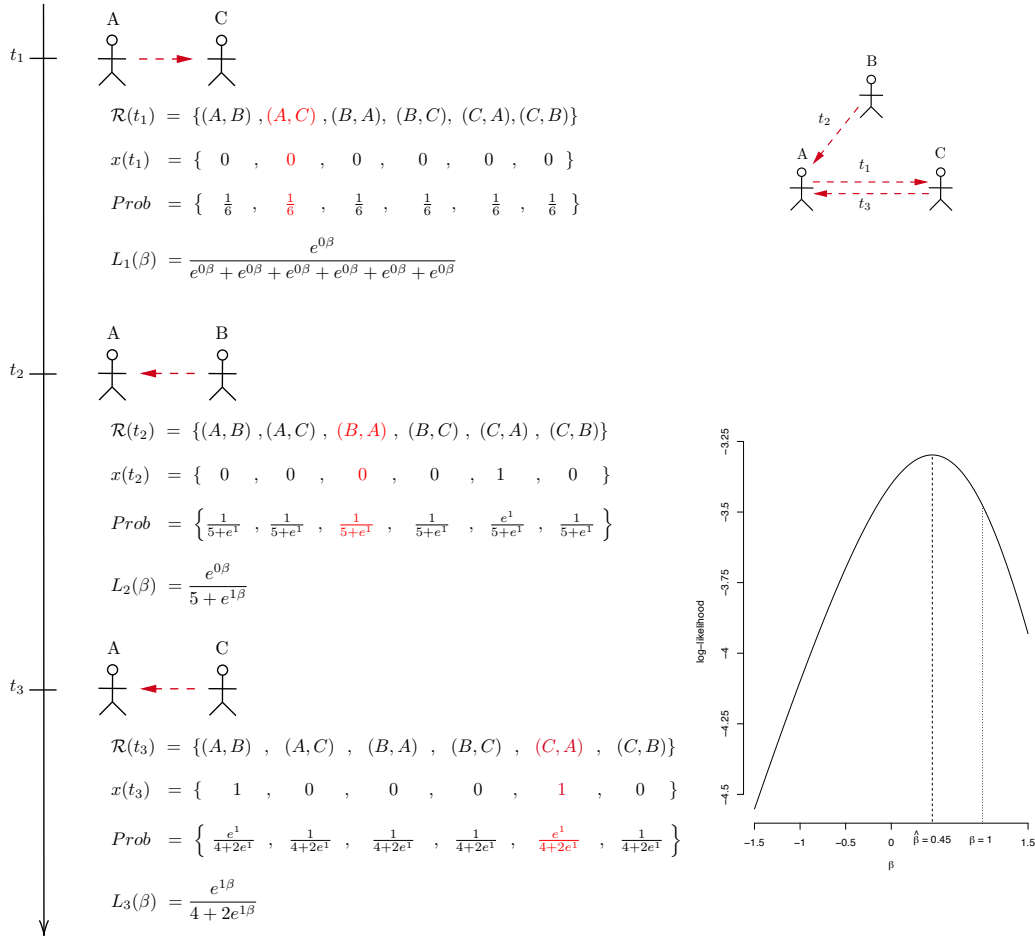


Figure 3.1: Example of a relational event process among three subjects  $\{A, B, C\}$  based on the hazard function  $\lambda_{sr}(t_i) = \lambda_0(t_i) \exp\{\beta x(t_i)\}$ , with  $\beta = 1$ , whereby the event rate is influenced by the reciprocation statistic  $x(t_i)$ . The left-hand side of the figure outlines the unfolding of the sequence of relational events according to the true probabilities  $Prob$ , mirrored in the graphical representation in the upper right. At each time point, the risk set  $\mathcal{R}(t)$  consists of all six possible interactions between pairs of individuals. At time  $t_1$ , an interaction from A to C is observed, while all reciprocation covariates are zero, as no interaction has occurred yet. The probability of each event is determined using Eq. 3.5.2, with partial likelihood  $L_1(\beta)$ . At time  $t_2$ , an interaction from B to A occurs. This event has partial likelihood  $L_2(\beta)$ , with one reciprocal non-event in the denominator. At  $t_3$ , the reciprocal interaction from C to A is observed, with partial likelihood  $L_3(\beta)$ . The overall partial likelihood  $L_p(\beta) = \prod_{i=1}^3 L_i(\beta)$  is maximized at  $\hat{\beta} = 0.45$ , as shown in the lower-right side of the figure, close to the actual effect size  $\beta = 1$ .

### Risk Set Sampling

The primary constraints in modeling the partial likelihood from Eq. 3.5.2 are found in the risk set size. The idea of adopting subsampling strategies to approximate the partial likelihood was noted by Butts [2008]. Vu et al. [2015] initially introduced a, more efficient, nested-case control sampling strategy [Borgan and Keogh, 2015] to mitigate the computational complexity involved in estimating the partial likelihood.

Nested case-control sampling consists of sampling from the current risk set  $\mathcal{R}(t)$  according to some probability  $\pi\{\cdot | \mathcal{R}(t)\}$  a set of non-events, or controls, for each event, or case. The sampled non-events together with the events are called the sampled risk set  $\tilde{\mathcal{R}}(t)$ . Borgan et al. [1995] show that the sampled partial likelihood  $\tilde{L}_p$ , accounting for the sampling probabilities is a valid likelihood. When this probability is assumed to be random, i.e.,  $\pi\{\cdot | \mathcal{R}(t)\} = 1/|\mathcal{R}(t) - 1|$ ,  $\tilde{L}_p(\beta)$  reduces to the simplified form, i.e.,

$$\begin{aligned} \tilde{L}_p(\beta) &= \prod_{i=1}^n \left( \frac{\pi((s_i, r_i) | \mathcal{R}(t_i)) \exp\{\beta^\top x_{s_i r_i}(t_i)\}}{\sum_{(s,r) \in \tilde{\mathcal{R}}(t)} \pi((s, r) | \mathcal{R}(t_i)) \exp\{\beta^\top x_{sr}(t_i)\}} \right) \\ &= \prod_{i=1}^n \left( \frac{\exp\{\beta^\top x_{s_i r_i}(t_i)\}}{\sum_{(s,r) \in \tilde{\mathcal{R}}(t)} \exp\{\beta^\top x_{sr}(t_i)\}} \right), \end{aligned}$$

where  $\tilde{\mathcal{R}}(t)$  is the sampled risk set.

Lerner and Lomi [2020b] employed nested case-control sampling to empirically showcase the efficiency of estimates on large networks, even when a limited number of non-events is sampled.

### Computational Aspects of Stratified Relational Event Models

Perry and Wolfe [2013] proposed the first REM that introduced sender stratification to expedite calculations and combined it with a customized method for maximizing the log partial likelihood. Vu et al. [2015] proposed a flexible stratification method allowing for data structures with many types of nodes and events, then showing the scalability of their approach to large data sets. Combining the two approaches, Bianchi and Lomi [2022] proposed a sender-stratified REM for high-frequency data, using nested case-control sampling to update the risk set at each new event. This model specification has been tested in empirical applications using millions of financial transactions. Filippi-Mazzola and Wit [2024b] proposed a receiver-stratified REM for the analysis of millions of patent citations, in which the hazard is modeled via smooth functions of the covariates using a

spline approach.

### 3.5.2 Baseline Hazard Estimation

With the advancement of REMs, the prevailing method for their estimation is based on the partial likelihood method [Cox, 1975], which treats the baseline hazard as a nuisance parameter. However, gaining insights into the temporal variations of the underlying event rates can be valuable for visualizing the baseline hazard.

Two common approaches to estimate the baseline hazard are the Breslow estimator [Breslow, 1972] and the Nelson–Aalen estimator [Nelson, 1972; Aalen, 1978]. Both estimate the cumulative hazard function only at the observed event times, and so does not capture the continuous nature of the underlying baseline hazard function. Meijerink-Bosman et al. [2022a] estimated the baseline hazard by assuming fixed baseline hazard rates within expanding windows. Juozaitienė and Wit [2022a] and Juozaitienė et al. [2022] improved smooth baseline hazard recovery by a spline-based approximation.

### 3.5.3 Model Comparison and Diagnostic Tools

Traditional approaches, such as likelihood ratio tests, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), are widely adopted for comparing REMs [Foucault Welles et al., 2014; Pilny et al., 2016]. Whereas model comparison methods are able to identify the best fitting model in a set of competing models, they do not indicate whether the fit is adequate.

Butts and Marcum [2017] propose an approach to model adequacy assessment based on deviance residuals and, so-called, surprise metrics. Similarly, Meijerink-Bosman et al. [2022a] suggest recall-based adequacy checking based on whether the observed events are in the top 5% of dyads with the highest predicted rates. Brandenberger [2019] proposes a procedure for model-based simulations of relational events. This method involves making predictions based on survival probabilities, which can then be used to generate new event sequences. In turn, by comparing the simulated event sequences with the original data, it becomes possible to assess whether the model can accurately replicate network characteristics.

Measures borrowed from survival analysis can also be used to assess the goodness of fit of REMs. Juozaitienė et al. [2022] propose using scaled Schoenfeld residuals to assess the proportional hazards assumption, deviance residuals to check for outliers and potential influential observations, and trends in martingale

residuals to check for non-linear effects of the covariates. Different model specifications are compared according to their deviance explained through pseudo  $R^2$  measures [Cox and Snell, 1989]. Boschi et al. [2023] extended the martingale score process to evaluate the goodness of fit for fixed, smooth, and random effects.

Guidelines on the statistical accuracy and precision of the REM are summarized in Schecter and Quintane [2021], and defined by conducting experiments on simulated sequences of relational events, to which different sampling and scaling procedures have been applied. Meijerink-Bosman et al. [2022b] showed that the accuracy and precision of REM estimates depend on the width of the selected temporal window.

#### 3.5.4 Bayesian Estimation of the Relational Event Model

REMs can be fitted also using Bayesian approaches [Butts, 2008]. DuBois et al. [2013b] extended the standard REM to incorporate multiple sequences, proposing a hierarchical model. Mulder and Leenders [2019] modeled multiple event sequences, estimating parameters that capture both within-sequence and between-sequence variations. This is particularly useful in multiplex networks, where multiple relational event sequences may be observed within the same network. Vieira et al. [2022] introduced a Bayesian hierarchical model that enables inference at the actor level, providing valuable insights into the drivers influencing actors' preferences in social interactions. Arena et al. [2022] and Arena et al. [2023] proposed different solutions for studying memory decay in REMs, showing that the Bayesian approach allows the estimation of short- and long-term memory effects on the model parameters in relational event sequences.

#### 3.5.5 Tools for Analyzing Relational Event Models

There are limited specialized software fitting REMs. The R-based packages `relevent` and `informR` [Marcum and Butts, 2015] are widely adopted, including all the REM features discussed in Butts [2008]. Another R-based option is the `rem` package [Brandenberger, 2018a], which allows the computation of endogenous covariates in signed one-, two-, and multi-mode event networks. Also the `remstats` package [Meijerink-Bosman et al., 2022b] assists empirical researchers in computing network covariates. It is typically used in combination with the `relevent` or `reestimate` packages for model estimation. A further R-based option is `goldfish` [Stadtfeld et al., 2017], which mainly supports DyNAM and is currently undergoing further development.

eventnet [Lerner and Lomi, 2020b] offers a reliable and scalable Java-based interface for the computation of endogenous and exogenous covariates that serve as inputs of proportional hazard regressions. Bauer et al. [2021] and Fritz et al. [2021] showed that time-stamped relations data can be fitted through the well-established R-package `mgcv` [Wood, 2017] for generalized additive models with penalized likelihoods after a proper data reorganization.

One potential challenge for future studies is the development of a comprehensive package capable of computing network covariates under different sampling schemes while accommodating different model assumptions. Ideally, such a tool should be integrated within a suite of packages encompassing different estimation techniques and diagnostic tools as well.

## 3.6 Applications of Relational Event Models

Empirical studies adopting REMs span a wide range of disciplines. In this section, we classify these studies into established subjects, presented below in alphabetical order. Within each category, we offer an illustrative rather than exhaustive collection of research questions that have been investigated using REMs.

### 3.6.1 Communication

Within the field of communication studies, broadly construed, REMs have been adopted to analyze conversational processes within and between organizations and teams as well as computer-mediated speeches.

Butts [2008] and Renshaw et al. [2023] demonstrated the practical value of REMs in a study of organizational communication in the context of emergency management. REMs make it possible to specify covariates that capture basic conversational norms [Gibson, 2003, 2005], such as the expectations of reciprocity in turn-taking.

Leenders et al. [2016] and Quintane and Carnabuci [2016] identified a number of challenges that hinder the identification of team dynamics and elaborated a REM-based analytic framework that supports a time-sensitive understanding of communication processes within teams. Similarly, Quintane et al. [2013], Pilny et al. [2016], and Schecter et al. [2018] revealed how REMs could be applied to studying the association between communication patterns and common indicators of process quality, coordination, and information sharing.

Computer-mediated communication is typically analyzed via two-mode REMs, which establish links between individuals and the situations they are involved



in, such as the questions they answer in online Q&A communities [[Stadtfeld and Geyer-Schulz, 2011](#)], or problems they attempt to resolve in open-source online projects [[Quintane et al., 2014](#); [Tonellato et al., 2023](#)]. In their study of communication instances in an online friendship network, [Foucault Welles et al. \[2014\]](#) adopt REMs to identify patterns of time dependence in data produced by online chats, focusing on how heterogeneous communication processes influence the creation, maintenance, and dissolution of communication ties over time.

### 3.6.2 Ecology

Studying behavior as sequences of relational events among animals promises to improve the understanding of key issues in behavioral ecology, such as how reciprocal giving and pro-social behavior emerge in small animal communities. Examples of this line of research include the study of [Tranmer et al. \[2015\]](#) on group interactions among captive jackdaws, with a focus on their food-sharing habits, and among cows struggling with the introduction of unfamiliar members in their community [[Patison et al., 2015](#)].

REMs have been adopted for studying ecological niche invasions [[Juozaitienė et al., 2022](#)] through the analysis of two-mode event networks linking invading species to territories. The relational event framework adopted in this study sheds light on potential risks associated with invasive species and develops insights into the ecological factors that may attract non-native species.

### 3.6.3 Health and Healthcare

In health and healthcare research, REMs have been adopted in the study of social interaction in surgery rooms and inter-hospital patient transfers, with the aim of understanding how collaborations among healthcare units may or may not improve the quality of care.

[Pallotti et al. \[2022\]](#) examined audio-visual recordings of task-related interactions among members of surgical teams to make sense of patterns of interpersonal communication among doctors and nurses organized around objects and technologies in the surgery room.

[Lomi et al. \[2014\]](#) studied inter-hospital mobility in a small community of hospitals and found that decentralized patient-sharing decisions ensure patients' access to higher-quality healthcare services. [Vu et al. \[2017\]](#) showed that patient transfers are usually organized around small clusters of hospitals including reciprocated patient exchange. Studying collaborations among hospitals in a regional community in Southern Italy, [Amati et al. \[2019\]](#) explained that the

generative mechanisms controlling change in event networks do not operate homogeneously and synchronously over time, but vary systematically over different days of the week. Similarly, [Zachrisson et al. \[2022\]](#) investigated the influence of characteristics such as reputation and institutional affiliation on the choice of destination hospital for emergency patients in the state of Massachusetts.

#### 3.6.4 Political Science

Within the field of political science, REMs are frequently employed in their two-mode version. In the typical application, political actors are linked to social activities, such as participation in cosponsoring events or support expressed for claims. [Brandenberger \[2018b\]](#) studied favor trading in congressional collaborations by examining the temporal dynamic of reciprocity, and found that the emergence of new collaboration clusters depends on the timing of mutual co-sponsorship. Due to the variety of actors involved in the political debate, [Haunss and Hollway \[2022\]](#) adopted a multimodal extension of the DyNAM framework to study the political discourse around Germany's nuclear energy phase-out. This work identifies the potential discursive mechanisms that may have influenced the debate, and when they may have operated.

REMs have occasionally been employed in criminology studies to investigate various aspects of offending behavior among individuals [[Niezink and Campana, 2022](#)]. These studies have explored phenomena such as illegal drug exchange in online markets [[Duxbury and Haynie, 2021, 2023](#)] and the replication tendency of gang violence [[Gravel et al., 2023](#)]. Overall, these analyses have demonstrated that criminal networks and personal attributes exert a substantial and temporally contingent influence on individual criminal acts.

#### 3.6.5 Sociology

Because of a general familiarity with social network methods and models, sociology is perhaps the field where REMs have found more extensive application. Restricting the focus on reciprocity, [Kitts et al. \[2017\]](#) and [Lomi and Bianchi \[2021\]](#) found that the mutual exchange of resources does not operate uniformly across different exchange regimes, time frames, and material settings defined, for instance, by the value of resources being exchanged [[Zappa and Vu, 2021; Bianchi and Lomi, 2022](#)].

[Lerner and Lomi \[2017\]](#) and [Lerner and Lomi \[2020a\]](#) analyzed the emergence of status and hierarchies under conditions of extreme decentralization characterizing the Wikipedia crowdsourced project, in which independent volunteers

interact by editing pages. In a study of individual editing behavior and collaboration/contention among Wikipedia editors, [Lerner and Lomi \[2019\]](#) examined the relation between team diversity, polarization, and productivity.

In the sociology of education, REMs have been applied to study how the nature of learning environments affects students' outcomes. [Vu et al. \[2015\]](#) studied educational experiences in massive open online courses by analyzing interactions between students with, and through an online learning interface, between students and their learning interface. [DuBois et al. \[2013b\]](#) investigated the social dynamics of high school classrooms by considering how the individual propensity to share information is affected by factors such as seating arrangements, teaching style, or sequences of participation shifts in conversation among students and teachers.

## 3.7 Open Issues and Challenges

In this section, we describe several open challenges in the context of relational event modeling. Although not exhaustive, it points to a number of interesting directions in which we expect significant progress in the near future.

### 3.7.1 Procedures for Assessing Goodness of Fit

A pressing issue in relational event modeling involves the absence of a comprehensive procedure for assessing goodness of fit. Methods have been proposed involving recall adequacy, and traditional residual approaches. However, there exists no general consensus regarding a formal testing paradigm. In general, what seems to be missing is an approach to goodness of fit of REMs consistent with established auxiliary variable approaches developed for assessing the goodness of fit of ERGMs [[Hunter et al., 2008](#)] and SAOMs [[Lospinoso and Snijders, 2019](#)].

### 3.7.2 Relational Big Data

Inference for REMs encounters a computational bottleneck as the number of relational events and, specifically, the number of actors increases. This presents practical implications, as it limits the applicability of REMs to large-scale networks. Addressing this limitation is an essential unresolved matter for REMs. Building upon the risk set sampling concepts introduced in [Vu et al. \[2015\]](#) and [Lerner and Lomi \[2020b\]](#), [Filippi-Mazzola and Wit \[2024b\]](#) have proposed a

*Stochastic Gradient Relational Event Additive Model* (STREAM) to analyze the network of patent citations in a dataset consisting of over 100 million events and 8 million nodes. By integrating case-control sampling with deep learning techniques, they have successfully achieved significant computational efficiency and real-time estimation of the model parameters.

### 3.7.3 Current Developments of Relational Event Models

Fritz et al. [2023] introduced a *Relational Event Model for Spurious Events* (REMSE) to address the issue of events recorded through machine-coding errors, which can give rise to false positives and false negatives. The REMSE is presented as a robust tool suitable for studying relational events generated in potentially error-prone contexts. Its intensity is decomposed into two components: one associated with true events and the other with spurious events. This decomposition allows for a more accurate representation and understanding of the underlying dynamics in the presence of potentially erroneous data.

Within the realm of coauthorship networks, Lerner and Hâncean [2023] adapted the RHEM [Lerner et al., 2021] to settings where events have a measurable outcome, such as a performance measure. These outcomes can serve as additional explanatory variables in the RHEM or can be used as response variable. This extension, known as the *Relational Hyperevent Outcome Model* (RHOM), implies that event rates and relational outcomes are determined by the same explanatory variables utilized in the RHEM.

Incorporating relational event dynamics into a latent space or latent clustering allows for novel hypotheses about drivers of network formation. DuBois et al. [2013a] combined ideas from stochastic block modeling and REMs by partitioning the node-set, where event dynamics within and between blocks evolve in distinct ways. Matias et al. [2018] developed a variational expectation-maximization method to estimate the latent groups. Moreover, combining REMs with latent space modeling allows the representation of actors as points in space, whose mutual distances drive the relational event process. Artico and Wit [2023] proposed a Kalman filter-based approach to estimate the trajectories of an actor in a latent space. An alternative approach is discussed in Rastelli and Corneli [2021], where the likelihood of an event given the current latent positions is maximized by stochastic gradient descent.

## 3.8 Concluding Remarks

Since their introduction fifteen years ago [Butts, 2008], REMs have undergone considerable refinement [Brandes et al., 2009; Perry and Wolfe, 2013], encouraged important extensions [Lerner and Lomi, 2023], and enabled development of substantive applications [Vu et al., 2015, 2017; Lerner and Lomi, 2020a]. Progress on these various fronts contributed to establish relational event modeling as one of the most promising frameworks for the analysis of dynamic network data.

REMs add to the existing set of statistical models for the analysis of dynamic networks the possibility of using the information contained in the sequential order of social interaction events when social interaction events are transformed into network ties defined at more aggregate time scales. Conversation [Gibson, 2003], financial transactions [Bianchi and Lomi, 2022], technology-mediated communication [Butts, 2008], problem-solving [Tonellato et al., 2023], disaster management [Renshaw et al., 2023], medical emergencies [Zachrisson et al., 2022], are only few examples of processes where the sequential timing of relational events is essential for understanding the underlying observation-generating mechanisms. In situations characterized by comparable sequential constraints on social interaction, statistical models that assume the concurrency of network “ties” leave unresolved problems related to the fact that network mechanisms operate over different time frameworks, and are regulated by different time-clocks [Bianchi et al., 2022].

This review outlined the core properties and the mathematical underpinnings of REMs by tracking the development of the original model since its appearance in 2008. We devoted special attention to the challenges posed by the estimation of REMs, and discussed the computational approaches proposed to address the complexities of the original model and its successive variants. We emphasized the flexibility of the relational event modeling framework, which allows the empirical specification to account for endogenous as well as exogenous covariates that may affect observed patterns of interaction. We discussed the various ways in which time may influence the impact that past events may have on future events — an issue that may be considered an empirical feature of the data that should be accounted for, or an opportunity to develop theoretically inspired hypotheses about how time affects social interaction.

We intended our review to appeal to a broad audience comprising both empirically minded researchers confronting problems posed by the analysis of relational event data with complex temporal dependencies, and statisticians interested in the analytical opportunities offered by recent advances in dynamic stochastic

models for social interaction. We expect that future progress in the modeling of relational events, and social networks more generally, will depend on the extent to which members of these communities will continue to discover areas of intersection for their interests.

## Chapter 4

# A Stochastic Gradient Relational Event Additive Model for Modelling US Patent Citations from 1976 until 2022

The following chapter was published as:

Filippi-Mazzola, E. & Wit, E. C., (2024). A Stochastic Gradient Relational Event Additive Model for modelling US patent citations from 1976 until 2022. *Journal of the Royal Statistical Society: Series C*, qlae023.

### 4.1 Introduction

Patents are not only a means of protecting intellectual property but also provide valuable information about the state of the art in technology and the evolution of knowledge and innovation over time [Trajtenberg and Jaffe, 2002]. The patent citation network captures the relationships between patents, where each citation represents a connection between two patents, indicating that the citing patent has built upon the knowledge contained in the cited patent [Sharma and Tripathi, 2017].

Patents represent a significant investment for many companies, and understanding the competitive landscape, and the strengths and weaknesses of competitors' patent portfolios can be essential for making strategic decisions about technology development, licensing, and litigation [Lerner, 1994]. Analyzing the factors that lead to a patent being cited can provide valuable insights into the underlying mechanisms driving innovation. Additionally, understanding the drivers of patent citation can inform decision-making in various contexts, such as technology development, intellectual property management, and innovation policy

[Ernst, 2003]. However, patent data analysis is a complex and challenging task, requiring advanced techniques and tools for managing and analyzing large and complex datasets.

The relational event model (REM) [Butts, 2008; Perry and Wolfe, 2013] has emerged as a powerful tool for modelling complex relational data. Although REMs were first introduced in the social sciences as a way of modelling the temporal dependencies between interactions in social networks, they have been applied in many different contexts, such as two-mode networks [Vu et al., 2017], animal behavioral interactions [Tranmer et al., 2015], and more recently, financial transactions [Bianchi et al., 2022] and invasive species analysis [Juozaitienė et al., 2023]. Following these examples, REMs can be a valuable tool for analyzing citation networks of patents, as they allow us to model the complex relationships between citing and cited patents, identifying the factors that influence the diffusion of knowledge and innovation. However, the practical applicability of REMs is limited by their runtime complexity [Welles et al., 2014], a problem rooted in the denominator of the partial likelihood on which the estimation of most REMs is based. There have been some early attempts to model citation networks through a REM-like approach [Vu et al., 2011]. Recently Lerner and Lomi [2020b] tackled the inherent computational issues by showing the robustness of REM estimation when controls and cases are sub-sampled through a nested case-control approach [Borgan et al., 1995]. This was first introduced in REMs by Vu et al. [2015].

The standard log-linear formulation of a REM is a convenient simplification that does not always suffice. For this purpose, Fritz et al. [2023] introduced non-linear effects to model non-linear structures. Nevertheless, as shown by Bauer et al. [2021], introducing non-linear effects when relational event models are applied to the patent citation network reaches practical limitations in memory management and optimization. Standard approaches fail to model the full event-set and result in extensive computing times.

The stochastic gradient relational event additive model (STREAM) presents a solution to these challenges. STREAM approximates the likelihood of REMs using a logistic regression. This allows for a more versatile modeling approach, where each predictor can be represented by a smooth effect through B-splines [Schoenberg, 1946, 1969; De Boor, 1972]. To address the estimation challenges in large networks, particularly when using smooth effects, STREAM employs the Adaptive Moment (ADAM) optimizer [Kingma and Ba, 2017] for estimating the model's coefficients. Overall, STREAM captures non-linear relationships between variables, providing more valuable interpretations of time-varying effects while identifying the most influential factors driving patent citations.



For our analysis, we used data obtained from the United States Patent and Trademark Office (USPTO), the federal agency responsible for granting patents and registering trademarks in the United States. The USPTO data is one of the most comprehensive sources of patent information in the world as it contains precise information contained in standard digitalized formats on all patents issued in the United States since 1976. While there are limitations to extrapolating the USPTO data to other regions, it is still a good proxy for global patent activity as well as a source for studying innovation and technological progress. Overall, by using STREAM, we gain important insights into the dynamics of patent citations while opening the road to further speculations on the current state of the innovation process.

In this chapter, we start by describing the USPTO patent data in section 4.2 on which this analysis is based. After developing the theoretical foundation in section 4.3, we apply the framework to the patent citation network in section 4.4. Although STREAM was specifically designed to work with citation networks, this modelling framework can easily be applied to model general relational event data.

## 4.2 Patent citations as event history data

A patent citation is an essential element of the patent system as it provides a means of demonstrating the novelty, non-obviousness, and importance of an invention. Indeed, a patent citation is crucial for both patent examiners and inventors, as it allows the examiner to evaluate the claims made in the patent application, and it helps the inventor establish the scope and value of their invention. In this regard, in many jurisdictions, applicants are legally obliged to cite those patents on which the patent builds forth as part of a patent deposition. The triple consisting of the instance of deposition, the citing, and cited patents can be seen as an instance of a relational event. Collections of patent citations constitute a citation network, which is a particular kind of temporal-directed graph, where new actors join the network and bind to existing nodes. In most situations, the citation is due to content similarity or other exogenous drivers. This is in contrast to classic social network architectures, where tie formation is a more endogenous process, based on, e.g., repetition, reciprocity, or triadic effects.

In large jurisdictions, patent citation networks consist of millions of time-stamped recorded citation events. The generative process of the US patent deposition gives important clues for modelling the resulting citation network. When a patent is filed, the owners have a legal requirement to fulfill the duty of dis-

closure. This consists of providing within the application a list of existing technologies or scientific discoveries that are related or considered to be fundamental for the creation of the patenting invention. Patent office examiners will only grant the patent if the application meets the uniqueness requirement and if the invention is fully disclosed in the documentation presented. The patent citation process conforms to the specific structure of event history data. The event set consists of a citation-based relationship between a specific sending, deposited patent and a receiving, pre-issued patent.

The patent citation network suffers from several boundary issues, relating to both space and time. With regard to space, different national or transnational jurisdictions have different application processes. Despite their similarities, slight differences in the juridical procedures make the citation-generating process both country- or region-specific. A clear example of this is the following difference between the citation procedures between the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO). In the latter, the examiner committee has to integrate additional documents and patent citations. EPO examiners, on the other hand, do not include any citations but evaluate if the invention has been properly disclosed by the cited documents. This difference results in USPTO patent citations typically surpassing the EPO patent citations by a large amount, sometimes referred to as a “patent office bias” [Bacchiocchi and Montobbio, 2010]. We focus in this study on the USPTO patent citation network. Concerning the time boundary, the electronic recording of patent citations has only started relatively recently. Although some sporadic efforts have been undertaken to record historic patent citations, this is far from complete. We focus our analysis on those patents issued by the USPTO between 1976 and 2022. The starting year of our observed period coincides with the initialization of the digitization process of US-patents.

In our analysis, we make use of the original USPTO online repository (<https://bulkdata.uspto.gov/>). This makes the raw material of this analysis as much standardized as possible in terms of general information available. Although there are various distributions available of the USPTO data, after careful evaluation we decided to avoid any third-party pre-processing. The raw USPTO XML files were processed in a uniform manner and combined to obtain CSV files through open-source software available at <https://github.com/efm95/patents>.

The resulting USPTO patent citation dataset consists of more than 8 million issued patents that generated 190 million citations. Despite the in-house processing by the USPTO, we have applied some data-cleaning procedures as a result of some specific features of the USPTO patents. First, by focusing our view on

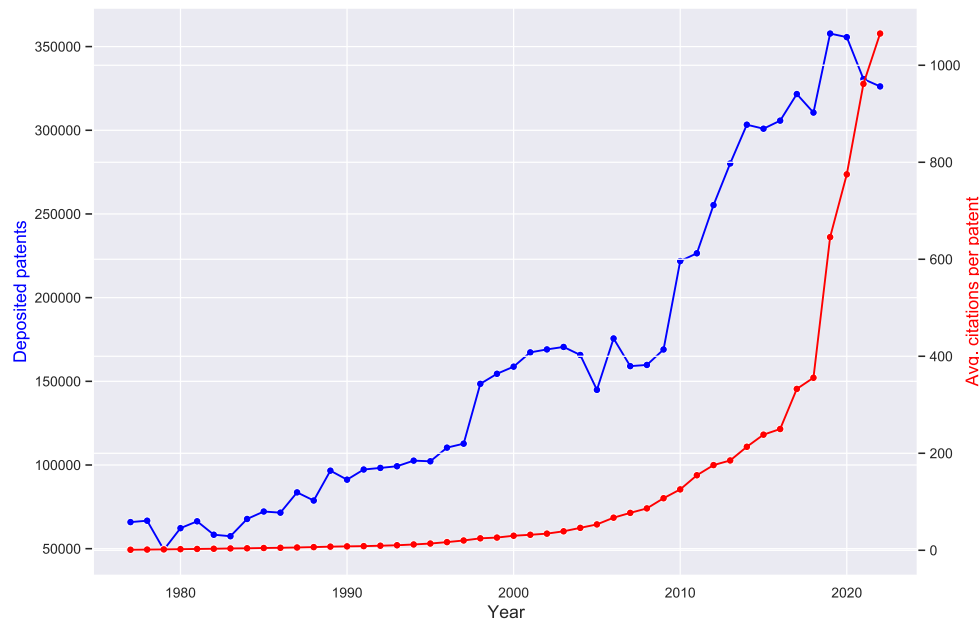


Figure 4.1: Number of deposited patents per year and the number of patent citations per patent per year since 1976.

patents issued only by the USPTO means losing track of those citations that go to patents outside the US jurisdictions. Secondly, in the same way as Whalen et al. [2020] and Filippi-Mazzola et al. [2023], we excluded all non-utility patents, such as plant and design patents, as these differ in many structural ways from the utility patents. With these two additional steps, our final dataset consists of around 100 million citations issued by a network of 8.3 million patents. The data preprocessing procedures for recreating the dataset can be found at [https://github.com/efm95/STREAM/tree/main/data\\_preprocessing](https://github.com/efm95/STREAM/tree/main/data_preprocessing).

Figure 4.1 shows that there has been a steady increase since 1976 in the number of deposited patents per year and a dramatic rise in the number of citations per patent. Various regulatory considerations have played a role. Failing to take those aspects into account will confound the picture of the true underlying innovation process. This paper aims to disentangle the causes that have contributed to this rise.

### 4.3 Stochastic Gradient Relational Event Additive Model (STREAM)

Relational event models (REMs) are a class of statistical models used to analyze event sequences and relationships between actors through a series of exogenous and endogenous effects based on the fine-grained event history process. In this section, we will extend the REM by developing the stochastic gradient relational event additive model for the network of patent citations.

#### 4.3.1 Relational event model

The temporal dynamic network is represented by a sequence of time-stamped events. Each event  $e_i$ , for  $i = 1, \dots, n$ , is recorded as the triple  $e_i = (s_i, r_i, t_i)$ , where  $s_i$  is sender,  $r_i$  the receiver and  $t_i$  the time at which the event takes place. As in [Perry and Wolfe \[2013\]](#), we define a counting process for the directed event that involves sender  $s$  and receiver  $r$  as

$$N_{sr}(t) = \#\{s \text{ interacts with } r \text{ up to time } t\}.$$

The counting process  $N_{sr}(t)$  is a local submartingale for which it is possible to define a predictable increasing process  $\Lambda_{sr}(t)$ , whose stochastic intensity function  $\lambda_{sr}(t_s)$  describes the tendency for  $s$  to interact with  $r$  at time  $t_s$ . Given the history of the network  $\mathcal{H}_{t^-}$  up to time  $t$ , it is possible to model the intensity function following the proportional hazard function [[Cox, 1972](#)]. The intensity function is given as the product of a baseline hazard  $\lambda_0$  and an exponential function of  $q$  covariates  $x_{sr}(t)$  with corresponding parameter  $\beta$ , i.e.,

$$\lambda_{sr}(t \mid \mathcal{H}_{t^-}) = \lambda_0(t) e^{\sum_{k=1}^q \beta_k x_{srk}(t)} \mathbb{1}_{\{(s,r) \in R(t_i \mid \mathcal{H}_{t^-})\}}, \quad (4.3.1)$$

where  $R(t_i \mid \mathcal{H}_{t^-})$  is the risk-set. The drivers of the relational process  $x_{sr}(t)$  refer to quantities that describe known statistics of the sender. These statistics can be either endogenous or exogenous. In the case of endogenous covariates, they depend on past interactions. On the other hand, covariates are considered exogenous if they depend on the characteristics of individual nodes (monadic covariates) or pairs of nodes (dyadic covariates). While events are assumed to be conditionally independent given the network of previous events, the inclusion of covariates in this model specification allows examining the impact of various drivers related to senders, receivers, or network topology. For a comprehensive overview of the most frequently analyzed statistics within REM applications, we

direct readers to the work of [Bianchi et al. \[2024\]](#), which provides an extensive list and discussion of these effects.

Given the difficulties that come with dealing with the full likelihood in (4.3.1), it is possible to estimate the coefficients through a partial likelihood approach [[Cox, 1975](#)], in which the baseline is treated as a nuisance parameter. The main idea of this approximation is to specify a partial likelihood that depends only on the order in which events occur, not the times at which they occur. Because the event time is by definition the publication date of the sender, the risk-set  $R(t | \mathcal{H}_{t-})$  consists of all potential receivers  $r$  that were present in the network at time  $t$  and, as a consequence, that could have been cited by the issued patent  $s$ . This results in the following partial likelihood,

$$L_P(\beta) = \prod_{i=1}^n \left( \frac{\exp \left\{ \sum_{k=1}^q \beta_k x_{s_i r_i k}(t_i) \right\}}{\sum_{r \in R(t_i | \mathcal{H}_{t_i-})} \exp \left\{ \sum_{k=1}^q \beta_k x_{s_i r k}(t_i) \right\}} \right). \quad (4.3.2)$$

In its logarithmic form (4.3.2) assumes a concave behavior, allowing the coefficients to be estimated via a Newton approach. The Partial-Likelihood in (4.3.2) is directly inspired by the model presented by [Butts \[2008\]](#) for temporal ordinal data. Indeed, (4.3.2) is a proper full-likelihood, as we model the probability of each subsequent event to occur as the product of multinomial probabilities.

We note that in the situation of the patent citation process, the event time and the appearance of the sender are equivalent. Although this changes the full likelihood, it retains the same partial likelihood formulation. One can argue that the citing process can be described as a dynamic egocentric network, where conditional on the publication process, the citation process is simply a multinomial selectional process described by the partial likelihood.

### 4.3.2 Case-control sampling of the risk-set and logit approximation

The practical applicability of the partial likelihood is compromised by runtime complexities in the computation of its denominator, involving the risk-set  $R(t | \mathcal{H}_{t-})$  [[Foucault Welles et al., 2014](#)]. As already noted by [Butts \[2008\]](#), the risk-set typically grows quadratically with the number of nodes in the network, making computations slow beyond a few hundred nodes. Even though the risk-set in our case consists only of alternative receivers, this still involves millions of patents, making the partial likelihood approach inaccessible for our problem.

The solution suggested by [Vu et al. \[2015\]](#) is to reduce computational complexity by applying nested case-control sampling on the risk set [[Borgan et al., 1995](#)].

The idea is to analyze all the observed events, i.e., citations or “cases”, but only a small number of non-events, i.e., non-citations or “controls.” [Borgan et al. \[1995\]](#) proved that maximum partial likelihood estimation with a nested case-control sampled risk-set yields a consistent estimator. This approach reduces the number of computing resources needed to build the risk set, however, it still makes heavy use of computer memory.

Empirical evidence presented by [Lerner and Lomi \[2020b\]](#) suggests that estimates are reliable with just one control per case. With a single control, the denominator in (4.3.2) is the sum of the rates for the cited patent with covariates  $x_{s_i r_i}$  and a randomly sampled non-cited patent with covariates  $x_{s_i r_i^*}$ . Then the nested case-control sampling version of the partial likelihood (4.3.2) is given as,

$$\tilde{L}_P(\beta) = \prod_{i=1}^n \left( \frac{\exp \left\{ \sum_{k=1}^q \beta_k \left( x_{s_i r_i k}(t_i) - x_{s_i r_i^* k}(t_i) \right) \right\}}{1 + \exp \left\{ \sum_{k=1}^q \beta_k \left( x_{s_i r_i k}(t_i) - x_{s_i r_i^* k}(t_i) \right) \right\}} \right), \quad (4.3.3)$$

which is the likelihood of a logistic regression with only ones as response and covariate levels  $x_{s_r k}(t) - x_{s_r^* k}(t)$ . This approximation reduces the amount of memory needed to analyze the full set of observed citations, while the concavity of the logit approximation ensures the convergence of any Newtonian optimizers.

### 4.3.3 Basis expansion of covariates

The core assumption of relational event modeling assumes that the rate of interaction between a sender  $s$  and a receiver  $r$  depends linearly on the covariates. Given the temporal complexity depicted in [Figure 4.1](#), it is reasonable to assume that could lead to an oversimplified representation of the patent citation process. From the logistic interpretation of the case-control partial likelihood, we propose to extend the REM via a generalized additive framework [[Hastie and Tibshirani, 1986](#)] by modelling single covariates via basis functions splines (B-splines) [[Schoenberg, 1946, 1969](#)].

B-splines are connected piece-wise polynomial functions of order  $p$  defined over a grid of knots  $u_0, u_1, \dots, u_m$ , such that  $u_{l-1} < u_l$ , for  $l = 1, \dots, m$ , on the parameter space that characterize the covariate  $x_{s_r k}(t)$ , for  $k = 1, \dots, q$ . In our modelling framework, we decided to place the knots evenly on the covariate support. Following [De Boor \[1972\]](#) recursive definition of basis function (see [Appendix A.1](#)), the B-spline effect associated to the  $k$ -th covariate  $x_{s_r k}$ , is then a

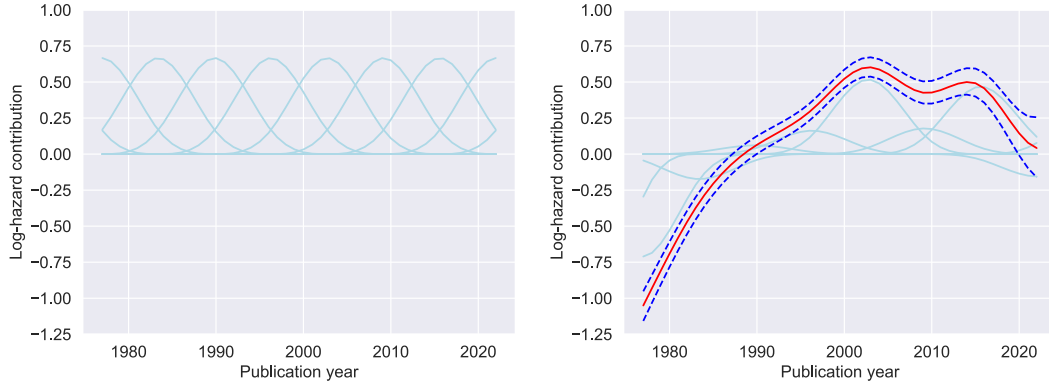


Figure 4.2: Left: receiver publication year uniform basis transformation. Right: receiver-publication year estimated effect, where the basis model-matrix is multiplied with its respective vector of estimated coefficients. The red line represents the estimated effect, while the dashed-blue lines are the confidence intervals.

linear combination of  $d$  coefficients with  $d$  basis functions, i.e.,

$$f_k(x_{srk}) = \sum_{j=1}^d \theta_{jk} B_{j,p}^k(x_{srk}).$$

Figure 4.2 shows a practical example on how B-splines are applied to the covariates. By substituting the basis expansion formulation in the hazard formulation in (4.3.1), the full model for the intensity function becomes

$$\lambda_{sr}(t | H_{t-}) = \lambda_0(t) e^{\sum_{k=1}^q f_k(x_{srk}(t))} \mathbb{1}_{\{(s,r) \in R(t_i | \mathcal{H}_{t-}^c)\}}, \quad (4.3.4)$$

where its partial likelihood approximation with one control is given as

$$\tilde{L}_p(\theta) = \prod_{i=1}^n \left[ 1 + \exp \left\{ - \sum_{k=1}^q \sum_{j=1}^d \theta_{jk} \left[ B_{j,p}^k(x_{s_i r_i k}(t_i)) - B_{j,p}^k(x_{s_i r_i^* k}(t_i)) \right] \right\} \right]^{-1}. \quad (4.3.5)$$

To smooth the estimated B-splines resulting from maximizing the partial likelihood in (4.3.5), various penalization terms can be added. One reliable option is the use of P-splines [Eilers and Marx, 1996], especially when dealing with flexibility at the boundaries of the covariate support. However, in order to calculate the penalty, a considerable number of bases must be generated. Using large degrees of freedom translates to high memory usage, as each predictor generates two matrices of dimension  $d$ . As a result, over-parametrizing each predictor

spline to provide smoothing can quickly exhaust the computer memory, making this procedure unsuitable for modelling large networks. In such situations, a cross-validation approach is preferred to select an appropriate number of basis functions, as memory constraints pose an upper limit on the number of degrees of freedom of the splines.

#### 4.3.4 Recovering baseline hazard

The partial likelihood approach avoids modelling the baseline hazard. Although this brings significant benefits in estimating the B-splines, it also loses information about the underlying rate of the process. The advantage of formulating the REM as a Cox regression is that we can rely on the survival modelling literature to estimate the cumulative baseline hazard post-hoc. We adapt the baseline estimator from the nested case-control sampling [Borgan et al., 1995] to estimate the underlying rate of the citation process. The adapted estimator for the cumulative baseline hazard is given as

$$\widehat{\Lambda}_0(t) = \sum_{t_i < t} \left[ \exp \left\{ \sum_{k=1}^q \hat{f}_k(x_{s_i, r_i, k}(t_i)) \right\} + \exp \left\{ \sum_{k=1}^q \hat{f}_k(x_{s_i, r_i^*, k}(t_i)) \right\} \right]^{-1} \frac{2}{n(t_i)}, \quad (4.3.6)$$

where  $n(t_i) = |\mathcal{R}(t_i | \mathcal{H}_{t_i^-})|$  is the number of events at risk at  $t_i$ , for  $i = 1, \dots, n$ , where  $t_i$  is the absolute time. A pointwise baseline hazard estimate can be calculated by taking differences between subsequent events of the cumulative hazard, i.e.,

$$\hat{\lambda}_0(t_i) = \frac{\widehat{\Lambda}_0(t_i) - \widehat{\Lambda}_0(t_{i-1})}{t_i - t_{i-1}}, \quad \text{for } i = 1, \dots, n. \quad (4.3.7)$$

#### 4.3.5 Parameter estimation using stochastic gradient descent

While the case-control partial likelihood helps to reduce computational complexity, it is not enough to overcome the optimization challenges presented by the size of the patent citation network. Most machine-learning techniques use stochastic gradient descent methods to address large optimization challenges. By separating the data stream into separate batches and adjusting the parameters after assessing each batch in succession optimization convergence can be achieved efficiently. As a result, even when working with large datasets, estimating the model parameters becomes manageable.

In this problem, we have opted for a stochastic gradient descent approach through the Adaptive Momentum (ADAM) optimizer [Kingma and Ba, 2017] to fit the



partial likelihood. Stochastic gradient descent has proved to be a reliable technique for estimating logistic regression models in large-scale scenarios [Lin et al., 2007; Bottou, 2010]. Different momentum-based approaches have been proposed in the last decade to solve problems connected to local minima, such as AdaGrad [Duchi et al., 2011] or ADADELTA [Zeiler, 2012]. Among these, ADAM has gained in popularity in the machine learning field for its scalability and its convergence reliability [Reddi et al., 2018]. ADAM uses adaptive learning rates that depend on estimates of the first and second moments of the gradients of the observed batch. It maintains an exponentially decaying average of past gradients and squared gradients, which it then uses to calculate the update step for the model parameters.

In many real-world scenarios, gradients are often sparse, which means that only a small fraction of the parameter’s partial derivatives are non-zero at any given time. In traditional gradient descent algorithms, these sparse gradients can result in slow convergence or even divergence. ADAM handles sparse gradients by incorporating a technique called moment correction, which adjusts the moment terms based on the frequency of non-zero gradients, which allows the optimizer to effectively use the sparse gradients. Although we did not experience any notable problems with sparse gradients in the optimization procedures, ADAM has been proven to be more stable than the classic SGD method. Let  $\nabla\tilde{L}_p(\theta)_b$  be the gradient evaluated on the partial likelihood on batch  $b$ . The ADAM routine updates the first and second moments according to the following routine:

$$\begin{aligned} m_b &\leftarrow \xi_1 m_{b-1} + (1 - \xi_1) \nabla\tilde{L}_p(\theta)_b \\ v_b &\leftarrow \xi_2 v_{b-1} + (1 - \xi_2) \nabla\tilde{L}_p(\theta)_b^2, \end{aligned}$$

where  $m$  and  $v$  are the first and second-moment gradients, respectively, and  $\xi_1$  and  $\xi_2$  are hyperparameters that control the importance of past information in updating the moments.

Furthermore, the ADAM algorithm uses bias correction to account for the bias introduced in the first and second moments of the gradients. The bias correction is necessary because the moving averages of the gradients (the first and second moments) are initialized to zero and thus biased towards zero, especially at the beginning of the training process. To correct this bias, ADAM applies a correction term to the moving averages, which is proportional to the learning rate and inversely proportional to the number of iterations. Let  $s$  be the current step of

the training process, then the first and second moments are corrected as follows,

$$\hat{m}_{b,s} = \frac{m_b}{1 - \xi_1^s}$$

$$\hat{v}_{b,s} = \frac{v_b}{1 - \xi_2^s},$$

where as  $s$  increases,  $\xi_1^s$  and  $\xi_2^s$  converge to 0. The model parameters are then updated according to the following rule:

$$\theta_b \leftarrow \theta_{b-1} + \psi \frac{\hat{m}_{b,s}}{\sqrt{\hat{v}_{b,s} + \epsilon}},$$

where  $\psi$  is the learning rate that determines the magnitude of each parameter update, while  $\epsilon$  represents a small scalar added to prevent division by zero (usually 1e-8). For our application,  $\psi$  has been kept constant.

The optimization procedure is repeated until the algorithm reaches the maximum point and the gradient becomes zero. At this stage, the method converges to a stationary distribution, indicating that the parameters have achieved a stable state where further parameter updates do not improve the model performance. It is important to note that the optimization process can be stopped earlier if the performance on a validation set starts deteriorating or if the maximum number of iterations is reached.

Overall, ADAM has demonstrated its effectiveness as a reliable optimizer for various machine learning applications as its computational complexity involves a constant number of operations that do not depend on the number of covariates. This gives STREAM a computational complexity on each batch of  $O(qbd)$ , where  $q$  is the number of covariates,  $b$  represents the batch size, and  $d$  is the basis dimensions. As a result, STREAM is estimated efficiently for a large number of observations even with the addition of additive components described by B-splines.

## 4.4 Modeling patent citations

The key question we seek to answer is what are the drivers of patent citations. The mechanisms that produce the patent citation network can be endogenous and exogenous. We will begin with the effects that we considered and how models including various effects have been compared. We then discuss a description of the model implementation. We complete the section with a discussion of the

results we have found and their implications for the patent citation process.

#### 4.4.1 Potential drivers of patent citations

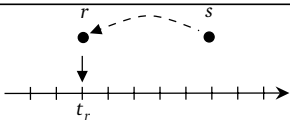
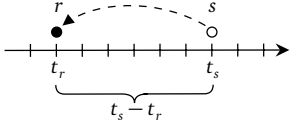
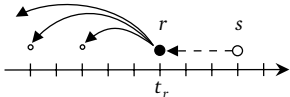
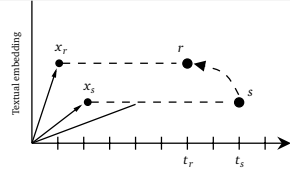
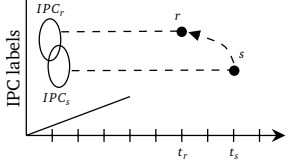
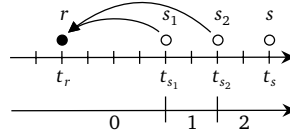
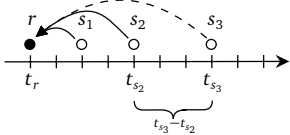
In this section, we describe the type of drivers we consider in the patent citation process. The specific nature of citation networks prevents the emergence of typical network effects that REMs commonly capture. This is primarily because each patent can only cite other patents from the past and only at the time of its own publication. While many fundamental network effects are absent by definition in this context, the ones described in this section adequately capture distinct factors of the patent citation process. We divide the type of tested statistics into node effects, patent similarity effects, and time-varying effects, related to viral and saturation considerations of patent citations. Table 4.1 contains an overview of all the effects and their respective mechanisms. Absent from the table are triadic effects. However, one may argue that due to the size of the patent citation network open triangles are unlikely to exert influence. Extensive model selection analysis can be found in the *appendix A.2*.

*Node effects.* Nodal effects refer to specific information about the cited patent, such as the publication year of the cited patent. A non-linear *cited patent publication year* effect can uncover any tendencies where patents issued in specific years are being cited more consistently. This can potentially indicate a period of significant technological advancement.

In addition, the time that has elapsed between the publication of the patents and the moment these have been cited can also be a driving factor in the patent citation network. This *time lag* effect can provide insight into whether patents tend to cite more recent material, reflecting the current state of technological innovation. By counting the number of days between the citing patent issue date and the receiver publication date, we can model this *time lag* effect and account for the time that has passed between the two nodes appearing in the network.

Not all patents are equally important in terms of their connectivity. One hypothesis may be that if a particular patent summarizes a lot of older knowledge, it may attract more citations. This hypothesis is sometimes referred to as the *cumulative process of knowledge creation* [Scotchmer, 1991]. To test for this hypothesis, we use the *receiver outdegree* as a proxy for centrality. In this context, the outdegree of a patent represents the number of patents itself cites at the time of its publication.

Table 4.1: Effects and their corresponding mechanisms. Where  $r$  represents the cited patent, i.e. the receiver;  $s$  represents the citing patent, i.e. the sender. Respectively,  $t_s$  and  $t_r$  represents the issue date for the sender for the receiver. Note:  $\mathbf{x}_r$  and  $\mathbf{x}_s$  are the patents abstracts embedding vectors, while  $IPC_r$  and  $IPC_s$  are vectors of IPC patent classes.

Effect	Mechanism	Statistic
Receiver publication year		$t_r$
Time lag		$t_s - t_r$
Receiver outdegree		$\sum_{r' \in R(t_r   \mathcal{H}_{t_r^-})} \mathbb{1}\{(r, r', t_r)\}$
Textual similarity		$\frac{\mathbf{x}_r \cdot \mathbf{x}_s}{\ \mathbf{x}_r\  \ \mathbf{x}_s\ }$
IPC relatedness		$\frac{ IPC_r \cap IPC_s }{ IPC_r \cup IPC_s }$
Cumulative citations received		$\sum_{t_i < t_s} \mathbb{1}\{(s, r, t_i)\}$
Time from last event		$\min_{t_i < t_s} \{(t_s - t_i) \mid \mathbb{1}\{(s_i, r, t_i)=1\}\}$

Patent similarity effects. Citations in patents arise from the assumption that there exists some technological similarity between the citing and the cited patent. However, technological relatedness is not a particularly tangible concept. Two distinct types of relatedness are often used to capture this idea.

The first type is a direct textual similarity between the citing and cited patents. Although there have been debates about the reliability of textual similarity as a measure of technological relatedness, recent studies [Kuhn et al., 2020; Filippi-Mazzola et al., 2023] have shown that it plays an important role in patent citation networks, when combined with other metrics. Following the same procedure described in Filippi-Mazzola et al. [2023], we calculate textual similarities of a pair of patents through a pre-trained Sentence-BERT neural network [Reimers and Gurevych, 2019]. It uses a vectorized loop to compute pairwise similarities of the abstracts of citing and cited patents.

Another important measure of technological relatedness is the overlap in technology classes between the citing and the cited patents. Patent classification systems, such as the International Patent Classification (IPC) scheme, are designed to facilitate the search for related technologies by classifying patents into a systematic hierarchical structure. Deeper levels of the classification indicate higher levels of specificity in the technological field. However, patent classes present several challenges. Patents often straddle various technological fields, and, as a result, may be allocated to multiple IPC classes. Furthermore, new IPC classes have been created, or existing ones have been merged or split since the creation of the USPTO [Younge and Kuhn, 2016], leading to a somewhat organic organization of technology classes. Despite these challenges, patent classes remain a crucial element in the patent-issuing process. To test the hypothesis that the assigned labels play a role in the citation process, we calculate the *Jaccard similarity index* for the IPC classes of the cited and citing patents [Yan and Luo, 2017]. The Jaccard index measures the similarity of the patent classes between two patents, taking into account the sub-class levels of the IPC classification. Filippi-Mazzola et al. [2023] have shown that both the main component *section* and the third component *sub-class* share similar importance in analyzing patent classes.

Time-varying effects. We will also consider two time-varying effects. Citation networks have distributional characteristics that are consistent with a viral process [Redner, 1998]. Popular patents, for whatever reason, may be more likely to draw more citations. We define for every patent the *cumulative number of citations received*. This is also known in network science as *receiver indegree* or *preferential attachment*.

This popularity effect may experience saturation. For this reason, we also consider the *time from the last event*, i.e., the last time the patent was cited. This variable captures how long it has been since the patent was last cited, and its influence on the rate of receiving a new citation.

Including time-varying effects to the model specification presents an additional challenge. Specifically, each time a new control is sampled, the time-varying covariates need to be updated within the risk-set according to the current observed event time  $t$ . Consequently, this complicates the creation of the model matrix. To overcome this problem, we used a similar approach of the “caching” data structures method proposed by [Vu et al. \[2011\]](#). Rather than uniformly sampling  $x_{sr \times k}(t)$  from  $R(t | \mathcal{H}_{t-})$ , we select a subset of control candidates from the risk set, denoted by  $\tilde{R}(t | \mathcal{H}_{t-}) \subseteq R(t | \mathcal{H}_{t-})$ , such that for each event-time  $t$ , we sample  $c$  potential receivers as control candidates. For each element in  $\tilde{R}(t | \mathcal{H}_{t-})$ , we update its relative time-varying effect at every observed time  $t$ . Depending of the size of  $c$ , we can store  $\tilde{R}(t | \mathcal{H}_{t-})$  in memory, without needing to update the full risk-set  $R(t | \mathcal{H}_{t-})$  every time a non-event is sampled. This significantly reduces the burden of creating the model-matrix.

Overall, incorporating time-varying effects in our model specification improves the accuracy and robustness of our analysis by accounting for the dynamic nature of patent citation behavior over time.

#### 4.4.2 Implementation

Although the process of generating basis functions from events and estimating the coefficients can be tackled by well-optimized R algorithms like the `gam` function in the `mgcv` package [[Wood, 2011](#)], it is unable to deal with 100 million patent citations [[Oancea and Dragoescu, 2014](#)]. The R software memory management system struggles with large data objects, resulting in limitations to the practical applicability of routines, such as `gam`. This complicates the estimation of the coefficients through the optimizers in `mgcv` as they would require a considerable amount of time to reach convergence. Spline basis expansions require the storage in memory of as many  $n \times d$  matrices as there are covariates in the model. In fact, in the REM partial likelihood formulation (4.3.5), this involves  $2q$  matrices for both cases and controls.

The model fitting problem will, therefore, be divided into two parts: (i) defining an efficient way to compute the basis function for millions of rows, and (ii) avoiding generating matrices that exceed the available memory.

To tackle the first problem, we create a vectorized recursive algorithm that efficiently generates basis functions from millions of elements in a vector. Dividing

the input data into batches is analogous as taking random samples from a larger population. The value associated with each observation following the basis function transformation is invariant to the position of the event in the observed set. Rather than applying the basis function transformation on the entire observed stream of the events, these can be computed directly on each batch when the gradient needs to be computed. This reduces memory usage at the expense of a small increase in computational costs. The implementation relied on the Python suite PyTorch [Paszke et al., 2019], which also provides access to the computational benefits of Graphic Processor Units (GPUs) to scale matrix multiplications and gradient computations. The vectorized nature of Pytorch and the use of GPU computational power are particularly suited for the recursive algorithm, drastically reducing the computational time for generating B-splines. Then, by dividing the stream of data into different batches, we can efficiently estimate the coefficients by iteratively updating them with respect to the back-propagated gradients, computed using the negative log-partial likelihood (4.3.5) as our loss metric. The code for STREAM can be found at <https://github.com/efm95/STREAM>.

#### 4.4.3 Interpretation of results on USPTO patent citation data 1976-2022

The stochastic component of the optimizing ADAM method introduces some additional randomness into the estimation of the model parameters but given the size of the data we obtain highly concentrated estimates. Figures 4.2-4.5 show the fitted splines with 10 degrees of freedom. Uncertainty estimates are provided via pointwise quantile confidence intervals estimated through 100 non-parametric bootstrap resamples. The y-axes indicate the log-hazard contribution to the citation rate of an individual patent. An increase by 0.7 on this scale indicates a doubling of the citation rate.

Node effects. One remarkable result can be seen in the *receiver publication year* curve in Figure 4.2. Contrary to the widely reported continuous increase in the patent depositing and patent citation process [Kuhn et al., 2020; Whalen et al., 2020], the rate with which an individual patent gets cited possesses a distinct peak. The peak occurs shortly after the year 2000. This means that, after accounting for all other effects, patents that were published around 2000 are, individually, attracting more citations than at any other period from 1976 to 2022. Patents from around 2000 tend to attract 70% more citations than those published around 2022, and more than 5 times more citations than those published

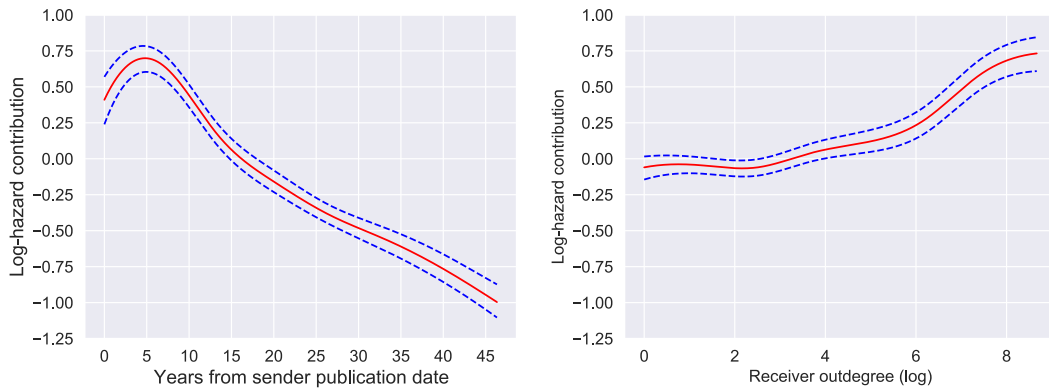


Figure 4.3: Splines associated to the nodal effects. Left: time lag. Right: receiver outdegree in log terms.

in 1976. While this study is limited to a macro-level network analysis, we hypothesize several key technical breakthroughs may have occurred around 2000. Park et al. [2023] also reported the recent decline in the disruptiveness of patents. However, in contrast to their findings, we find clear evidence for monotone increasing innovation from 1976 before peaking around the year 2000. It may be that by failing to take into account the growing patent network, their initial decline is an artifact.

The *temporal lag* spline in Figure 4.3 indicates at which time in the future patents are most likely to be cited. The curve shows that there is a peak around year 5. This indicates the presence of a sweet spot of approximately 3 to 7 years after the original publication of the patent where citations are most likely to arise. It is important to note that this temporal lag effect could be influenced by various factors such as the pace of technological development, the lifespan of technology, and the overall trends in the field. This effect provides valuable insights into the timing of patent publications and their impact on the citation network. By identifying this sweet spot where citations are more likely to arise, inventors can strategically plan their patent filing and publication strategies to increase their chances of being cited and recognized in the field. Furthermore, the inclusion of *temporal lag* into the model deals with the boundary problem, as it accounts for the fact that recently published patents are unlikely to have gathered a significant number of citations.

The *receiver outdegree* effect in Figure 4.3 shows that patents that cite a lot of other patents are more likely to be cited themselves. This finding highlights the importance of citing all relevant patents in one's patent application, as it makes a patent more visible and accessible to other inventors, increasing the likelihood of



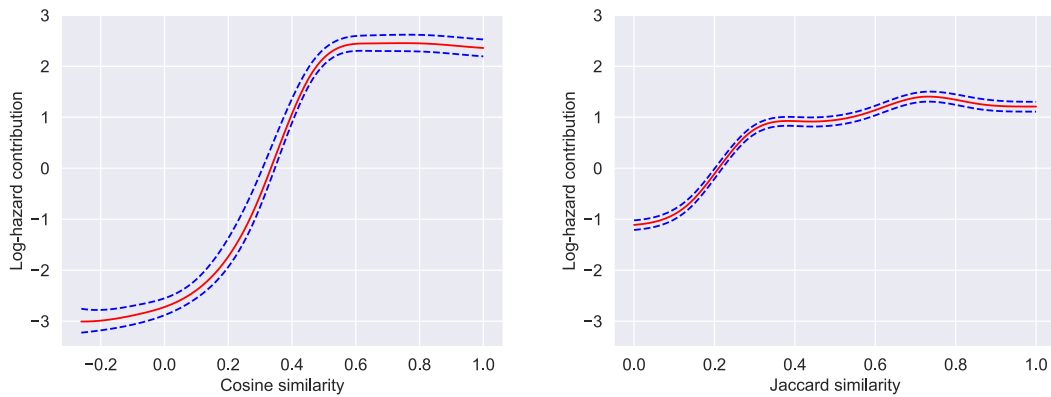


Figure 4.4: Splines associated to the similarity effects. Left: textual similarity. Right: IPC relatedness.

being cited. This result is consistent with previous research that has emphasized the importance of network position in predicting innovation outcomes [Uzzi, 1997]. Furthermore, this finding has practical implications for policymakers and inventors who may wish to increase the likelihood of their patents being cited. By fostering collaboration and networking opportunities, inventors can improve their chances of connecting to other inventors and increase their outdegree, thus increasing the visibility of their work.

Similarity effects. The curves for both *textual similarity* and *IPC relatedness* shown in Figure 4.4 demonstrate the significance of the technological similarity between the citing and cited patents. It highlights how patents that are more closely related are more likely to cite each other. The *textual similarity* curve indicates a stronger tendency towards citing patents that share linguistically similar abstracts. The *IPC relatedness* curve, on the other hand, indicates that patents that share even a limited number of technology classes have a higher probability of being cited.

Furthermore, the weight placed on the *textual similarity* effect is noteworthy. Compared to patents that share a linguistic similarity less than 0.2, patents that share a similarity larger than 0.5 are 60 times more likely to cite each other. While the *IPC relatedness* effect is not as strong as the *textual similarity* effect, it still increases the citation rate by more than 7 times, between patents that share at least 0.3 IPC classification on the Jaccard scale.

These findings confirm results from previous studies (e.g. Trajtenberg and Jaffe [2002]). Despite the structural changes over time in the technological similarity across cited and citing patents [Kuhn et al., 2020; Whalen et al., 2020], patents

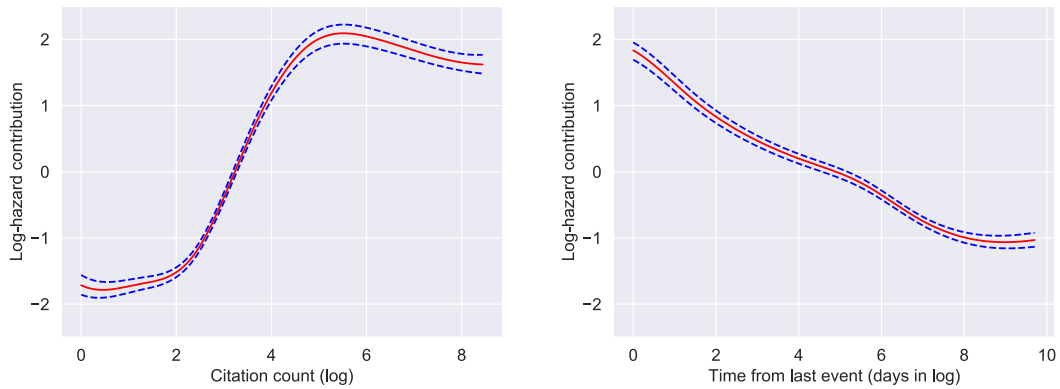


Figure 4.5: Splines associated with the time-varying effects. Left: cumulative citations received in log scale. Right: time from last event in days in log scale.

with greater technological similarity remain more likely to cite each other.

Time varying effects. The two time-varying effects in Figure 4.5 demonstrate the dynamic nature of patent citations. The *cumulative citation count* effect reveals how the number of citations a patent has received so far influences its likelihood of being cited in the future. This effect is particularly notable as the log-hazard contribution shows a rapid increase after receiving more than 20 citations, indicating a positive feedback loop where the more citations a patent receives, the more likely it is to receive additional citations. This snowball effect is a crucial factor in determining the significance of a patent within the network, and it underscores the importance of early recognition and citation of relevant breakthroughs.

On the other hand, the *time from last event* effect highlights the inverse relationship between the time interval from the last citation and the likelihood of receiving subsequent citations. As the time interval grows longer, the probability of receiving additional citations decreases. This effect is shown by the steady decrease in log-hazard contribution. This trend underlines the importance of continuous recognition of relevant patents to maintain their significance and relevance within the citation network.

It is worth noting that these two effects work in together to shape the dynamic nature of patent citations within the network. The snowball effect of the *cumulative citation count* effect can counteract the decay caused by the *time from last event* effect, but only up to a certain point. Eventually, even the most significant patents will fade in relevance if they are not consistently recognized and cited within the network.

Similarly, patents with very high continuous citation could have invariably a short time since last citation and therefore receive an additional boost from the *time from last event effect*.

An alternative explanation for patents that are very popular for a short period but then fade out of the spotlight could be highlighting the use of a different version of the indegree covariate that only considers “recent indegree”. This could be done either through a sliding window approach or with a decay mechanism, such as the one proposed by Brandes et al. [2009], that allows modeling this effect.

#### 4.4.4 Estimated baseline hazard

To analyze the overall citation rates over time, we estimate the baseline hazard by differentiating the adapted version of the Borgan et al. [1995] estimator presented in (4.3.7), using the average coefficients obtained from the repeated STREAM fits. To capture the general trend and present a clearer picture of the base hazard, we applied a Gaussian filter to the estimated baseline. Figure 4.6 shows the estimated baseline hazard, which provides a visual representation of the overall pattern of the hazard rates over the observed period.

As anticipated, the curve demonstrates that the baseline rate of being cited increases over time. The general increasing trend of the curve indicates that patents have started to cite up to 5 times more since the 1980s. This may be attributed to the accumulation of knowledge and technological advancement over time. Moreover, this result underscores the importance of considering the temporal dimension when analyzing citation patterns and provides valuable insights into the dynamics of knowledge diffusion in patent systems.

Furthermore, the estimated baseline reveals an interesting trend in the patent citation network. Specifically, we note the sudden increase in the baseline hazard in the year 2010. One possible explanation for this observed increase is the legal changes in the applicant’s duty of disclosure that took place in 2010. As reported by Kuhn [2010], these legal changes led to a drastic increase in the number and scope of cited references in patent documents. Consequently, more citations were included that were further afield from the citing patent, resulting in a generally higher rate of patent citations.

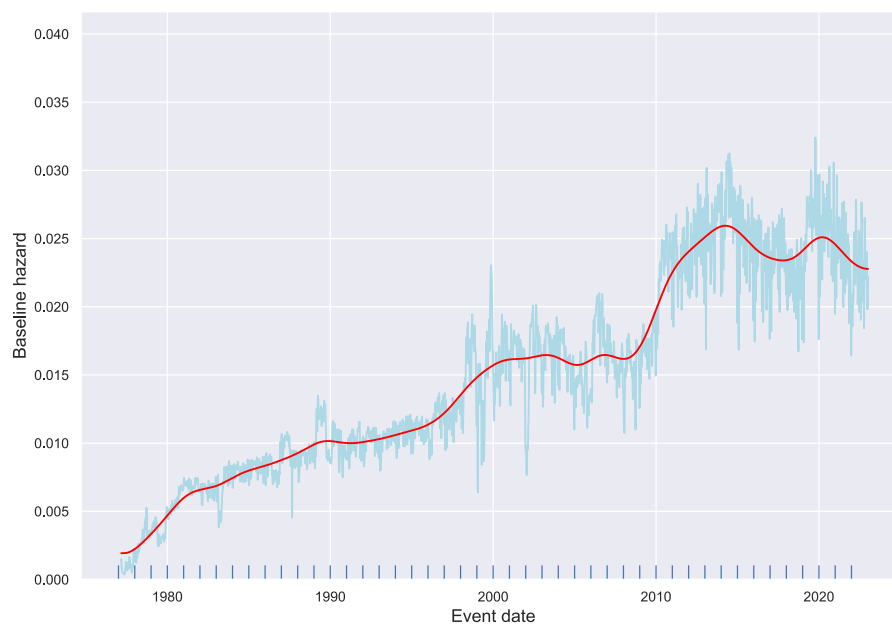


Figure 4.6: Baseline hazard estimated through the adapted estimator in (4.3.7) from [Borgan et al. \[1995\]](#). The red line is the application of a Gaussian filter to smooth the resulting estimate and capture the general trend of the baseline.

## 4.5 Conclusions

Relational event models are a sophisticated and effective approach for analyzing complex patterns in temporal network event data. In this study, we applied this framework to patent analysis to identify the drivers of patent citations. The use of REMs in studying large and intricate structures is limited by the computational complexity of modeling non-linear effects, which may result in an oversimplification of the network complex interplay of dynamic relationships. Furthermore, the inherent limitations of standard REM approaches in accommodating large datasets render them ineffective in managing the magnitude and complexity of the citation network.

To address these challenges, we introduced the Stochastic Gradient Relational Event Additive Model. This model integrates non-linear modelling with nested case-control sampling, effectively approximating the likelihood of a REM by logistic regression. By applying STREAM to a network of patent citations spanning from 1976 to the end of 2022 with over 8 million patents and over 100 million citations, we were able to identify patterns that affect the patent citation rate.

Our findings offer several interesting insights. While some effects are straightforward, others reveal peculiar patterns that require further investigation. For instance, we found that patents from around the year 2000 have been much more influential than from any other period. This suggests that there must have been several important technological innovations in those years.

There are several ways the analysis can be extended. Further research is required to assess which areas of technology have been innovating more and how this has developed through time. We could also consider a more sophisticated approach to incorporate the possible time decay of some effects. It would be interesting to evaluate the behavior of, e.g., the *textual similarity* curve in Figure 4.4 over the observed period, particularly in light of recent discussions on changes in the generative process of patent citations [Kuhn et al., 2020; Filippi-Mazzola et al., 2023]. However, such studies would require careful consideration with respect to the underlying changes in the legal patent framework. Approaches like the ones proposed by Juozaitienė and Wit [2022b] could be further investigated to be applied to the STREAM to assess the temporal decay of predictors.

In this work, the citation dynamics are modeled as a collection of dyadic interactions between patents. This is a simplification. Typically, when a citation occurs, a patent cites multiple receivers. This shows how further research could expand the current STREAM approach to modeling polyadic interactions among patents. Indeed, the flexibility of the STREAM approach could be combined with the newly proposed relational hyperevent model [Lerner and Lomi, 2023; Lerner et al.,

2023] to gain further understanding of the patent citation network's intricate dynamics.

Overall, the STREAM approach is a promising solution to overcome limitations of standard REMs in modeling complex non-linear effects in large event networks.

# Chapter 5

## Modelling Non-linear Effects with Neural Networks in Relational Event Models

The following chapter was published as:

Filippi-Mazzola, E. & Wit, E. C., (2024). Modeling non-linear effects with neural networks in Relational Event Models. *Social Networks*, 79, 25-33.

### 5.1 Introduction

Dynamic network modelling has emerged as an essential tool in social network studies, providing a nuanced perspective on the evolving nature of interactions and relationships. These networks capture the dynamism inherent in dynamic interacting structures by shedding light on how connections form, dissolve, or transform over time. Although the inclusion of the temporal dimension increases the complexity of the models, it provides richer insights, revealing patterns that static networks might miss.

Relational Event Models (REMs) [Butts, 2008; Perry and Wolfe, 2013; Bianchi et al., 2024] are an efficient and flexible framework for modelling dynamic networks, particularly in settings where events, or interactions between actors, occur sequentially over time. Unlike traditional network models, which focus on aggregated states or snapshots, REM focuses on micro-dynamics, tracking the chronological order of ties as they form or dissolve [Fritz et al., 2020]. The great versatility of REMs is underscored by the diverse fields to which they have been applied. In finance, Lomi and Bianchi [2021] and Zappa and Vu [2021] highlight that resource exchanges vary significantly across different trading conditions,

temporal contexts, and the values involved. Similarly, in healthcare, [Vu et al. \[2017\]](#) observed that patient transfers tend to form around tightly-knit hospital clusters that frequently reciprocate transfers. [Amati et al. \[2019\]](#) analyzed hospital collaborations in Southern Italy, revealing that interaction dynamics within these networks fluctuate considerably throughout the week. In ecology, [Tranmer et al. \[2015\]](#) investigated social behavioral dynamics, such as food sharing, among jackdaws, whereas [Patison et al. \[2015\]](#) used REMs to analyze adaptive behavior among cows when introduced to new herd members. In another ecological application, [Juozaitienė et al. \[2022\]](#) utilized REMs to explore the dynamics of ecological niche invasions by modelling interactions between invasive species and their new territories. Despite its easy adaptability, REM's practical applicability is limited by its computational complexity [[Welles et al., 2014](#)].

[Vu et al. \[2015\]](#) first tackled the problem by proposing various sampling strategies on the risk-set connected to the partial-likelihood denominator. [Lerner and Lomi \[2020b\]](#) demonstrated the robustness of REMs when the risk set is sub-sampled via a nested-case-control approach, demonstrating that when REMs are used to model large dynamic networks, only one control per case is sufficient to obtain reliable estimates. This sub-sampling strategy was used by [Filippi-Mazzola and Wit \[2024b\]](#) to approximate the REM partial likelihood by a logistic regression, which reduces computational complexity and allows for efficient modelling of non-linear effects.

Non-linear approaches to REMs were first tackled by [Bauer et al. \[2021\]](#) using B-splines [[De Boor, 1972](#)]. By design, these spline-based models require to storing multiple high-dimensional model matrices. When these factors are combined with large networks with millions of dyadic interactions, many REM computing frameworks suffer from both convergence and memory management issues.

Inspired by Neural Additive Models [[Agarwal et al., 2021](#)], we propose to model non-linear effects through a Deep Relational Event Additive Model (DREAM). DREAM strategically trades computational complexity with memory management by letting each effect be modelled by an independent neural network. By leveraging the higher computational power of graphic processor units (GPUs), DREAM can capture complex non-linear effects among variables. Each of these independent neural networks is trained at the same time using a Stochastic Gradient Descent (SGD) approach. SGD is especially renowned for its ability to handle large datasets and high-dimensional spaces as it iteratively refines model parameters to ensure optimal convergence. The simultaneous estimation of these neural networks not only increases computational efficiency but also ensures that interdependencies and mutual information among different effects are captured effectively.



In this chapter, we start by describing the methodological background on which REMs are built in section 5.2. After defining how DREAM is structured in section 5.3, we provide a comprehensive simulation study to test its robustness and efficiency in section 5.4. To conclude, we show an application on the analysis of the US patent citation network in section 5.5.

## 5.2 Background and Methods

REMs are a class of statistical models for sequences of social interaction events occurring over time. The primary focus of these models is to model the pattern and structure of relationships that emerge as a series of observed social interactions or events.

### 5.2.1 Relational Event Model

In REMs, the primary statistical units are a series of recorded dyadic interactions defined as events. These are denoted as  $e_i$  ( $i = 1, \dots, n$ ) and are typically represented by the triplet  $e_i = (s_i, r_i, t_i)$ , which denotes that an action was initiated by a sender  $s_i$ , targeted towards a receiver  $r_i$ , and occurred at a specific time  $t_i$ .

Following Perry and Wolfe [2013], it is possible to define a multivariate counting process  $N_{sr}(t)$  that records the number of directed interactions between  $s$  and  $r$ , up to time  $t$ ,

$$N_{sr}(t) = \sum_{i \geq 1} \mathbb{1}_{\{t_i \leq t; s_i = s; r_i = r\}}.$$

$N_{sr}(t)$  is then a local submartingale, where, through Doob-Meyer decomposition  $N_{sr}(t) = \Lambda_{sr}(t) + M_{sr}(t)$ , we can define its predictable increasing process  $\Lambda_{sr}(t)$ . REMs describe this predictable increasing process by assuming that the counting process is an inhomogeneous Poisson process, i.e.,

$$\Lambda_{sr}(t) = \int_0^t \lambda_{sr}(\tau) d\tau,$$

where  $\lambda_{sr}$  is the hazard function of the relational event  $(s, r)$ . Considering the history of prior events  $\mathcal{H}_t$  up to time  $t$ , a common method for modelling this intensity function relies on the log-linear model [Cox, 1972]. Consequently, the intensity function is expressed as the product of a baseline hazard  $\lambda_0(t)$  and an

exponential function of  $q$   $\mathcal{H}_t$ -measurable covariates  $x$ ,

$$\lambda_{sr}(t | \mathcal{H}_t) = \lambda_0(t)e^{f(x_{sr})}, \quad (5.2.1)$$

where  $f(x_{sr}) = \sum_{k=1}^q f_k(x_{srk})$  is some additive model. In the original formulation of the REM,  $f_k(x_{srk})$  is modelled as a linear function weighted by a coefficient  $\beta_k$ , such that  $f_k(x_{srk}) = x_{srk}\beta_k$ .

Given the network's prior history, the model definition assumes conditional independence of events. Incorporating covariates into this structural design allows for an in-depth investigation into a multitude of individual influences or factors that contribute to the occurrence of the event. These influential components are typically classified as exogenous or endogenous. Exogenous factors generally pertain to attributes or effects related to the sender or receiver, offering insights into external dynamics. These could include individual characteristics, roles within the network, or external circumstances that influence their actions. On the other hand, endogenous factors relate to the intrinsic micro-mechanisms within the network itself. These are patterns or tendencies that arise from the inherent structure and dynamics of the network, becoming visible as the series of events progressively unfold over time. By recognizing and studying these factors, we can gain an understanding of how the network's internal mechanisms shape the unfolding of events.

A further characteristic that lends significant appeal to the proportional hazard model is its absence of distributional assumptions concerning activity rates. This flexibility is a notable advantage over fully parametric models and enables to treat the baseline hazard  $\lambda_0$  as a nuisance parameter [Cox, 1975]. This strategic consideration helps simplify the computational complexities that emerge when trying to estimate the weights in the full-likelihood derived from (5.2.1). The resulting partial likelihood is expressed as follows,

$$L_p(\beta) = \prod_{i=1}^n \left( \frac{\exp \{f(x_{s_i r_i})\}}{\sum_{(s_i^*, r_i^*) \in \mathcal{R}(t_i)} \exp \{f(x_{s_i^* r_i^*})\}} \right), \quad (5.2.2)$$

where  $\mathcal{R}(t)$  is the risk-set, i.e., the set of all possible relational events that could have happened at time  $t$ .

### 5.2.2 Nested case control sampling

While the application of partial likelihood in (5.2.2) introduces significant simplifications to REM estimation, its practical application is constrained by the dimensionality of its denominator. The risk set  $\mathcal{R}(t)$  tends to increase quadratically with the number of nodes in traditional longitudinal networks, although it may vary depending on the specific context. For instance, in citation networks, the risk set tends to expand linearly as new nodes cite existing documents within the network [Vu et al., 2011; Filippi-Mazzola and Wit, 2024b]. However, irrespective of the individual scenarios under analysis, scalability remains a limiting factor for these models. Large networks, comprising millions of nodes, will inevitably pose computational challenges and potentially limit the model efficiency in such contexts.

A solution to this issue has been proposed by Vu et al. [2015], who introduced the idea of nested case-control sub-sampling the risk set [Borgan et al., 1995]. The central idea revolves around analyzing all the observed events, or “cases,” while only scrutinizing a smaller subset of non-events, termed “controls.” Borgan et al. [1995] demonstrated that using a nested case-control sampled risk set yields a consistent estimator. Building on this concept, Boschi et al. [2023] and Filippi-Mazzola and Wit [2024b] extended the empirical findings of Lerner and Lomi [2020b] to show that in scenarios with a large number of nodes, one control per case is sufficient to obtain reliable parameter estimates.

When only a single control per case is considered, the partial likelihood in (5.2.2) results in the likelihood of a logistic regression model where only successful outcomes are observed as responses. This key insight further enhances the practicality of REMs, reducing the computational complexity of estimating large and complex risk sets. With this transformation in place, the sub-sampled case-control version of the partial likelihood in (5.2.2) is given as,

$$\tilde{L}_p(\beta) = \prod_{i=1}^n \left[ \frac{\exp \{f(x_{s_i r_i}) - f(x_{s_i^* r_i^*})\}}{1 + \exp \{f(x_{s_i r_i}) - f(x_{s_i^* r_i^*})\}} \right], \quad (5.2.3)$$

where  $x_{s_i^* r_i^*}$  is a randomly sampled non-event for  $i$ th event sender  $s_i^*$  and receiver  $r_i^*$  from  $\mathcal{R}(t_i)$ .

## 5.3 Deep Relational Event Additive Model

Although standard REM formulations assume that the rate of interaction between  $s$  and  $r$  exhibits a linear dependence on the covariates, the relationship between predictors and event rates could be non-linear and exhibit a greater degree of complexity. If this is the case, deploying linear effects could inadvertently result in model oversimplification and the production of biased estimates. This highlights the necessity of exploring alternative modelling techniques beyond the traditional linear framework. In this section, we propose the Deep Relational Event Additive Model (DREAM) to estimate non-linear effects in large networks, that leverage machine-learning methods and graphical processor units for efficient computation.

### 5.3.1 Non-linear modelling with Neural Networks

Modelling non-linear effects in REMs was first tackled by [Bauer et al. \[2021\]](#) and [Filippi-Mazzola and Wit \[2024b\]](#). Both heavily rely on the use of B-splines [[De Boor, 1972](#)], a computational technique that represents curves as a series of interconnected piecewise polynomial functions. While splines are a standard tool in non-linear, additive modelling, their implementation comes with challenges in big data settings, especially with respect to memory: the fitting procedure necessitates the creation of a potentially huge model matrix.

DREAM strategically trades off memory usage with computational complexity. Following the recent developments of Neural Additive Models [[Agarwal et al., 2021](#)], DREAM leverages multi-layered neural networks to model non-linear effects, where each effect is modeled by an independent neural network. Let then  $f_k$  be a feed-forward Artificial Neural Network (ANN) [[Ripley, 1996](#)] with a single input and a single output, separated by  $L$  layers, for  $l = 1, \dots, L$ . Each of these layers contains  $m_1, \dots, m_L$  neurons. The output of each  $f_k$  is the result of a series of sequential operations, such as

$$\begin{aligned}
 a_k^{(1)} &= \phi(\beta_k^{(1)} x_{srk} + \beta_{0k}^{(1)}) \\
 a_k^{(2)} &= \phi(\beta_k^{(2)} a_k^{(1)} + \beta_{0k}^{(2)}) \\
 &\vdots \\
 a_k^{(L-1)} &= \phi(\beta_k^{(L-1)} a_k^{(L-2)} + \beta_{0k}^{(L-1)}) \\
 a_k^{(L)} &= \beta_k^{(L)} a_k^{(L-1)} + \beta_{0k}^{(L)},
 \end{aligned}$$

where  $a_k^{(l)}$  represents the output of the  $l$ -th layer for the  $k$ -th covariate,  $\beta_k^{(l)}$  is the weight matrix of size  $m_l \times m_{l-1}$ , and  $\beta_{0k}^{(l)}$  is the intercept or *bias* of size  $m_l \times 1$ .  $f_k$  is then an ANN with  $M = \sum_{l=1}^L (m_l \times m_{l-1} + m_l)$  number of parameters.  $\phi$  is a non-linear function, commonly referred as activation function. ANN models use a non-linear function to be able to model complex relationships in data, in a way that a linear function cannot. Noel et al. [2023] recently proposed the Growing Cosine Unit (GCU) activation function, as a way to deal with some of the drawbacks of standard activation functions. In our empirical evaluation, this function seems to perform well. A more detailed discussion on the activation function can be found in B.1.  $f_k$  can then be expressed as

$$f_k(x_{srk}) = \beta_k^{(L)} \left( \dots \phi \left( \beta_k^{(2)} \left( \phi \left( \beta_k^{(1)} x_{srk} + \beta_{0k}^{(1)} \right) \right) + \beta_{0k}^{(2)} \right) \dots \right) + \beta_{0k}^{(L)}. \quad (5.3.1)$$

Let then  $f(x_{sr})$  represent the collective sum of  $q$  independent ANN output effects for  $x_{srk}(t)$ , with  $k = 1, \dots, q$ , i.e.,  $f(x_{sr}) = \sum_{k=1}^q f_k(x_{srk})$ . Each of these ANNs is then trained simultaneously to maximize (5.2.3). Figure 5.1 describes the structure of how the information is passing through  $f(x_{sr})$ , offering a clear understanding of the DREAM framework. The most significant asset of this modelling technique lies in its interpretability. Through a visual examination of the individual functions  $f_k$ , one can develop a comprehensive understanding of the dynamic behavior of each effect  $x_{srk}$ , mimicking the interpretability of splines. Although this technique increases the computational complexity for evaluating the likelihood in (5.2.3) as the passage from one layer of the network to another requires multiple matrix multiplications, it eliminates the heavy memory usage associated with basis transformations. While modern frameworks allow efficient matrix operations, the efficiency of this approach is mainly guaranteed by recent technological advancements in the application of vectorized computations on GPUs.

Like most common machine-learning techniques, DREAM scalability in the estimation process is attained thanks to the adoption of a Stochastic Gradient Descent (SGD) approach. SGD is particularly effective for large datasets and complex models, as it updates model parameters iteratively based on subsets (batches) of data, rather than the entire dataset.

Among the available SGD methodologies, we use the ADAM optimizer [Kingma and Ba, 2017]. ADAM is easily scalable and has reliable convergence [Reddi et al., 2018]. Its computational merits are due to the way this approach updates the weights of the model, by calculating individual adaptive learning rates based on estimates of the first and second moments of the gradients. Details on ADAM can be found in the B.2.

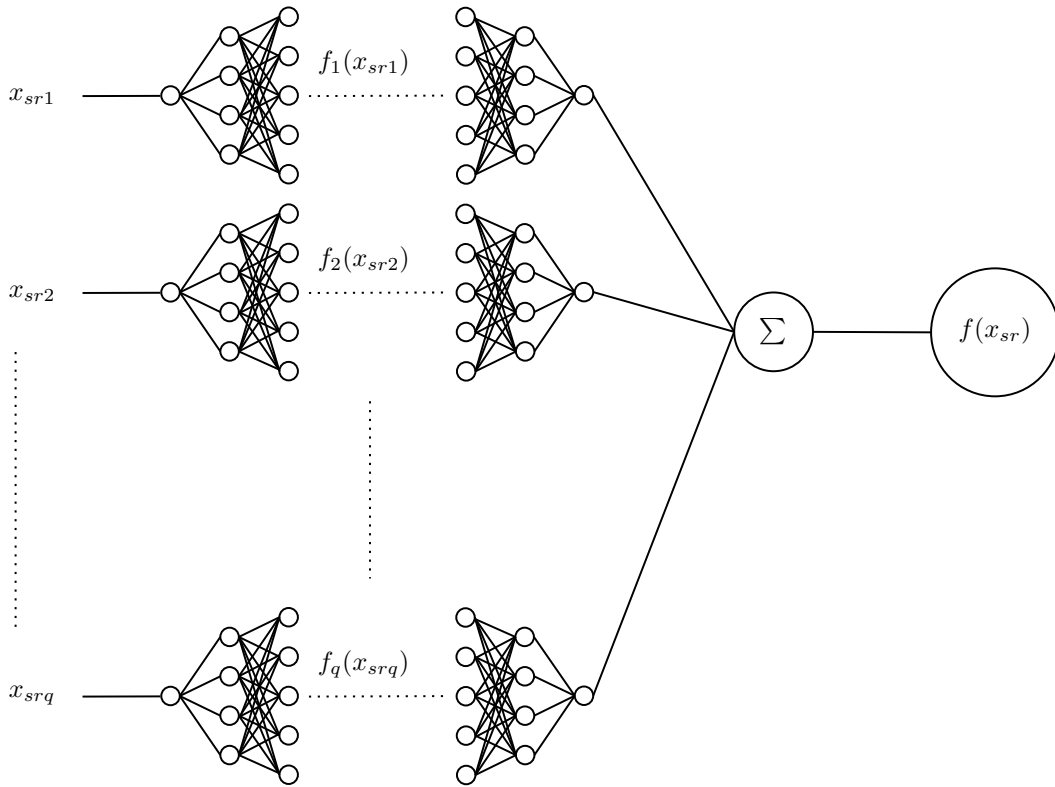


Figure 5.1: DREAM framework. Each effect is modeled via an independent ANN structure that captures its non-linearity. It is possible to extend this modelling technique with interaction effects simply by letting  $f_k$  have a multivariate input, such as  $f_k(x_{srk_1}, x_{srk_2})$ , and a univariate output.

While DREAM's flexibility allows for different regularization techniques to be used, in our modelling approach we used dropout [Srivastava et al., 2014] to prevent overfitting during the estimation process. By randomly omitting subsets of features or neurons during the training phase, dropout helps improve the robustness and generalizability of the neural network. Together with the number of layers, and the number of neurons per layer, dropout constitutes one of the main hyperparameters of the model to infer.

### 5.3.2 Uncertainty estimation with Gaussian Process Regression

Due to its over-parametrization, estimation of ANN hyper-parameters  $\{\beta_{0k}^{(l)}, \beta_k^{(l)}\}_{k,l}$  via SGD does not imply achieving convergence to a global optimum. As discussed in Goodfellow et al. [2016], neural network optimization often focuses on finding satisfactory parameters that perform well, rather than continuing the estimation

process until a theoretical optimum is reached. This approach recognizes the complex, high-dimensional landscapes in which these models operate, where multiple parameter sets can yield comparably effective outcomes. Consequently, the specific parameters of a neural network should not be interpreted in isolation since their values can vary across different runs of the model without necessarily affecting performance. Instead, our attention focuses on the behavior of the estimated functions  $\hat{f}_k$ , which provide meaningful insights into the data relationships modeled by the ANN. Consequently, our focus on uncertainty assessment pertains to the estimated functions, rather than the hyper-parameters.

A common solution for non-parametric models is to employ non-parametric bootstrap [Efron, 1979]. Such procedure estimates pointwise uncertainty intervals by generating a sufficient number of sampled copies from the original dataset. For each repetition, the model is re-trained on a new “bootstrapped” version of the original dataset. Consequently, each re-train yields distinct estimates of the non-linear effects. Once a sufficient number of repetitions are completed, confidence bands can be directly computed from the estimated curves using the desired percentiles.

While bootstrapping offers a flexible way to evaluate estimation uncertainty, it is computationally demanding, particularly when considering the training requirements of neural networks. To address this, we propose to post-process a limited number of bootstrap refits via Gaussian Processes Regression (GPR) [Rasmussen and Williams, 2006] to obtain robust confidence bands.

We assume that the estimated function can be represented by a Gaussian process  $\hat{f}_k \sim \mathcal{GP}(f_k, K_k)$  where  $f_k$  represents the true mean function of the Gaussian Process, while  $K_k$  is a Radial Basis Function (RBF) kernel with the form

$$K_k = \exp\left(-\frac{\|x - x'\|}{2l^2}\right), \quad (5.3.2)$$

where  $l$  is a scale parameter, while  $x'$  is a subsequent value of  $x$ . Computing the posterior mean and covariance matrix, we obtain an estimated mean function whose uncertainty is represented by the standard deviations computed by square rooting the diagonal of the posterior covariance matrix. The  $O(n^3)$  computational complexity of GPR models presents a limitation for large-scale applications. This higher computational cost is due to the inversion of the kernel matrix in the posterior. However, to obtain pointwise estimates of the curves, it is sufficient to evaluate the kernel matrix on a reduced sample of equidistant points on the support range of  $x$ . Knowing the upper and the lower bound of each covariate, we can generate a vector  $\tilde{x}$  of equidistant points. We can then

define  $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N)$ , where  $\tilde{x}_i = x_{\min} + (i - 1)\Delta$  for  $i = 1, \dots, N$ , with  $\Delta = \frac{x_{\max} - x_{\min}}{N-1}$  and  $N \ll n$ . With  $\tilde{x}$ , we can compute the posterior estimates of the Gaussian Process,

$$\hat{\mu}(\tilde{x}) = K_k^\top [K_k + I]^{-1} \hat{f}(\tilde{x}), \quad (5.3.3)$$

$$\hat{\Sigma}(\tilde{x}) = K_k - K_k^\top [K_k + I]^{-1} K_k, \quad (5.3.4)$$

where  $I$  is the identity matrix that improves numerical stability during inversion. While in many scenarios it is possible to scale this identity matrix by multiplying  $I$  to a constant term, we have noticed in our applications that the results remain roughly invariant to such scaling. Overall, this alternative approach is designed to offer a more computationally efficient alternative to a full bootstrap approach.

## 5.4 Simulation study

In this section, we present a series of simulation studies that highlight DREAM's ability in accurately identifying non-linear effects in dynamic networks. We focus on three specific aspects. The initial simulation study illustrates DREAM's ability to reconstruct the true generating functions behind observed effects. Secondly, given their close similarity, it is natural to compare additive Neural Network models with spline-based additive models such as Generalized Additive Models (GAMs). In our comparison, we use the `gam` function from the `mgcv` package in R. Finally, we present a study on the time-complexity of our method. The full code has been written in Python within the Pytorch suite [Paszke et al., 2019] and it is publicly available in a GitHub repository (<https://github.com/efm95/DREAM>) together with all the simulations and applications.

### 5.4.1 True function recovery

We simulated relational event data under the assumption that each node possesses both a sender and a receiver covariate simulated as uniform variables  $\mathcal{U}(0, 1)$ . The effect of each covariate is given in red in Figure 5.2. A network with 5,000 nodes and 500,000 edges is sampled. The fitted curve was estimated via five bootstrap refits, followed by the application of a GPR model using the `scikit-learn` Python library [Pedregosa et al., 2011]. The ANN architecture was determined via CV (details can be found in B.5, while formulas of the true generating functions can be found in B.4).



The results show how the estimated curves follow the true generating functions behavior, while the estimated confidence intervals obtained through this approach encompass the true functions over the variables support.

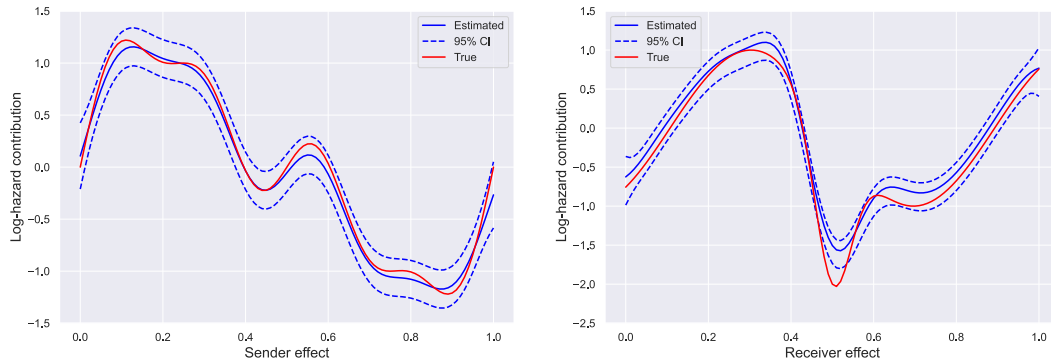


Figure 5.2: True and estimated effects along with their confidence intervals. The red lines denote the actual effects, whereas the blue lines are the estimated effects, and the dashed-blue lines represent the confidence intervals. Confidence intervals are calculated by adding and subtracting twice the local standard error from the estimated functions.

#### 5.4.2 Accuracy comparison with GAM

As previously noted, non-linear effects in a REM can be estimated by using a logistic regression additive model approach. A computationally efficient choice is the `gam` function within the `mgcv` package in R, where the smooth terms are estimated via penalized b-splines. The great advantage of `mgcv` is that the degrees of freedom of the splines are automatically selected, thus reducing the number of hyperparameters that are required to be set.

Comparing models using traditional information criteria such as AIC or BIC may not provide a fair assessment in the context of ANNs. This is because ANNs incorporate a substantially larger number of parameters compared to GAMs or other classical statistical models. To address this challenge, we adopt an alternative metric for model comparison. Specifically, in Table 5.1, we report the maximized log-partial-likelihood values obtained by DREAM and a GAM, alongside the log-partial-likelihood computed from the sampled population with the true generating functions. By using the Kullback–Leibler (KL) divergence assessed with the sampled population, we provide a comparison that considers model fit and the ability to accurately reconstruct the true generating functions.

	GAM	DREAM	Population
$\log \tilde{L}_p$	-232'991.59	-232'102.28	-231'634.90
$KL(\text{Pop.}  \text{Model})$	1794.71	1647.98	-

Table 5.1: Log-partial-likelihood values for each estimated compared with the one computed from the sampled population using the true generating model. Subsequently, the KL-divergence values are assessed in relation to this same sampled population.

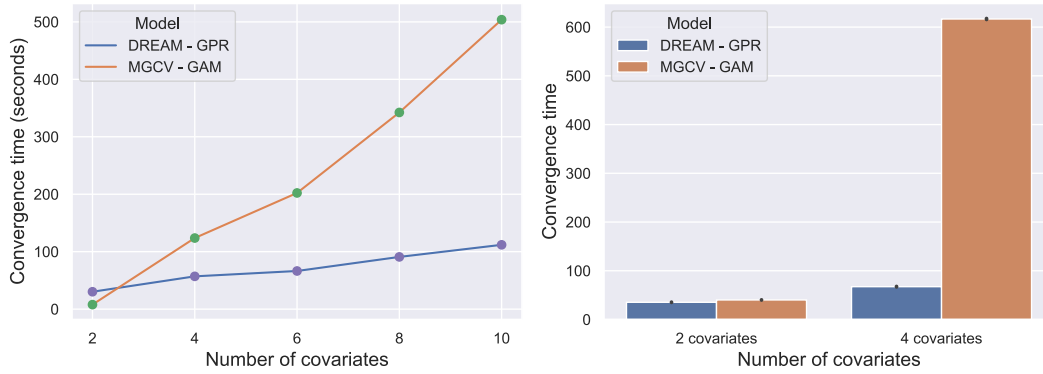
From the results in Table 5.1, it is possible to notice DREAM attains a log-partial-likelihood score that more closely attains the one of the sampled population. As a consequence, this is also reflected in its KL-divergence scores. For this scenario, GAM with smooth terms is slightly outperformed. However, the proximity of the performances between DREAM and GAM suggests that both models offer a similar level of accuracy in approximating the true model.

### 5.4.3 Time efficiency comparison with GAM

Estimation in the `mgcv` package uses highly optimized Newtonian solvers, written in C routines to achieve rapid convergence. However, the computational efficiency of these solvers is frequently undermined by R less-than-optimal memory management system [Kotthaus et al., 2015]. This results in computational bottlenecks, prolonging the time required for the algorithm to converge. In some instances, the inefficiency in memory management can even lead to computational overflow, further complicating the estimation process.

In contrast, DREAM relies on the PyTorch suite [Paszke et al., 2019], a deep learning framework that excels in handling vectorized operations. PyTorch is specifically designed to leverage the computational capabilities of Graphics Processing Units (GPUs). Using Google Colab free GPUs (Nvidia Tesla T4 with 15GB of memory), we run two sets of simulations to compare `mgcv` convergence times with DREAM with the implementation of the GPR approach to estimate the curves.

It is important to consider that the convergence time of DREAM should not be evaluated in isolation, as it depends on various hyperparameters such as the learning rate and number of epochs. To address this, we always fitted DREAM with the default learning rate of 0.001, and we employed an early stopping technique to stop the training process. The convergence timings presented in this section include not only the duration required to achieve a convergence but also the multiple iterations needed to fit the Gaussian process for uncertainty estima-



(a) 100'000 events with 1'000 actors.

(b) 500'000 events with 5'000 actors.

Figure 5.3: Comparison of convergence times between MGCV and DREAM across two simulated REM data scenarios.

tions. Figure 5.3a compares the convergence times of mgcv and DREAM using generated REM data that comprised 1'000 nodes and 100'000 events. We gradually augmented the complexity by adding sequentially covariates, thus increasing the number of non-linear effects each model needed to estimate. We carried out the fitting procedure ten times. While mgcv convergence time initially appeared faster with only two covariates, its performance rapidly degraded as the complexity grows, revealing the computational bottleneck within R. Conversely, DREAM exhibited only a modest uptick in convergence time as the complexity increased. Figure 5.3b presents for a larger dataset comprising 5'000 nodes and 500'000 events. While the C routines lend mgcv stability in its convergence times, it becomes noticeably strained by including 4 covariates, taking considerably longer. It is to be noted that with this data size, we could not fit a model with more covariates as the algorithm failed to converge.

## 5.5 US patent citation network

To demonstrate DREAM practical applicability on large networks, we model non-linear effects in the US patent citation network that contains nearly 100 million citations and almost 8 million patents from 1976 to 2022. We chose this specific application not only because of its size and complexity, but also because the data preprocessing procedures and the computation of the statistics are well-defined [Filippi-Mazzola and Wit, 2024b], making the study more accessible and simple to replicate. The preprocessing of the patent citation network can be found at

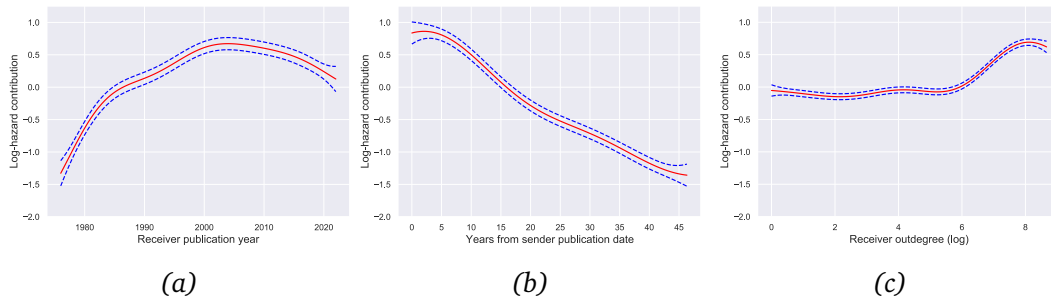


Figure 5.4: Nodal effects consisting of a receiver publication year effect (a), a time-difference effect (b) and a receiver outdegree effect (c).

<https://github.com/efm95/STREAM>. A detailed model selection for the application to the US patent citation network is extensively covered in B.5.

In order for a patent to be formally issued, the applicant must disclose all relevant prior art. As a result, the US patent citation network consists of patents that cite earlier works in relation to their issuance date. This results in a dynamic network that is constantly growing and expanding. Within this network, nodes are represented by patents, and as they are published, they establish connections to pre-existing nodes in the network via citations. This modelling exercise aims to identify what drives a patent  $s$  to cite a patent  $r$  at time  $t$ .

Filippi-Mazzola and Wit [2024b] proposed to model the network via three different set of statistics: patent effects, patent similarity effects, and endogenous temporal effects. The first set of effects is portrayed in Figure 5.4 and consists of the receiver publication year, the time-difference between the sender issue date and the receiver publication date, and the receiver outdegree. Figure 5.4a shows a maximum around the year 2000. Potentially, this indicates that increased technological innovation happened during that time. The time-difference effect identifies a period of approximately 5 years following the patent publication date when citations are most likely. Finally, the receiver outdegree confirms that patents with a higher number of citations at the time of publication tend to play a more central role in the network, consequently increasing their chances of accumulating more citations over time.

The second set of effects, shown in Figure 5.5, delves into the patent similarity characteristics that contribute to a citation. The first statistic is textual similarity. We embed patent abstracts in an Euclidean space using a pre-trained SBERT model [Reimers and Gurevych, 2019], and calculate pairwise cosine similarities. The resulting non-linear effects conclusively demonstrate that patents are more likely to cite each other when their abstracts share significant textual similarities.

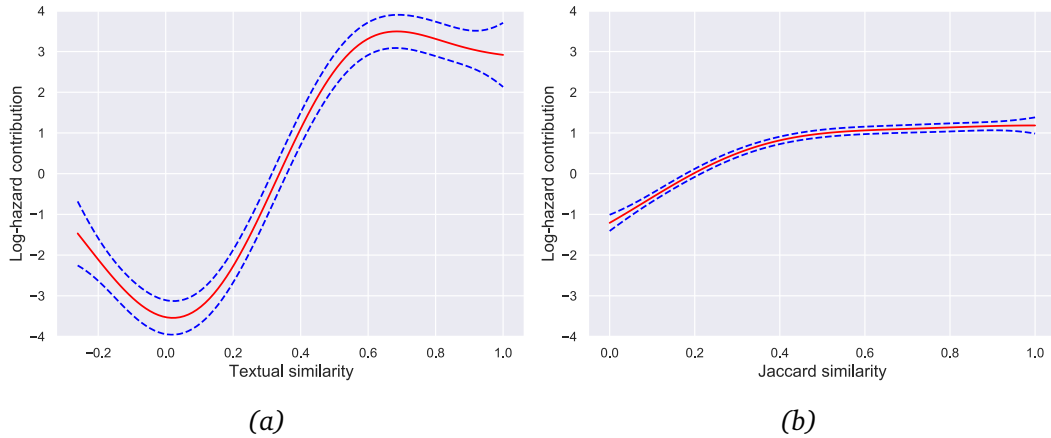


Figure 5.5: Similarity effect consisting of the textual similarity effect (a) and the technological relatedness effect (b).

Specifically, the hazard of a citation occurring is 60 times higher for patents who share a textual similarity larger than 0.5 when compared to those who share a similarity of 0.2. Secondly, we consider the technological relationship between the two patents, as indicated by their shared International Patent Classification (IPC) classes. We capture the proportion of shared classes to the total classes observed across both patents by computing the Jaccard similarity among these IPC classes. According to the results Figure 5.5b, the rate of one patent citing another increases as the number of technology classes increases. Indeed, when comparing citations with a Jaccard index of 0.4 to those with 0, the hazard rate increases by about 7 times.

Figure 5.6 captures time-varying factors that influence the rate of a patent being cited. The first of these is the cumulative citations a patent has received, illustrating that as a patent accumulates more citations, its probability of receiving additional ones increases, until it reaches a plateau around  $e^{5.4} \approx 221$ . The second effect evaluates the time elapsed since a patent most recent citation. This indicates that the longer the duration since the last citation, the less probable it becomes for the patent to be cited again.

For an alternative interpretability of the model's outputs, section B.6 includes figures that present the effects of our fitted effects expressed in terms of Hazard contributions rather than Log-hazard contributions.

Following the classification system proposed by Jake Olivier and Bell [2017], the contributions of effects to the hazard can be categorized into *large*, *medium*, or *small* based on their hazard ratio sizes. Observations from our analysis reveal significant variations in hazard contributions across the support of these effects.

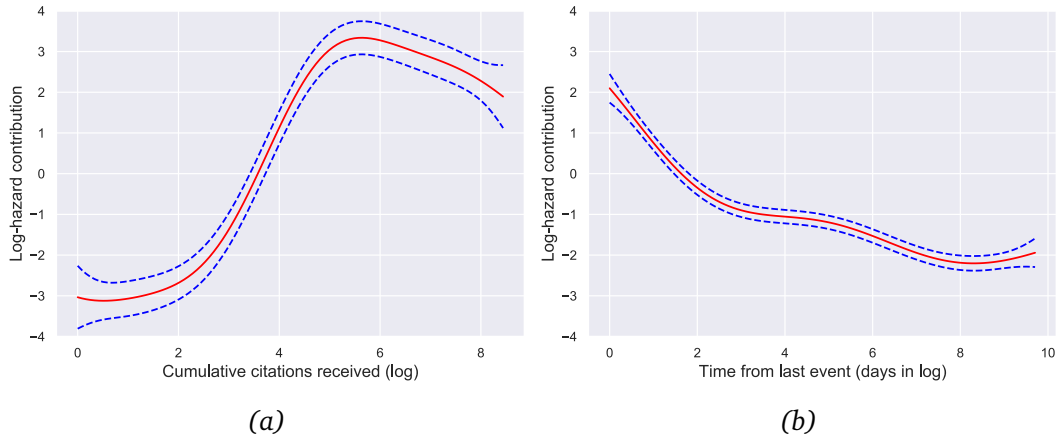


Figure 5.6: Time-varying effects consisting of the cumulative citations received (a) and the time from last event (b).

Most notably, the majority of these hazard ratio sizes fall within the *medium* or *large* categories. This indicates not only the substantial impact of these effects on the model’s outcomes but also underscores their relevance in describing the overall dynamics of the patent citation network.

## 5.6 Conclusions

Relational Event Models offer a versatile framework for modelling dynamic networks. Yet, their real-time application often faces challenges due to the computational complexity in their fitting procedures. Such challenges tend to amplify as the volume of observed events increases. In this study, we present a solution to these computational issues by introducing the Deep Relational Event Additive Model (DREAM). In DREAM, the non-linear behavior of each covariate is captured by an independent Artificial Neural Network, providing both precision and efficiency in capturing network dynamics. We proposed two distinct methods in DREAM for estimating areas of uncertainty. The first method entails non-parametrically bootstrapping the observed dataset and then refitting the model multiple times. The second method, employs Gaussian Process Regression based on a small subset of non-parametric bootstrap refits, offering a more efficient way to handle uncertainty while maintaining robustness in our estimations.

Throughout a series of simulation studies, we introduced and tested the capabilities of DREAM, emphasizing its ability in capturing non-linear effects within large dynamic networks. The robustness and efficiency of DREAM became clear

when compared to existing methods such as GAMs from the MGCV package in R. DREAM strength lies not only in its ability to accurately model nonlinear effects, but also in its fast convergence, which is accomplished by leveraging the computational advantages of Pytorch and GPUs. We further demonstrated the practical significance of DREAM by modelling a patent citation network, which encompasses nearly 100 million events and about 8 million actors.

In our study, we have not addressed the complex challenge of assessing model fit for REMs, particularly when applied to large-scale and complex datasets. The metrics employed, like KL divergence, while effective under controlled simulation conditions, fall short in empirical scenarios where the true underlying function is unknown. Consequently, we used a held-out set and cross-validation techniques in both our simulation study and our empirical setting, to evaluate the fit of our chosen model. However, this approach falls short when discerning which non-linear relationships are most accurately represented. This underscores the importance of methodological innovation in the area of model evaluation in REMs to keep pace with the evolving complexity of data structures and analysis techniques in social network research.

DREAM not only offers an efficient and scalable approach to analyzing longitudinal networks and capturing complex non-linear effects, but it also offers remarkable flexibility to customize model complexity. This adaptability includes compatibility with traditional regularization methods like dropout, ridge, and lasso. As speculated in section 5.3.1, DREAM has the potential to be expanded to capture complex non-linear interactions among covariates. Currently, our architecture processes each covariate separately. However, future iterations of this model could explore the implementation of a single, wider neural network that processes the entire  $q$ -dimensional covariate vector  $(x_{sr1}, \dots, x_{srq})$ . This approach would directly transform the vector into  $f(x_{sr})$ , potentially enhancing the model's ability to capture and interpret higher-order interactions without the combinatorial increase in complexity seen with models that estimate functions on pairs of covariates. Such a configuration would not only simplify the architecture but could also offer more profound insights by varying covariate combinations systematically, fixing others at their means to isolate effects. The inherent capability of neural networks to manage high-dimensional inputs suggests that this could be a feasible and valuable direction for future research, particularly as it might overcome the limitations associated with more traditional methods like B-splines in modelling complex interactions within large datasets. Moreover, DREAM flexibility allows it to be easily adapted to address multi-cast interactions [Perry and Wolfe, 2013] and further extended to model hyperedges [Lerner and Lomi, 2023].

## Chapter 6

# Analyzing Non-linear Network Effects in the European Interbank Market

The following chapter was published as:

Filippi-Mazzola, E., Bianchi, F., & Wit, E. C. (2024). Analyzing non-linear network effects in the European interbank market. In *2024 11th IEEE Swiss Conference on Data Science (SDS)*, IEEE, 16–22.

### 6.1 Introduction

Financial markets are complex and constantly evolving ecosystems. Unlike traditional markets, where roles and interactions follow predictable patterns, financial markets present a unique landscape where the roles of participants and the nature of transactions are constantly changing [Cetina and Preda, 2004]. This fluidity is not just a feature of these markets, it is the very essence that drives their operations and influences global economies [Beckert, 2010]. The fast-paced environment of financial markets, driven by high-frequency transactions and rapid shifts in roles and relationships, offers a fertile ground for examining the underlying principles of market behavior [Preis et al., 2011]. Here, traditional sociological theories intersect with real-world financial data, presenting opportunities for novel insights and a deeper understanding of market dynamics [Carruthers and Kim, 2011]. Our focus narrows to the European interbank market, a critical component of the global financial system. The interbank market is key in maintaining liquidity and financial stability across nations [Gabbi et al., 2013]. The importance of this market extends beyond regional financial stability. It is instrumental in the execution and efficacy of monetary policies, especially those enacted by the European Central Bank. Moreover, it serves as a barometer for



the health of the broader European economy, reflecting underlying trends and potential financial risks.

The Relational Event Model (REM) [Butts, 2008; Bianchi et al., 2024] is a family of statistical models that are particularly advantageous for modeling dynamic networks due to their ability to capture the temporal evolution of interactions. These models excel in analyzing complex, high-frequency data by accounting for the sequence and timing of events, which is essential in understanding network dynamics [Fritz et al., 2020]. REMs also allow for the integration of various covariates and actor attributes. Additionally, they address endogeneity by considering dependencies among subsequent events, thereby offering an accurate and intuitive understanding of network evolution. The great versatility of REMs is emphasized by the various fields they have been applied to, including finance [Zappa and Vu, 2021], healthcare [Vu et al., 2017; Amati et al., 2019], and ecology [Juozaitienė et al., 2022; Boschi et al., 2023]. Similarly to Iori et al. [2015], we propose to analyze the European interbank market by using a network-based approach. While this methodology is not new to the field [Temizsoy et al., 2015], we propose to expand the study of Lomi and Bianchi [2021], which is rooted in the concepts of REMs, by modeling a series of network effects using a novel non-linear approach. Although the dimensionality of the networks that can be analyzed has always been the drawback of this modeling technique due to their run-time complexities [Welles et al., 2014], newer models have recently overcome such issues [Lerner and Lomi, 2020b; Filippi-Mazzola and Wit, 2024b]. As opposed to other parametric alternatives [Bauer et al., 2021], the Deep Relational Event Additive Model (DREAM) [Filippi-Mazzola and Wit, 2024a], inspired by the recent Neural Additive Model [Agarwal et al., 2021], models complex non-linear effects by aggregating the output of multiple independent neural networks. Thus, by leveraging the computational power of Graphical Processing Units (GPUs), thereby enhancing the efficiency and scalability of REMs in analyzing large complex network dynamics. Another major advantage of this non-parametric alternative that stands out from the original REM is the ability of DREAM to model non-linear effects. This allows to capture behavior that linear models might oversimplify, thereby providing a more realistic representation of the underlying dynamics of the European interbank market. We start in section 6.2 by describing the empirical setting of the European interbank market, together with a description of the data. After a description of the methodological background on which REMs are built in 6.3, we proceed to define how DREAM is structured. We then conclude with a description of the fitted effects shown in 6.4, together with some speculation on what caused these results.

## 6.2 The European Interbank Market

The European interbank market serves as a critical platform for financial transactions between banks, facilitating the borrowing and lending of funds on a short-term basis. It operates under the broader regulatory and monetary framework established by the European Central Bank, which sets key interest rates and provides liquidity to the system. National central banks act as intermediaries, implementing ECB policies at the local level and ensuring smooth operation within their respective countries. Overall, this market has two main objectives [Gabrieli, 2011]. First, it addresses financial institutions' short-term liquidity needs, managing both anticipated and unexpected liquidity imbalances. Second, it assists banks in meeting the European Central Bank's reserve requirements, thereby promoting effective liquidity management strategies.

The relevance of the European interbank market in the global economy cannot be understated. Interbank rate fluctuations have immediate and far-reaching consequences, influencing borrowing conditions for both businesses and households [Gabbi et al., 2013]. These rates are integral to derivative contracts such as interest rate swaps or short-term interest rate futures, commonly utilized by financial institutions to guard against fluctuations in short-term interest rates. Therefore, a smoothly operating interbank market is a prerequisite for central banks to effectively manage liquidity, control interest rates, and implement monetary policy.

The interbank market is instrumental in reallocating liquidity that was originally supplied by national central banks [Wiemers and Neyer, 2003]. This reallocation process was vital to provide heterogeneous market participants with accessible liquidity. Indeed, borrowing from the central bank has different costs for credit institutions, depending on their ability to provide adequate collateral. In contrast, the interbank market does not impose the cost of holding eligible assets. The e-MID market, as the sole electronic market for interbank deposits in the Euro area, exemplifies this. It functions as a multilateral screen-based market where registered banks can transfer assets electronically through dedicated credit lines.

The e-MID market offers a wide range of credit contract maturities, extending from overnight (which constitutes approximately 85% of the transactions) up to one year. This platform distinguishes between regular and large transactions based on the transaction amount. Transactions are categorized as "regular" for amounts as low as EUR 0.05 million and "large" for transactions of EUR 100 million or more. The platform's design ensures transparency. Trades are publicly visible in terms of duration, amount, rate, and time. Participants can observe all

negotiations on the platform, enhancing their understanding of market dynamics.

Quoters and aggressors are the two roles that facilitate communication through the market interface, with the e-MID being a quote-driven market. This setup allows market participants to express their interest in trading openly, either seeking or offering capital, thus enhancing market efficiency.

The e-MID market records transactions in time-stamped datasets in which each line reports the distinctive features of the corresponding credit contract. These transactions occur frequently, with their time stamps being precise to the second. Similar to [Lomi and Bianchi \[2021\]](#), we have collected data from the e-MID trading platform, generating a dataset in which each entry is a time-stamped transaction that includes public information such as its duration, exact time, and the amount involved (in millions of EUR). The data is then composed of 1,468,463 overnight credit transactions from 355 credit institutions across 16 European countries recorded from January 4, 1999, to December 31, 2015. The time window we analyzed encompasses significant market events. Among these, this work focuses its analysis on the effects of the 2008 financial crisis. To capture the distinct network dynamics before and after this pivotal event, the observation period was divided into two phases: pre-crisis and post-crisis, with the separation date being September 15, 2008. This division allows for an in-depth examination of how the interbank network's behavior and interactions evolved due to the crisis.

## 6.3 Background and methods

### 6.3.1 The Relational Event Model

In Relational Event Models (REMs) [[Butts, 2008](#); [Perry and Wolfe, 2013](#)], interactions between pairs are called events and are the fundamental units of analysis. Each event, represented as  $e_i$  (where  $i = 1, \dots, n$ ), is typically expressed by the triplet  $e_i = (s_i, r_i, t_i)$ , where an initiator  $s_i$  directs an action towards a recipient  $r_i$  at a certain time point  $t_i$ .

Following [Perry and Wolfe \[2013\]](#), we define  $N_{sr}(t)$  as a multivariate counting process that counts the number of directed interactions between  $s$  and  $r$ , up to time  $t$ ,

$$N_{sr}(t) = \sum_{i \geq 1} \mathbb{1}_{\{t_i \leq t; s_i = s; r_i = r\}}.$$

$N_{sr}(t)$  is then a local submartingale. Through Doob-Meyer decomposition, this

counting process can be expressed as  $N_{sr}(t) = \Lambda_{sr}(t) + M_{sr}(t)$ , where  $\Lambda_{sr}(t)$  is the predictable increasing process. Then, it is possible to define a predictable continuous process  $\lambda_{sr}$  such that

$$\Lambda_{sr}(t) = \int_0^t \lambda_{sr}(\tau) d\tau.$$

The process  $\lambda_{sr}$  represents then the stochastic intensity function of  $N$ , i.e., the hazard function of the relational event  $(s, r)$ . Considering the history of prior events  $\mathcal{H}_t$  up to time  $t$ , a common method for modeling this intensity function relies on the log-linear model of Cox [1972]. Consequently, the intensity function is expressed as the product of a baseline hazard  $\lambda_0(t)$  and an exponential function of  $q$  covariates  $x$ ,

$$\lambda_{sr}(t | \mathcal{H}_t) = \lambda_0(t) e^{\sum_{k=1}^q f_k(x_{sr})}, \quad (6.3.1)$$

where  $f_k(x_{sr})$  is a function that maps coefficients with network data. Given the network's prior history, the model assumes that events occur independently of each other. Eq. (6.3.1) accommodates various covariates, which can be both exogenous and endogenous.

Exogenous covariates are typically associated with characteristics or influences of the initiator or recipient of the action, shedding light on the external dynamics at play. Such factors encompass personal attributes, roles held within the network, or external situations impacting their behavior. Conversely, endogenous factors are concerned with the innate social mechanisms inherent to the network. These encompass emergent patterns or propensities that emerge directly from the sequence of relational events. Examples of endogenous covariates are repetition of past events, reciprocity, and transitivity.

The fundamental information associated with relational events  $\{e_i : i = 1, \dots, n\}$  is captured within the full likelihood function from (6.3.1). In REMs, this likelihood is formulated as a combination of the conditional generalized exponential event time densities along with their corresponding multinomial relational event probabilities. The task of estimating the parameters of REMs by optimizing the full likelihood presents multiple challenges, particularly in its complex definition. Indeed, the likelihood function is sophisticated as it involves the direct integration over the unknown risk function and aggregating across the extensive set of potential events that might have happened at  $t$ , known as the risk set.

The proportional hazard model [Cox and Oakes, 1984] presents a compelling option compared to fully parametric models, primarily because of its absence

of specific distributional assumptions about activity rates. Consequently, these are instead treated as nuisance parameters. This approach significantly simplifies the complete REM likelihood by employing the concept of partial likelihood [Cox, 1975] to counting processes occurring on network edges. The resulting likelihood focuses solely on the probabilities of multinomial events, i.e.

$$L_P(\beta) = \prod_{i=1}^n \left( \frac{\exp \left\{ \sum_{k=1}^q f_k(x_{s_i r_i k}) \right\}}{\sum_{(s_i^*, r_i^*) \in \mathcal{R}(t_i)} \exp \left\{ \sum_{k=1}^q f_k(x_{s_i^* r_i^* k}) \right\}} \right), \quad (6.3.2)$$

where  $\mathcal{R}(t)$  the risk-set, and  $f_k$  is a function that maps parameters  $\beta$  with network data.

### 6.3.2 Nested case control sampling

Applying partial likelihood to estimate Eq. (6.3.2) simplifies it. However, its practical implementation is often limited by the size of its denominator, the risk set  $\mathcal{R}(t)$ . As already observed by Butts [2008], the risk set typically grows quadratically with the number of nodes in standard longitudinal networks, although this growth rate can vary in different contexts. For example, in citation networks, the risk set usually expands linearly as new nodes refer to existing documents [Vu et al., 2011; Filippi-Mazzola and Wit, 2024b]. Despite the context, the scalability of these models is a common challenge, particularly in large networks with millions of nodes, where computational difficulties and limitations in model effectiveness may arise.

Vu et al. [2015] proposed a solution for this issue by suggesting the use of a nested case-control sub-sampling strategy to reduce the size of the risk set. This method involves keeping all observed events, or “cases”, and sampling only a select subset of non-events, or “controls”. According to Borgan et al. [1995], using a nested case-control sampled risk set can lead to consistent estimators. Lerner and Lomi [2020b] built on this idea and pointed out that in large networks that contain millions of nodes and events, selecting one control per case is often sufficient for obtaining reliable parameter estimates.

Similarly, Boschi et al. [2023] and Filippi-Mazzola and Wit [2024b] have shown that when the model is adjusted to include just one control per case, the partial likelihood as in (6.3.2) can be reformulated as the likelihood of a logistic regression model with only successful outcomes as observed responses. With this transformation in place, the sub-sampled case-control version of the partial

likelihood is given as,

$$\tilde{L}_p(\beta) = \prod_{i=1}^n \left[ \frac{\exp \{f(x_{s_i r_i}) - f(x_{s_i^* r_i^*})\}}{1 + \exp \{f(x_{s_i r_i}) - f(x_{s_i^* r_i^*})\}} \right], \quad (6.3.3)$$

where  $f(x_{s_i r_i}) = \sum_{k=1}^q f_k(x_{s_i r_i k})$ , and  $x_{s_i^* r_i^*}$  is a non-event for  $i$ th event with randomly sampled sender  $s_i^*$  and sampled receiver  $r_i^*$  from  $\mathcal{R}(t_i)$ .

### 6.3.3 Modeling non-linear effects with neural networks

Conventional REM formulations assume a linear relationship between the logarithmic  $(s, r)$  event rate and the covariates. However, the actual relationship may be more intricate and non-linear. Incorporating non-linear effects into REMs was initially addressed by [Bauer et al. \[2021\]](#) and [Filippi-Mazzola and Wit \[2024b\]](#). These studies extensively utilized B-splines, a method that interprets data through a series of interlinked piecewise polynomial functions ([De Boor \[1972\]](#)). Although the use of splines is beneficial for precisely identifying non-linear trends, they present certain implementation challenges. One significant challenge is the increased memory requirements during the model fitting process, as each new effect introduced necessitates the generation of an additional multi-dimensional matrix.

Building upon the recent advancements in Neural Additive Models [[Agarwal et al., 2021](#)], [Filippi-Mazzola and Wit \[2024a\]](#) solved the computational memory issues caused by the introduction of splines by proposing the Deep Relational Event Additive Model (DREAM). DREAM strategically trades off memory usage with computational complexity by leveraging multi-layered neural networks to model non-linear effects in REMs. Each covariate in (6.3.3) is then modeled by an independent neural network. Let then  $f_k$  (for  $k = 1, \dots, q$ ) be a feed-forward Artificial Neural Network (ANN) [Ripley \[1996\]](#) with a single input and a single output. The aggregation of the function  $f(x_{s,r})$  represents then the collective sum of the outputs from the ANNs, which is used to maximize (6.3.3). [Figure 6.1](#) from [Filippi-Mazzola and Wit \[2024a\]](#) describes the structure of how the information passes and is aggregated through  $f(x_{s,r})$ , offering a clear understanding of the DREAM framework.

DREAM has a higher computational complexity compared to classic spline approaches. However, it scales efficiently thanks to the higher computational power of Graphic Processor Units (GPUs) on which the fitting procedure is based, as demonstrated in [Filippi-Mazzola and Wit \[2024a\]](#). Additionally, each of the  $k$  in-

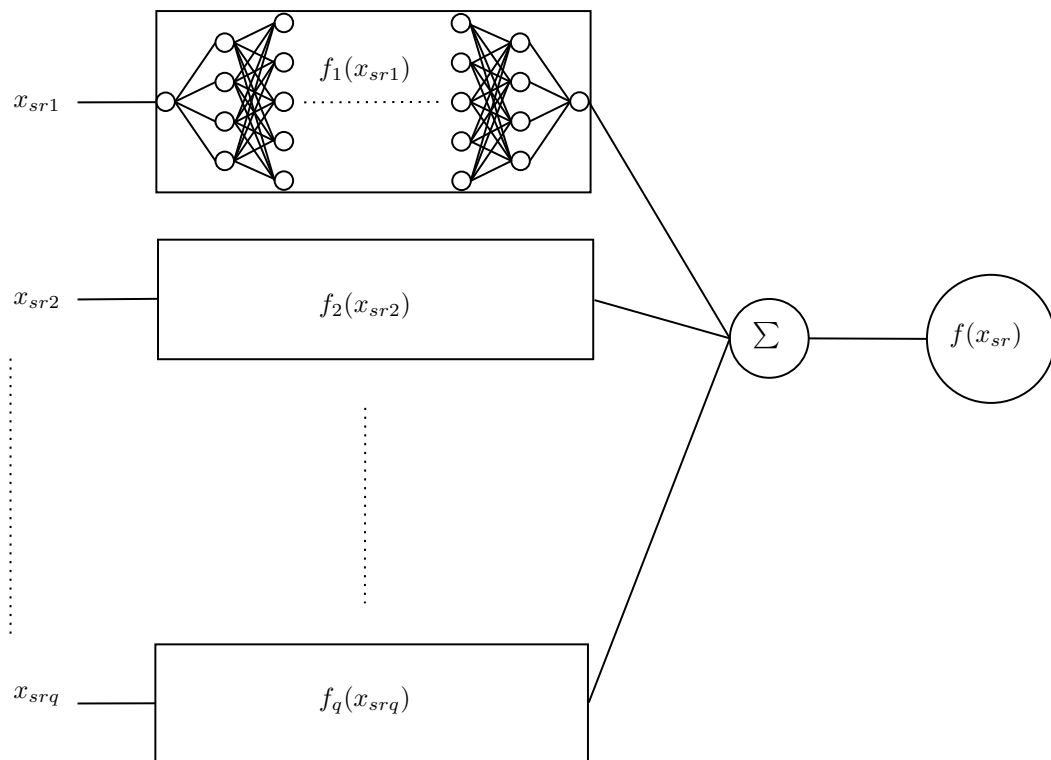


Figure 6.1: DREAM framework from [Filippi-Mazzola and Wit \[2024a\]](#). Each effect is modeled via an independent Neural Network structure that captures its non-linearity.

dependent ANNs is trained simultaneously using ADAM Kingma and Ba [2017], a Stochastic Gradient Descent (SGD) approach. The simultaneous estimation of these neural networks not only increases computational efficiency but also effectively captures interdependencies and mutual information among different effects. Estimates of uncertainty for the curves were obtained using the methodology described by Filippi-Mazzola and Wit [2024a], which employs a combination of bootstrap resamples and Gaussian process Rasmussen and Williams [2006].

#### 6.3.4 Network statistics

In REMs, network effects represent temporal patterns associated with structures of network dependence. Drawing upon Vu et al. [2015] and based on Lomi and Bianchi [2021], the current study posits that the significance of a relational event diminishes over time in a manner described by a power law distribution Brandes et al. [2009], such as

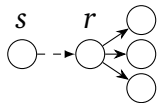
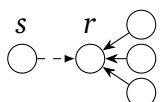
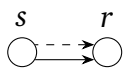
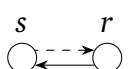
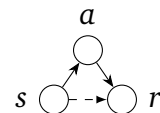
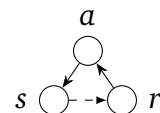
$$s(t, t_i, \alpha) = (t - t_i)^{-\alpha}, \quad (6.3.4)$$

where  $t_i$  is the exact time of the relational event on the edge  $(s_i, r_i)$ , and  $\alpha$  is the time-decay parameter. When  $\alpha = 0$ , all the past events contribute equally to the computation of network statistics. When  $\alpha > 0$ , recent events have a greater impact. Therefore, the larger  $\alpha$ , the lower the impact of past events on the statistic. This assumption allows for a nuanced understanding of how past events exert influence over a period, gradually waning in their relevance. This approach not only adheres to established practices in the field but also provides a more comprehensive framework for analyzing the dynamics of relational events within networks. Such a framework is essential for understanding the underlying patterns and trends that drive interactions in complex event networks. The decay parameter in our analysis was set to  $\alpha = 0.5$ , as explained in Lomi and Bianchi [2021], to reflect the weights of past events following the European Central Bank calendar for three-month longer-term refinancing operations.

Table 6.1 contains a list of the network statistics computed for this analysis, where node  $a$  is considered as a third (alter) trading counterpart of the sender-receiver pair  $(s, r)$ . The term  $N_{sr}(t^-)$  is the number of relational events flowing from sender  $s$  to receiver  $r$  right before time  $t$ , while  $s(t, t_i, \alpha)$  is the decay function in (6.3.4) accounting for the temporal relevance of previous events. Solid line arrows refer to past relational events, while dashed arrows indicate current events. Being associated with preferential attachment, *out-degree* and *in-degree* statistics refer to the tendency of nodes within a network to form new connections based



Table 6.1: List of network covariates and their relative mechanism.

Network Covariate	Mechanism	Formula
Out-degree		$\sum_{s \neq r} N_{rs}(t^-)$
In-degree		$\sum_{s \neq r} N_{sr}(t^-)$
Repetition		$\sum_{i=1}^{N_{sr}(t^-)} s(t, t_i, \alpha)$
Reciprocity		$\sum_{i=1}^{N_{rs}(t^-)} s(t, t_i, \alpha)$
Transitive closure		$\sum_{i=1}^{N_{sa}(t^-)} \sum_{i=1}^{N_{ra}(t^-)} s(t, t_i, \alpha)$
Cyclic closure		$\sum_{i=1}^{N_{as}(t^-)} \sum_{i=1}^{N_{ra}(t^-)} s(t, t_i, \alpha)$

on their level of interaction with other actors. Specifically, out-degree measures the number of outgoing transactions a financial institution initiates. A higher out-degree suggests that a bank is more active in outsourcing to others, potentially indicating an influence or prominence within the network. As banks engage more frequently with others, their out-degree increases, which can be interpreted as a growing likelihood of initiating additional connections. In-degree effects count the number of incoming connections a node receives. A higher in-degree is indicative of an institution's attractiveness or popularity within the network. *Repetition* in the context of liquidity trading among banks refers to their propensity to engage in similar transactions repeatedly over time. This pattern indicates that when banks have previously traded liquidity, they are likely to do so again in the future. This inertia plays a crucial role in stabilizing the flow of transactions in the interbank market. It implies a level of predictability and reliability in the interactions between banks. *Reciprocity* pertains to a behavioral pattern observed among banks in the realm of liquidity management. It denotes the inclination of banks that have historically acted as providers of liquidity to subsequently assume the role of receivers of liquidity at a later time. In our study, reciprocity is a potentially relevant indicator for the underlying interbank market functionality, as banks that supply liquidity to others may eventually find themselves in need of liquidity.

*Transitive closure* in the context of REMs assesses the degree to which the likelihood of future interactions between two entities (a sender and a receiver) is influenced by their past interactions through a mutual third party. This concept is rooted in the idea that relationships in a network are not isolated but are instead interconnected through various nodes, forming a web of interactions. This statistic explores how the previous interactions of a sender with a third party, along with that third party's history with a recipient, can establish a way that enhances the likelihood of a direct interaction between the original sender and the recipient in the future. It implies that if Bank A frequently interacts with Bank B, and Bank B regularly interacts with Bank C, then the likelihood of Bank A starting a direct interaction with Bank C is higher.

*Cyclic closure* refers to the tendency for sequences of relationships to form closed loops or cycles. In simpler terms, cyclic closure occurs when a series of relational transactions eventually leads back to the original node, thus creating a circular pattern of interactions. The significance of cyclic closure lies in its ability to reveal the complex interdependencies and reciprocal nature of relationships within a network. It provides insights into the resilience and stability of the network, as cyclic patterns often indicate robustness in the network structure. In financial networks, the presence of cyclic closures can imply a balanced system of

credit and liquidity exchange, suggesting a healthy level of interconnectedness and mutual support among different entities.

## 6.4 Results

DREAM is currently developed in Python3, utilizing the PyTorch suite [Paszke et al., 2019], and is openly accessible on GitHub at <https://github.com/efm95/DREAM>. In this analysis, each ANN consists of three layers of neurons featuring layer sizes of [32, 64, 16]. For the training process, we used a freely available GPU on Google Colab (Nvidia Tesla T4 with 15 GB of memory). The results from the fitted model are in Figure 6.2, where we highlighted the different effects between the two-time frames we considered, where the second corresponds to the crisis period. *Phase 1*: from January 4, 1999 to September 15, 2008. *Phase 2*: from September 16, 2008 to December 31, 2015. Figure 6.2 shows that most

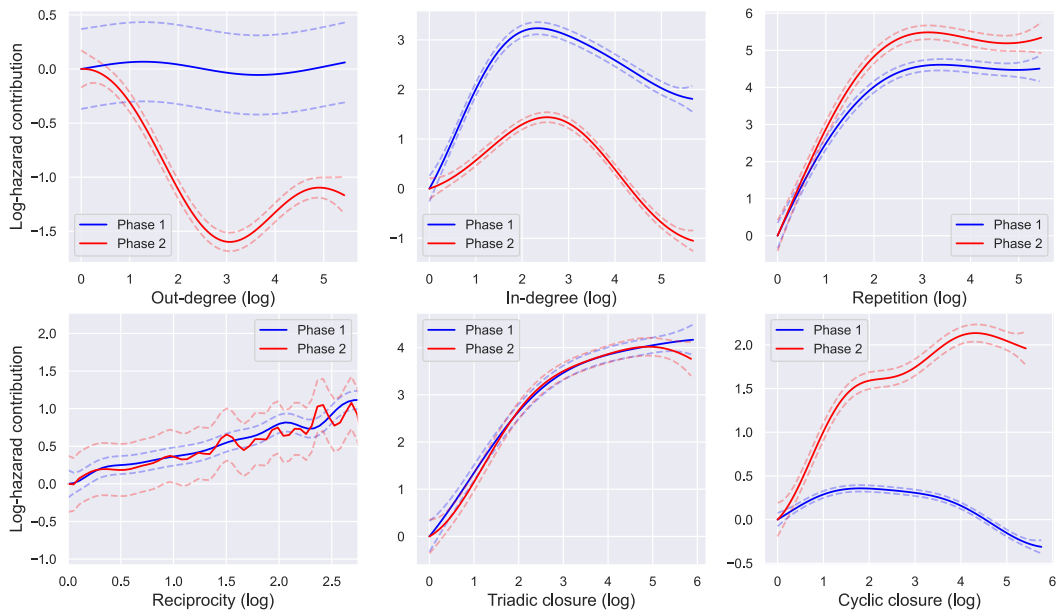


Figure 6.2: DREAM results. Among all the estimated effects, we highlighted the two different phases that have been considered for this analysis. Phase 1: from January 4, 1999 to September 15, 2008. Phase 2: from September 16, 2008 to December 31, 2015. Dashed lines represent 95% confidence intervals. Up-left: out-degree. Up-middle: in-degree. Up-right: repetition. Down-left: reciprocity. Down-middle: triadic closure. Down-right: cyclic closure.

effects differ depending on the two phases that have been taken into considera-

tion. The following paragraph will explore the outcome of the estimated effects in figure 6.2.

*Out-degree.* While there is almost no effect in the first phase, as the curve is almost flat, a more pronounced effect is more visible in the second phase. The curve exhibits a decrease, which levels off at the point corresponding to  $\exp(3) \approx 20$ , suggesting that, during the second phase, the more a bank borrows, the less likely it is to borrow again, a pattern not observed before the crisis. From a pointwise interpretation, it is possible to notice from the curve that the likelihood of a bank engaging in another lending transaction after its first is almost three times higher than a bank that has already carried out seven transactions.

*In-degree.* The pattern of the two curves is very similar, although they differ in magnitude. For both, there is a consistent trend where the probability for a bank to request loans increases with the number it has already received up to a certain threshold. This plateau occurs around the 12th transaction ( $\exp(2.5)$ ) that occurred within the observed time window, for both phases. Nevertheless, there is a noticeable difference in the tendency to seek additional loans. The probability of a bank receiving a second loan is nearly three times higher before the crisis than after it. From the curve, it can be observed that in phase 1, the likelihood for a bank to receive a second loan after its initial transaction is nearly three times greater than in phase 2.

*Repetition.* The estimated statistic increases sharply with the repetition for both phases, which indicates a strong tendency for banks to re-engage in transactions with previous partners. While the two trajectories share considerable similarities, it is noteworthy that in Phase 2, there appears to be an even stronger propensity to repeat transactions with the same institution after the second transaction has occurred. This tendency could be attributed to increased trust and reliability factors, which may have become more critical in the decision-making process of banks following the initial transaction.

*Reciprocity.* The plot shows comparable patterns across both phases, indicating that the economic crisis did not significantly alter the reciprocal nature of transactions. This similarity suggests that the propensity of banks to alternate roles between lenders and borrowers remained stable despite the financial upheaval. In the latter half of the plot, phase 2 displays more erratic behavior.

*Transitive closure.* The estimated effect shows an almost identical increasing trend for both phases. This might imply the crisis did not affect transitive interactions.

*Cyclic closure.* The effect of cyclic closure displays a marked disparity between the two phases. In Phase 2, there is a notable and consistent rise, indicating a period of steady contribution once a certain level is reached. Conversely, in Phase 1, we

observe a small peak, giving way to a slight but consistent decline. While before the crisis, the network might have lessened the emphasis on, or the frequency of, such cyclic patterns. After the crisis, the highlighted pattern suggests that cyclic transactions played a more prevalent role in the network after the crisis, where trust and relationships among institutions were more crucial.

## 6.5 Conclusions

In this work, we analyzed the transactions that occurred on the European interbank network through the lenses of Relational Event Models. Thanks to the use of the Deep Relational Event Additive Model, we were able to model a series of network effects non-linearly with the use of neural networks.

In this analysis, we highlighted the remarkable differences in the liquidity trading network dynamics before and after the 2008 economic crisis. In the post-crisis phase, banks demonstrated a decreased propensity to borrow as their previous borrowing activity increased, a trend that was not evident before the crisis. This suggests a shift in the borrowing strategies and possibly tighter liquidity management or risk aversion behaviors following the crisis.

Furthermore, this study emphasizes the importance of modeling such exchange mechanisms nonlinearly to extract more information. In this regard, the application of DREAM provided a deeper understanding of the underlying dynamics. Therefore, this study expands the current work of [Lomi and Bianchi \[2021\]](#), by offering a more comprehensive and detailed picture of global trading dynamics and the intricate interactions within financial networks.

In segmenting our data into periods before and after the 2008 financial crisis, we aimed to isolate the effects of this pivotal event on the European interbank market. However, we recognize that other significant events within this two-time segmentation also shape market dynamics. Future analyses could benefit from considering multiple segmentation depending on different financial events to refine our understanding of the market's behavior over time.

Future research could also focus on dissecting the causes behind these observed shifts. Additionally, extending the analysis to include even exogenous effects concerning characteristics from the institutions themselves could provide further insights into the effects of the crisis and their subsequent recovery.

# Chapter 7

## Conclusions

This chapter briefly summarizes the key contributions of this thesis, and the insights that have been learned from it. Furthermore, it describes the open challenges and the limitations of the work.

### 7.1 Summary of Key Findings

This dissertation has explored various aspects of dynamic network modeling, with a particular focus on patent citation networks. The research projects presented in each chapter contribute to a deeper understanding of the complexities inherent in these networks, introduce novel methodologies to offer new analytical perspectives, and propose solutions to overcome the limitations of existing models.

**Text-based Similarity and Patent Relatedness.** In the first study, we proposed a novel approach to compute textual similarity scores by generating embeddings from patent abstracts using a pre-trained Natural Language Processing (NLP) model. This method proved to be not only efficient but also effective in bypassing the computational bottlenecks associated with traditional techniques by utilizing the concept of transfer learning. Our analysis further explored the application of Generalized Additive Models (GAMs) to uncover non-linear relationships between patent similarities and various influencing factors. By incorporating both fixed and non-linear effects, we demonstrated that the observed downward trend in patent similarity scores is not constant but instead characterized by oscillating behavior over time. Key findings include the significant impact of the time lag between citing and cited patents and the critical role of citation behavior and class effects in shaping patent similarities.

Relational Event Models and Their Evolution. The second project involved a comprehensive review of the current state of the art in the field of Relational Event Models (REMs), which have become a prominent framework for analyzing dynamic network data over the last decade. We reviewed the core properties and mathematical underpinnings of REMs, highlighting their flexibility across multiple scientific fields, ranging from ecology and healthcare to political science. In this review, we also addressed a series of challenges and open issues, particularly concerning procedures for assessing goodness of fit, as well as ongoing development projects within the REM community.

Stochastic Gradient Relational Event Additive Model. The subsequent chapter of this dissertation addresses the limitations of traditional REMs by introducing the Stochastic Gradient Relational Event Additive Model (STREAM). This model integrates non-linear modeling of covariates through B-splines with nested case-control sampling, enabling the efficient approximation of REM likelihoods through logistic regression. STREAM also incorporates techniques borrowed from deep learning to overcome limitations in estimating such large models. The model was applied to the extensive network of U.S. patent citations from 1976 to 2022, encompassing over 8 million patents and 100 million citations, revealing significant patterns in patent citation rates. Notably, patents issued around the year 2000 were found to be particularly influential, suggesting a period of heightened technological innovation. Although STREAM was specifically designed for modeling this patent citation network, its mathematical framework is flexible enough to be employed in analogous situations.

Deep Relational Event Additive Model. In response to the computational challenges posed by large dynamic networks, the next chapter introduced the Deep Relational Event Additive Model (DREAM). DREAM leverages neural networks to model non-linear effects in covariates, offering a scalable and efficient solution for analyzing extensive datasets. Through a series of simulation studies, we demonstrated DREAM's robustness and efficiency compared to traditional methods. Additionally, we successfully replicated the results obtained from the STREAM model, validating DREAM's effectiveness.

Applying DREAM to the European Interbank Market. In the final chapter, we demonstrated the flexibility of DREAM by applying it to the European interbank network. This analysis uncovered significant shifts in liquidity trading dynamics following the 2008 financial crisis. The study not only revealed the emergence

of clustering patterns, likely driven by a trust system that developed within the market post-crisis, but also highlighted the importance of non-linear modeling in large datasets for capturing complex network behaviors.

## 7.2 Open Challenges and Limitations

The research presented in this dissertation contributes to both the field of patent analysis and network science by extending the theoretical and practical application of REMs. By incorporating non-linear modeling techniques and deep learning approaches, we have contributed to expanding the methodological capabilities of REMs, enabling them to more accurately capture the complexities of dynamic networks. From a theoretical perspective, our work underscores the importance of considering non-linear relationships in the analysis of network interactions. The integration of splines and neural networks into REMs provides deeper insights into how event dynamics unfold over time and how various factors influence such dynamics. Despite the potential advancements this research offers, there remain several challenges and limitations that need to be addressed. One of the primary challenges encountered relates to the computation of complex network statistics. While models like STREAM and DREAM address some of the issues associated with estimating REMs in large datasets, the calculation of intricate network effects, such as triadic or cyclic closure, remains problematic. Current approaches to computing these statistics inherit a computational complexity that is prohibitive for large-scale scenarios. Although [Lerner and Lomi \[2020b\]](#) proposed a series of techniques to overcome such limitations, these methods are currently confined to the Java package *Eventnet*, making them difficult to implement outside the scope of this software.

Another significant limitation involves to the assessment of model fit in REMs, particularly when applied to large and complex datasets. Traditional metrics proved effective in controlled simulation settings, but less reliable in empirical scenarios where the true underlying functions were unknown. As a result, our reliance on cross-validation techniques, while practical and limited to the model selection approach, does not fully address the need for more sophisticated methods to properly evaluate the model's goodness of fit. This highlights a critical area for future research: the development of robust, scalable approaches to model evaluation that can accommodate the intricacies of large, dynamic networks. Novel approaches such as those proposed by [Amati et al. \[2024\]](#) and [Boschi and Wit \[2024\]](#), reflect the growing focus within the community on addressing this issue. Additionally, the models we developed consistently assumed dyadic interactions



between entities. While this approach is at the core of Relational Event Modeling, it may oversimplify real-world dynamics. As demonstrated by [Lerner et al. \[2023\]](#), citation networks and other dynamic systems often involve more complex, multi-party interactions. Expanding the scope of our analyses to incorporate polyadic or hyperedge interactions remains an important direction for future research. Such extensions could provide a more accurate representation of the processes within these networks, where interactions frequently involve multiple actors.

# Appendix A

## Supplementary materials for Chapter 4: “A Stochastic Gradient Relational Event Additive Mode for modelling US patent citations from 1976 until 2022”

### A.1 B-spline recursive formulation

Let  $B_{j,p}^k(x_{srk})$  be the  $j$ -th basis transformation ( $j = 1, \dots, d$ ), i.e. a series of connected piece-wise polynomial functions of order  $p$  defined over a grid of knots  $u_0, u_1, \dots, u_m$ , such that  $u_{l-1} < u_l$ , for  $l = 1, \dots, m$ , on the parameter space that characterize the covariate  $x_{srk}(t)$ , for  $k = 1, \dots, q$ .  $B_{j,p}^k(x_{srk})$  can be defined recursively [De Boor, 1972], as

$$B_{j,p}^k(x_{srk}) = \frac{x_{srk}(t) - u_j}{u_{j+p} - u_j} B_{j,p-1}^k(x_{srk}) + \frac{u_{j+p+1} - x_{srk}(t)}{u_{j+p+1} - u_{j+1}} B_{j+1,p-1}^k(x_{srk}), \quad (\text{A.1.1})$$

where  $d$  are the number of splines that represent the degrees of freedom of the B-spline transformation for covariate  $x_{srk}(t)$  and where

$$B_{j,0}^k(x_{srk}) = \begin{cases} 1 & \text{if } u_j \leq x_{srk} < u_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

Table A.1: Model selection considerations using AIC and BIC. Effects have been divided into three groups: nodal (*No*), similarity (*Si*), and time-varying (*Tv*).

Effect group	AIC	BIC
No	100'547'264	100'547'552
Si	21'212'756	21'212'952
Tv	58'466'908	58'467'104
No + Si	17'783'270	17'783'762
No + Tv	46'950'016	46'950'508
Si + Tv	9'530'301	9'530'694
No + Si + Tv	8'303'071	8'303'759

## A.2 Model selection

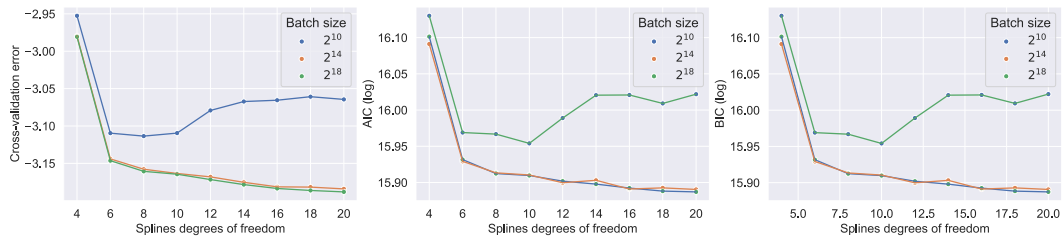
Various model formulations have been compared with each other on the available data. For simplicity, we grouped the statistics into *nodal* (*No*), *similarity* (*Si*), and *time-varying* (*Tv*) effects, we sequentially add those groups of statistics to our model. On Table A.1, the estimated AIC and BIC values for each fitted model are reported. The results highlight the significant contribution that similarity statistics make to the model when they are included. It also suggests that the complete model is the best in describing the underlying data. Clearly, with so much data available, these relatively sensible statistics all significantly contribute to improving the fit of the relational event model.

Furthermore, two hyperparameters need to be specified: batch sizes and the spline degrees of freedom. Batch sizes refer to the number of events used in each batch during the model fitting process. We tested three different batch sizes:  $2^{10}$ ,  $2^{14}$  and  $2^{18}$ . The choice of batch size affects the trade-off between computational efficiency and accuracy of model fit. Smaller batch size induce noisier gradients in the parameter updates, but are computationally more efficient.

The degrees of freedom refers to the flexibility of the nonparametric model used to estimate the non-linear effects. We tested degrees of freedom ranging from 4 to 20 to find the optimal level of flexibility. A lower degree of freedom may result in a less flexible model that underfits the data, while a higher degree of freedom may result in a model that is too flexible and, potentially, overfits the data.

To determine the best values for these hyperparameters, we compared the results from a 6-fold cross-validation, the AIC and the BIC where we test these three different batch sizes while changing the degrees of freedom from 4 to 20. As a test-error metric for the cross-validation, we evaluate the negative log-likelihood for each held-out set in the cross-validation. These values were then re-scaled by

Figure A.1: Model selection performed to compare three different batch sizes with varying degrees of freedom for the B-splines ranging from 4 to 20. Left: 6-fold cross-validation. Center: AIC scores (in log terms). Right: BIC scores (in log terms).



the size of their validation set for comparison purposes.

The results are shown on the right of Fig A.1. The plots indicate that there is no substantial difference between the models fitted with batch sizes of 2<sup>14</sup> and 2<sup>18</sup>; both achieve similarly lower scores across the three evaluation metrics. Conversely, the model with a batch size of 2<sup>10</sup> exhibits higher test scores on average. However, this model also demonstrates that metrics increase beyond the 10th spline degree of freedom, suggesting a potential overfitting issue. This overfitting could be obscured by the larger batch sizes used in the latter two models. Consequently, we determine that the optimal configuration is a batch size of 2<sup>14</sup> with 10 degrees of freedom. This configuration strikes the most effective balance between stability, accuracy, and convergence speed.

## Appendix B

# Supplementary materials for Chapter 5: “Modelling Non-linear Effects with Neural Networks in Relational Event Models”

### B.1 Oscillatory activation functions

Among the prevalent choices for activation functions lies the family of Rectified Linear Units (ReLU) [Agarap, 2019], known for their simplicity and computational efficiency. Despite the attractive features, ReLU functions are often affected by the issue known as the “dying ReLU” [Lu, 2020]. This emerges in many empirical applications when certain neurons within the network become perpetually inactive, i.e., they continuously output zeros for specific regions of the input support space. This behavior makes the affected neurons essentially irrelevant during the training phase as once a neuron enters in this state, the gradient at that point becomes zero. Consequently, during the backpropagation phase, no updates are made to the weights connected to that neuron. The absence of any weight adjustment leads to a state of inertia where the neuron remains inactive, never contributing to the model again.

Within the family of non-linear activation functions, two noteworthy alternatives to the ReLU are the Sigmoid [Narayan, 1997] and Hyperbolic Tangent (tanh) [Namin et al., 2009]. While both functions share similar sigmoidal curves, they exhibit distinct behaviors concerning their derivatives. Specifically, the sigmoid function derivative quickly approaches zero on both right and left sides. This behavior translates to smaller gradients, which in turn leads to protracted training

periods. Furthermore, this rapid decline in the derivative magnitude opens to the vanishing gradient problem during backpropagation, which poses a significant challenge in achieving swift and stable convergence in the neural network. In contrast, the tanh function mitigates some of these difficulties. Its derivative is characteristically sharper and maintains non-zero values over a more extended range on both ends. This design helps in alleviating the vanishing gradient problem to some extent. However, tanh is not without its limitations. Its adaptability is limited by its rigidity in defining the non-linear transformation shape. As a result, the tanh function sometimes struggles to model more intricate and nuanced patterns present in complex datasets.

The challenges associated with previously discussed activation functions prompted a rigorous exploration of alternative units. This led to the adoption of the Growing Cosine Unit (GCU) [Noel et al., 2023]. Initially conceptualized to mitigate the “dying ReLU” problem in convolutional neural networks, GCUs have emerged as a promising contender among oscillatory activation functions. Defined as

$$\phi(a_k^{(l-1)}) = (\beta_k^{(l)} a_k^{(l-1)} + \beta_{0k}^{(l)}) \cos(\beta_k^{(l)} a_k^{(l-1)} + \beta_{0k}^{(l)}), \quad (\text{B.1.1})$$

where  $a_k^{(l-1)}$  represents the output from the previous layer, GCUs exhibit a unique property. Unlike ReLU units, which typically yield a singular decision boundary, GCU neurons decision boundary comprises infinitely many parallel hyperplanes. This is attributable to the GCU activation function infinite zeros. Additionally, GCUs offer consistent and favorable derivatives, acting as a countermeasure against the vanishing gradient issue. Furthermore, this trait produces a more efficient training process, marked by reduced duration and improved convergence rates.

## B.2 ADAM

Consider  $\nabla \tilde{L}_p(\beta)_b$  as the gradient of the partial likelihood for batch  $b$ . In the ADAM optimization process, the first and second moment estimations are updated as follows:

$$\begin{aligned} m_b &\leftarrow \xi_1 m_{b-1} + (1 - \xi_1) \nabla \tilde{L}_p(\theta)_b, \\ v_b &\leftarrow \xi_2 v_{b-1} + (1 - \xi_2) \nabla \tilde{L}_p(\theta)_b^2, \end{aligned}$$

where  $m$  and  $v$  represent the first and second moment gradients, respectively. The hyperparameters  $\xi_1$  and  $\xi_2$  are instrumental in determining the extent to

which past gradients influence the current moment updates.

ADAM incorporates bias correction to adjust for the initial bias in the first and second moments of the gradients. This correction is crucial because the moving averages of these gradients start from zero, leading to an initial bias toward zero, particularly noticeable at the early stages of training. To counteract this, ADAM modifies the moving averages with a correction factor that is directly related to the learning rate and inversely related to the iteration count. Denoting the current training step as  $s$ , the first and second moments undergo bias correction as follows:

$$\hat{m}_{b,s} = \frac{m_b}{1 - \xi_1^s},$$

$$\hat{v}_{b,s} = \frac{v_b}{1 - \xi_2^s},$$

with  $\xi_1^s$  and  $\xi_2^s$  approaching zero as  $s$  increases. Consequently, the model parameters are updated by:

$$\beta_b \leftarrow \beta_{b-1} - \psi \frac{\hat{m}_{b,s}}{\sqrt{\hat{v}_{b,s} + \epsilon}},$$

where  $\psi$  denotes the learning rate, determining the step size of each parameter update, and  $\epsilon$  is a small constant (typically  $1e-8$ ) to avoid any division by zero. It is important to note that adjusting the learning rate during training can further refine the estimation of the weights. However, due to the complexity involved in determining an optimal learning rate decay, we chose to maintain a constant learning rate, denoted as  $\psi$ , throughout the training process in our simulation studies and application.

### B.3 Simulation study model selection

To determine the optimal configuration for DREAM, we conducted a simulation study that examines four neural network architectures, each incorporating different degrees of dropout to apply varying levels of regularization. In this setting, the choice of dropout level is a classic trade-off between bias and variance. Too little dropout may not provide sufficient regularization, leading to overfitting. Conversely, too much dropout may lead to underfitting. Analogous to regularizing splines in GAMs, increasing penalties yield to less flexible curves, resulting in more linear representations.

Table B.1 we report the four neural network architectures that have been tested.

Each of these is designed with an incrementally increasing number of neurons and layers, thereby escalating the model complexity. This progressive complexity, much like the role of splines in GAMs, allows for greater flexibility in the resulting curves. The architectures range from the simplest, Model 1, with a configuration of (64, 128, 64), to the most complex, Model 4, with an expansive (512, 1024, 512, 256, 128) layout, enabling us to scrutinize the trade-off between model complexity and curve adaptability.

Model	Framework
Model 1	(64, 128, 64)
Model 2	(128, 256, 64)
Model 3	(256, 512, 256, 128)
Model 4	(512, 1024, 512, 256, 128)

*Table B.1: Summary of neural network frameworks with varying complexities.*

Examining the insights provided by Figure B.1, a clear pattern emerges indicating that the model with the highest complexity is most adequate in capturing the non-linearities. While a comparison within Model 4 reveals minimal variance between dropout levels of 0 and 0.05, our preference inclines towards the model iteration with no penalization. By avoiding the regularization imposed by dropout, we aim to preserve the model sensitivity in capturing more data, ensuring a more accurate interpretation of the underlying dynamics.

## B.4 Complex polynomials for simulation study

The true sender function effect is defined as

$$f(x) = \sin(2\pi x) + \frac{1}{2} \sin(4\pi x) + \frac{1}{4} \sin(8\pi x).$$

The true receiver function effect is defined as

$$f(x) = -\sin(2(4x - 2)) - 2 \exp(-16^2(x - 0.5)^2).$$

## B.5 US patent citation network model selection

Building upon our prior simulation study, detailed in Section B.3, we extend our analysis to the U.S. Patent citation network. Utilizing the same architectural



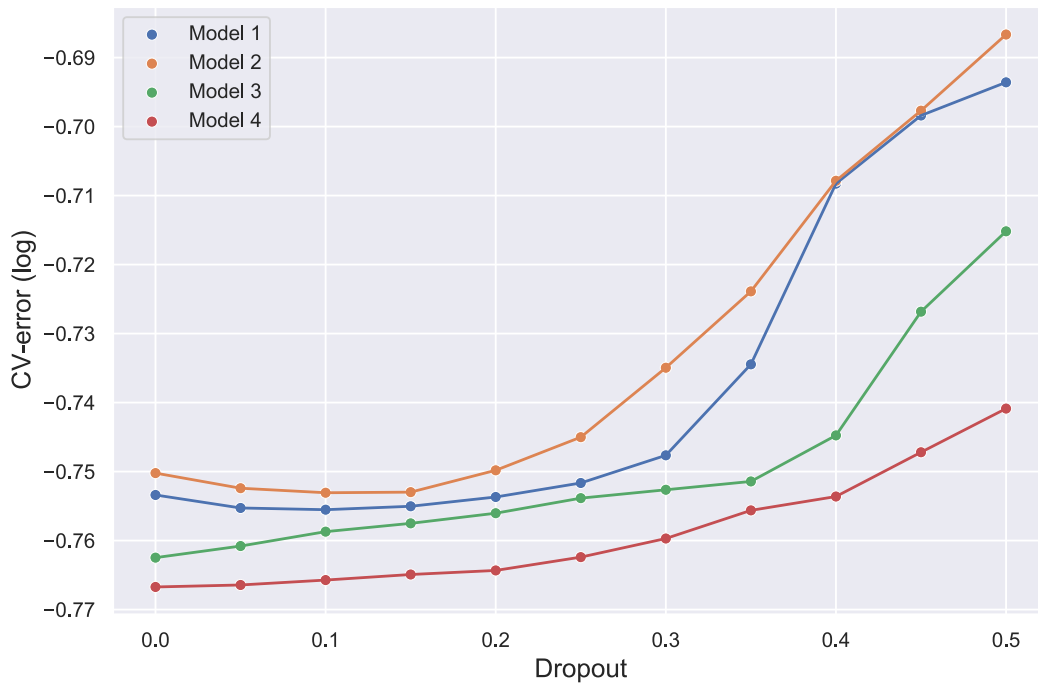


Figure B.1: Results from the simulation study illustrating the performance of various model configurations.

frameworks, we conducted a simulation study to assess performance across the network. The comparative results are visually represented in Figure B.2, where the spread of performance metrics is evaluated using a 10-fold cross-validation method. From our analysis, it becomes evident that Model 2 outperforms the others. Notably, this model demonstrates how increased computational complexity correlates with elevated cross-validation errors. Given this insight, we decided not to proceed with Model 4. The results indicated diminishing returns with higher complexity, thus reinforcing our decision to halt further simulations at Model 3. Consequently, Model 2 with a dropout of 0.05 was selected for our application due to its optimal balance of complexity and error minimization.

## B.6 US patent citation network fitted effects

For an alternative interpretability of the model's effects, Figure B.3 includes the effects of our model expressed in terms of Hazard contributions. This alternative representation translates the log hazard contributions from the primary analysis into their exponential form.

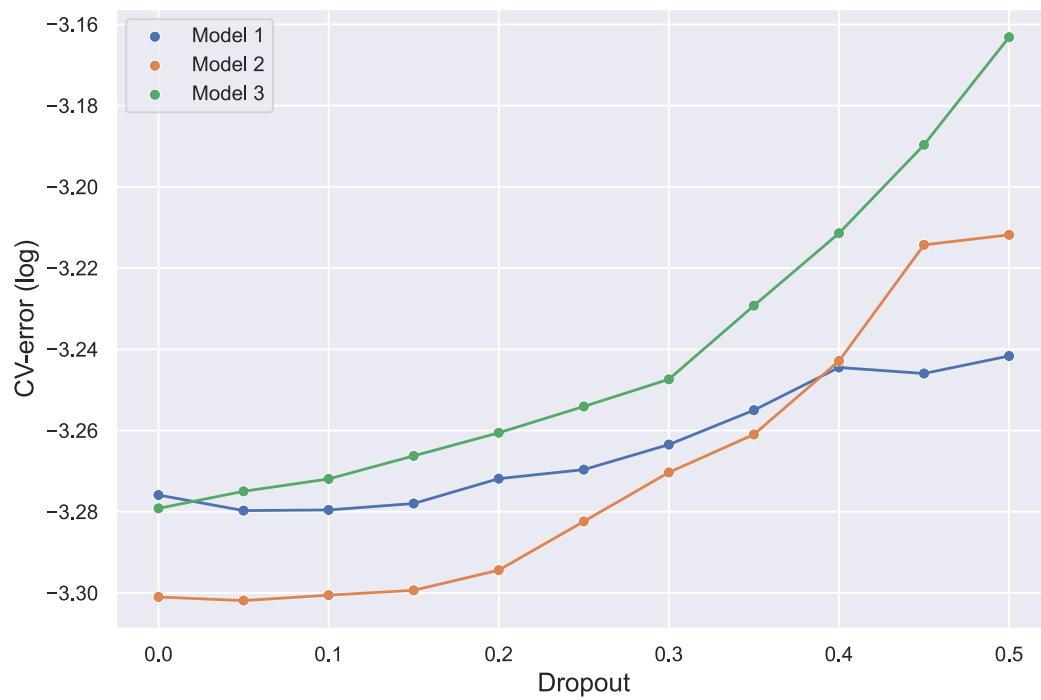


Figure B.2: Cross-validation Performance Spread for Model Selection in the U.S. Patent Citation Network.

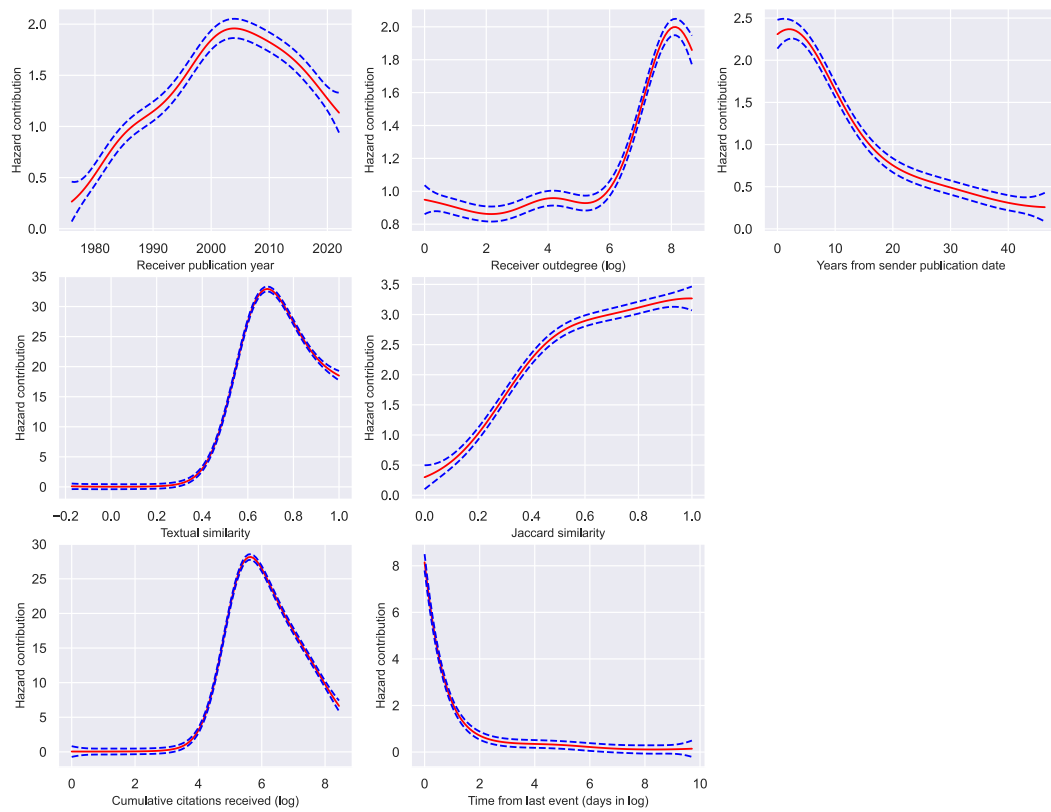


Figure B.3: Fitted effects on US patent citation network.

# Bibliography

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.
- Aalen, O., Borgan, Ø., and Gjessing, H. (2008). *Survival and Event History Analysis: a Process Point of View*. Statistics for Biology and Health. Springer, New York.
- Abbott, A. (1992). From causes to events: Notes on narrative positivism. *Sociological Methods & Research*, 20(4):428–455.
- Acemoglu, D., Akcigit, U., and Kerr, W. R. (2016). Innovation network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488.
- Agarap, A. F. (2019). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.
- Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. (2021). Neural Additive Models: Interpretable Machine Learning with Neural Nets. *arXiv preprint arXiv:2004.13912*.
- Albert, M. B., Avery, D., Narin, F., and McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3):251–259.
- Amati, V., Lomi, A., and Mascia, D. (2019). Some days are better than others: Examining time-specific variation in the structuring of interorganizational relations. *Social Networks*, 57:18–33.
- Amati, V., Lomi, A., and Mira, A. (2018). Social network modeling. *Annual Review of Statistics and Its Application*, 5:343–369.
- Amati, V., Lomi, A., and Snijders, T. A. B. (2024). A goodness of fit framework for relational event models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, qnae016.

- An, J., Kim, K., Mortara, L., and Lee, S. (2018). Deriving technology intelligence from patents: Preposition-based semantic analysis. *Journal of Informetrics*, 12(1):217–236.
- An, X., Li, J., Xu, S., Chen, L., and Sun, W. (2021). An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics*, 15(2):101135.
- Andersen, P., Borgan, Ø., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
- Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.
- Arena, G., Mulder, J., and Leenders, R. T. A. (2023). How fast do we forget our past social interactions? understanding memory retention with parametric decays in relational event models. *Network Science*, 11(2):267–294.
- Arena, G., Mulder, J., and Leenders, R. T. A. J. (2022). A Bayesian Semi-Parametric Approach for Modeling Memory Decay in Dynamic Social Networks. *Sociological Methods & Research*.
- Artico, I. and Wit, E. C. (2023). Dynamic latent space relational event model. *Journal of the Royal Statistical Society Series A: Statistics in Society*.
- Arts, S. and Veugelers, R. (2014). Technology familiarity, recombinant novelty, and breakthrough invention. *Industrial and Corporate Change*, 24(6):1215–1246.
- Bacchiocchi, E. and Montobbio, F. (2010). International Knowledge Diffusion and Home-bias Effect: Do USPTO and EPO Patent Citations Tell the Same Story? *Scandinavian Journal of Economics*, 112(3):441–470.
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.
- Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.
- Barnes, J. A. and Harary, F. (1983). Graph theory in network analysis. *Social Networks*, 5(2):235–244.

- Bauer, V., Harhoff, D., and Kauermann, G. (2021). A smooth dynamic network model for patent collaboration data. *AStA Advances in Statistical Analysis*, 106(1):1–20.
- Beckert, J. (2010). How do fields change? the interrelations of institutions, networks, and cognition in the dynamics of markets. *Organization Studies*, 31(5):605–627.
- Bekamiri, H., Hain, D. S., and Jurowetzki, R. (2021). Hybrid model for patent classification using augmented SBERT and KNN. *arXiv preprint arXiv:2103.11933*.
- Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.
- Besag, J. E. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.
- Bianchi, F., Filippi-Mazzola, E., Lomi, A., and Wit, E. C. (2024). Relational event modeling. *Annual Review of Statistics and Its Application*, 11(1):297–319.
- Bianchi, F. and Lomi, A. (2022). From ties to events in the analysis of interorganizational exchange relations. *Organizational Research Methods*.
- Bianchi, F., Stivala, A., and Lomi, A. (2022). Multiple clocks in network evolution. *Methodological Innovations*, 15(1):29–41.
- Borgan, Ø., Goldstein, L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the cox proportional hazards model. *The Annals of Statistics*, 23(5):1749–1778.
- Borgan, Ø. and Keogh, R. (2015). Nested case-control studies: Should one break the matching? *Lifetime Data Analysis*, 21(4):517–541.
- Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916):892–895.
- Boschi, M., Juozaitienė, R., and Wit, E.-J. C. (2023). Smooth alien species invasion model with random and time-varying effects. *arXiv preprint arXiv:2304.00654*.

- Boschi, M. and Wit, E.-J. C. (2024). Goodness of fit of relational event models. *arXiv preprint arXiv:2407.08599*.
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, Germany. Physica-Verlag HD.
- Brandenberger, L. (2018a). *Rem: Relational Event Models*. R package version 1.3.1.
- Brandenberger, L. (2018b). Trading favors—examining the temporal dynamics of reciprocity in congressional collaborations using relational event models. *Social Networks*, 54:238–253.
- Brandenberger, L. (2019). Predicting network events to assess goodness of fit of relational event models. *Political Analysis*, 27(4):556–571.
- Brandes, U., Lerner, J., and Snijders, T. A. B. (2009). Networks evolving step by step: Statistical analysis of dyadic event data. In *International Conference on Advances in Social Network Analysis and Mining*, pages 200–205. IEEE.
- Breschi, S. and Lissoni, F. (2004). Knowledge networks from patent data. In *Handbook of Quantitative Science and Technology Research*, pages 613–643. Springer.
- Breslow, N. E. (1972). Discussion of professor Cox’s paper. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):216–217.
- Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1):155–200.
- Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science*, 325(5939):414–416.
- Butts, C. T., Lomi, A., Snijders, T. A. B., and Stadtfeld, C. (2023). Relational event models in network science. *Network Science*, 11(2):175–183.
- Butts, C. T. and Marcum, C. S. (2017). A relational event approach to modeling behavioral dynamics. In Pilny, A. and , Poole, M. S., editors, *Group Processes: Data-Driven Computational Approaches*, pages 51–92. Springer.
- Carpenter, M. P. and Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. *World Patent Information*, 5(3):180–185.

- Carruthers, B. G. and Kim, J.-C. (2011). The sociology of finance. *Annual Review of Sociology*, 37(1):239–259.
- Cetina, K. K. and Preda, A. (2004). *The Sociology of Financial Markets*. Oxford University Press, Oxford.
- Chakraborty, M., Byshkin, M., and Crestani, F. (2020). Patent citation network analysis: A perspective from descriptive statistics and ergms. *PLOS ONE*, 15(12):1–28.
- Choi, S., Lee, H., Park, E., and Choi, S. (2022). Deep learning for patent landscaping using transformer and graph embedding. *Technological Forecasting and Social Change*, 175(C).
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.
- Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, London.
- Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, New York.
- Cressie, N. A. C. (2015). *Statistics for Spatial Data (Revised Version)*. John Wiley & Sons, New York.
- Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods*. Springer, New York.
- De Boor, C. (1972). On calculating with b-splines. *Journal of Approximation Theory*, 6(1):50–62.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. 41(6):391–407.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.



- DuBois, C., Butts, C., and Smyth, P. (2013a). Stochastic blockmodeling of relational event dynamics. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 31 of *Proceedings of Machine Learning Research*, pages 238–246. PMLR.
- DuBois, C., Butts, C. T., McFarland, D., and Smyth, P. (2013b). Hierarchical models for relational event sequences. *Journal of Mathematical Psychology*, 57(6):297–309.
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(7):2121–2159.
- Duxbury, S. W. and Haynie, D. L. (2021). Shining a light on the shadows: Endogenous trade structure and the growth of an online illegal market. *American Journal of Sociology*, 127(3):787–827.
- Duxbury, S. W. and Haynie, D. L. (2023). Network embeddedness in illegal online markets: Endogenous sources of prices and profit in anonymous criminal drug trade. *Socio-Economic Review*, 21(1):25–50.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2).
- Elmer, T. and Stadtfeld, C. (2020). Depressive symptoms are associated with social isolation in face-to-face interaction networks. *Scientific Reports*, 10(1):1–12.
- Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, 25(3):233–242.
- Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21(4):825–839.
- Filippi-Mazzola, E., Bianchi, F., and Wit, E. C. (2023). Drivers of the decrease of patent similarities from 1976 to 2021. *PLOS ONE*, 18(3):1–13.

- Filippi-Mazzola, E. and Wit, E. C. (2024a). Modeling non-linear effects with neural networks in relational event models. *Social Networks*, 79:25–33.
- Filippi-Mazzola, E. and Wit, E. C. (2024b). A Stochastic Gradient Relational Event Additive Model for modelling US patent citations from 1976 until 2022. *Journal of the Royal Statistical Society Series C*, 73(4):1008–1024.
- Fleming, L. and Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy*, 30(7):1019–1039.
- Foucault Welles, B., Vashevko, A., Bennett, N., and Contractor, N. (2014). Dynamic Models of communication in an online friendship network. *Communication Methods and Measures*, 8(4):223–243.
- Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.
- Freeman, L. C., Romney, A. K., and Freeman, S. C. (1987). Cognitive structure and informant accuracy. *American Anthropologist*, 89(2):310–325.
- Fritz, C., Lebacher, M., and Kauermann, G. (2020). Tempus volat, hora fugit: A survey of tie-oriented dynamic network models in discrete and continuous time. *Statistica Neerlandica*, 74(3):275–299.
- Fritz, C., Mehrl, M., Thurner, P. W., and Kauermann, G. (2023). All that glitters is not gold: Relational events models with spurious events. *Network Science*, 11(2):184–204.
- Fritz, C., Thurner, P. W., and Kauermann, G. (2021). Separable and semiparametric network-based counting processes applied to the international combat aircraft trades. *Network Science*, 9(3):291–311.
- Gabbi, G., Germano, G., Hatzopoulos, V., Iori, G., and Politi, M. (2013). Market microstructure, banks’ behaviour, and interbank spreads. Working paper.
- Gabrieli, S. (2011). The functioning of the european interbank market during the 2007–08 financial crisis. Working paper, CEIS.
- Gibson, D. R. (2003). Participation shifts: Order and differentiation in group conversation. *Social Forces*, 81(4):1335–1380.
- Gibson, D. R. (2005). Taking turns and talking ties: Networks and conversational interaction. *American Journal of Sociology*, 110(6):1561–1597.

- Gile, K. J. and Handcock, M. S. (2017). Analysis of networks with missing data with application to the national longitudinal study of adolescent health. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 66(3):501.
- Golder, S. A., Wilkinson, D. M., and Huberman, B. A. (2007). Rhythms of social interaction: Messaging within a massive online network. In *Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference, Michigan State University*, pages 41–66. Springer.
- Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.
- Gravel, J., Valasik, M., Mulder, J., Leenders, R. T. A. J., Butts, C., Brantingham, P. J., and Tita, G. E. (2023). Rivalries, reputation, retaliation, and repetition: Testing plausible mechanisms for the contagion of violence between street gangs using relational event models. *Network Science*, 11(2):324–350.
- Gress, B. (2010). Properties of the USPTO patent citation network: 1963–2002. *World Patent Information*, 32(1):3–21.
- Griliches, Z., Pakes, A., and Hall, B. H. (1986). The value of patents as indicators of inventive activity. NBER Working Paper 2083, National Bureau of Economic Research.
- Hage, P (1979). Graph theory as a structural model in cultural anthropology. *Annual Review of Anthropology*, 8(1):115–136.
- Hage, P and Harary, F (1984). *Structural Models in Anthropology*. Cambridge Studies in Social and Cultural Anthropology. Cambridge University Press, Cambridge.
- Hain, D. S., Jurowetzki, R., Buchmann, T., and Wolf, P. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.
- Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The nber patent citation data file: Lessons, insights, and methodological tools. NBER Working Paper 8498, National Bureau of Economic Research, Cambridge, MA.
- Hanneke, S., Fu, W., and Xing, E. P (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 5:585–605.

- Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297–310.
- Haunss, S. and Hollway, J. (2022). Multimodal mechanisms of political discourse dynamics and the case of germany’s nuclear energy phase-out. *Network Science*, 11(2):205–223.
- Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r d relationship. *Econometrica*, 52(4):909–938.
- Henderson, R. M. and Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative science quarterly*, pages 9–30.
- Hoffman, M., Block, P., Elmer, T., and Stadtfeld, C. (2020). A model for the dynamics of face-to-face interactions in social groups. *Network Science*, 8(S1):S4–S25.
- Holland, P. W. and Leinhardt, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, 5(1):5–20.
- Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.
- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.
- Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.
- Immordino, S. C. (2019). *Comparing Similarity of Patent Textual Data Through the Application of Machine Learning*. PhD thesis, University of Illinois at Chicago.
- Iori, G., Mantegna, R. N., Marotta, L., Micciché, S., Porter, J., and Tumminello, M. (2015). Networked relationships in the e-mid interbank market: A trading model with memory. *Journal of Economic Dynamics and Control*, 50:98–116.

- Jaffe, A. (1986). Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value. Technical Report w1815, National Bureau of Economic Research.
- Jaffe, A. B. (1989). Characterizing the “technological position” of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2):87–97.
- Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3):577–598.
- Jake Olivier, W. L. M. and Bell, M. L. (2017). Relative effect sizes for measures of risk. *Communications in Statistics - Theory and Methods*, 46(14):6774–6781.
- Jennings, H. H. (1948). Sociometry in group relations; a work guide for teachers. *American Council on Education*.
- Juozaitienė, R., Seebens, H., Latombe, G., Essl, F., and Wit, E. C. (2022). Analysing ecological dynamics with relational event models: The case of biological invasions. *Network Science*, 11(2):205–223.
- Juozaitienė, R. and Wit, E. C. (2022a). Nodal heterogeneity may induce ghost triadic effects in relational event models. *arXiv preprint arXiv:2203.16386*.
- Juozaitienė, R. and Wit, E. C. (2022b). Non-parametric estimation of reciprocity and triadic effects in relational event networks. *Social Networks*, 68:296–305.
- Juozaitienė, R., Seebens, H., Latombe, G., Essl, F., and Wit, E. C. (2023). Analysing ecological dynamics with relational event models: The case of biological invasions. *Diversity and Distributions*, 29(10):1208–1225.
- Keiding, N. (2014). Event history analysis. *Annual Review of Statistics and Its Application*, 1:333–360.
- Kejžar, N., Korenjak-Černe, S., and Batagelj, V. (2011). Clustering of distributions: A case of patent citations. *Journal of Classification*, 28(2):156–183.
- Kim, B., Schein, A., Desmarais, B. A., and Wallach, H. (2018). The hyperedge event model. *arXiv preprint arXiv:1807.08225*.
- Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

- Kitts, J. A., Lomi, A., Mascia, D., Pallotti, F., and Quintane, E. (2017). Investigating the temporal dynamics of interorganizational exchange: Patient transfers among Italian hospitals. *American Journal of Sociology*, 123(3):850–910.
- Koskinen, J. and Snijders, T. A. B. (2023). Multilevel longitudinal analysis of social networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*.
- Kotthaus, H., Korb, I., Lang, M., Bischl, B., Rahnenführer, J., and Marwedel, P. (2015). Runtime and memory consumption analyses for machine learning R programs. *Journal of Statistical Computation and Simulation*, 85(1):14–29.
- Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46.
- Kuhn, J. (2010). Information overload at the U.S. Patent and Trademark Office: Reframing the Duty of Disclosure in Patent Law as a Search and Filter Problem. *Yale JL & Tech*, 13:89–140.
- Kuhn, J., Younge, K., and Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1):109–132.
- Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago and London.
- Laumann, E. O., Marsden, P. V., and Prensky, D. (1989). The boundary specification problem in network analysis. *Research Methods in Social Network Analysis*, 61(8):18–34.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Lee, J.-S. and Hsiang, J. (2020). Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965.
- Leenders, R. T. A. J., Contractor, N., and DeChurch, L. A. (2016). Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review*, 6(1):92–115.
- Lerner, J. (1994). The importance of patent scope: An empirical analysis. *The RAND Journal of Economics*, 25(2):319–333.

- Lerner, J. and Hâncean, M.-G. (2023). Micro-level network dynamics of scientific collaboration and impact: Relational hyperevent models for the analysis of coauthor networks. *Network Science*, 11(1):5–35.
- Lerner, J., Hâncean, M.-G., and Lomi, A. (2023). Relational hyperevent models for the coevolution of coauthoring and citation networks. *arXiv preprint arXiv:2308.01722*.
- Lerner, J. and Lomi, A. (2017). The third man: Hierarchy formation in Wikipedia. *Applied Network Science*, 2(1):2–24.
- Lerner, J. and Lomi, A. (2019). Team diversity, polarization, and productivity in online peer production. *Social Network Analysis and Mining*, 9(1):1–17.
- Lerner, J. and Lomi, A. (2020a). The free encyclopedia that anyone can dispute: An analysis of the micro-structural dynamics of positive and negative relations in the production of contentious Wikipedia articles. *Social Networks*, 60(1):11–25.
- Lerner, J. and Lomi, A. (2020b). Reliability of relational event model estimates under sampling: How to fit a relational event model to 360 million dyadic events. *Network Science*, 8(1):97–135.
- Lerner, J. and Lomi, A. (2023). Relational hyperevent models for polyadic interaction networks. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*.
- Lerner, J., Lomi, A., Mowbray, J., Rollings, N., and Tranmer, M. (2021). Dynamic network analysis of contact diaries. *Social Networks*, 66:224–236.
- Lévi-Strauss, C. (1971). *The Elementary Structures of Kinship*. Beacon Press, Boston.
- Levin, R. C. (1988). Appropriability, r&d spending, and technological performance. *The American Economic Review*, 78(2):424–428.
- Lin, C.-J., Weng, R. C., and Keerthi, S. S. (2007). Trust region Newton methods for large-scale logistic regression. In *Proceedings of the 24th international conference on Machine learning*, pages 561–568, Corvallis Oregon USA. ACM.
- Liu, B., Cai, Y., Guo, Y., and Chen, X. (2021). Transtailor: Pruning the pre-trained model for improved transfer learning. *arXiv preprint arXiv:2103.01542*.

- Lomi, A. and Bianchi, F. (2021). A time to give and a time to receive: Role switching and generalized exchange in a financial market. *Social Networks*.
- Lomi, A., Mascia, D., Vu, D. Q., Pallotti, F., Conaldi, G., and Iwashyna, T. J. (2014). Quality of care and interhospital collaboration: A study of patient transfers in Italy. *Medical Care*, 52(5):407–414.
- Lospinoso, J. and Snijders, T. A. B. (2019). Goodness of fit for stochastic actor-oriented models. *Methodological Innovations*, 12(3):1–18.
- Lu, L. (2020). Dying ReLU and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706.
- Lusher, D., Koskinen, J., and Robins, G., editors (2012). *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press.
- Machlup, F. (1980). *Knowledge: Its Creation, Distribution and Economic Significance, Volume I: Knowledge and Knowledge Production*. Princeton University Press, Princeton, NJ.
- Marco, A. C. (2007). The dynamics of patent citations. *Economics Letters*, 94(2):290–296.
- Marcum, C. S. and Butts, C. T. (2015). Constructing and modifying sequence statistics for relevant using informR in R. *Journal of Statistical Software*, 64(5):1–36.
- Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16(1):435–463.
- Matias, C., Rebafka, T., and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press New York, New York, NY, USA.
- Meijerink-Bosman, M., Back, M., Geukes, K., Leenders, R. T. A. J., and Mulder, J. (2022a). Discovering trends of social interaction behavior over time: An introduction to relational event modeling. *Behavior Research Methods*, 55(3):1–27.



- Meijerink-Bosman, M., Leenders, R. T. A. J., and Mulder, J. (2022b). Dynamic relational event modeling: Testing, exploring, and applying. *PLOS One*, 17(8):e0272309.
- Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press, Chicago, IL.
- Meyer, M. (2000). What is special about patent citations? differences between scientific and patent citations. *Scientometrics*, 49(1):93–123.
- Meyer, P.-A. (1962). A decomposition theorem for supermartingales. *Illinois Journal of Mathematics*, 6(2):193–205.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Moreno, J. L. (1934). *Who Shall Survive?: A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co, Washington.
- Mulder, J. and Hoff, P. D. (2021). A latent variable model for relational events with multiple receivers. *arXiv preprint arXiv:2101.05135*.
- Mulder, J. and Leenders, R. T. A. J. (2019). Modeling the evolution of interaction behavior in social networks: A dynamic relational event approach for real-time analysis. *Chaos, Solitons & Fractals*, 119:73–85.
- Namin, A. H., Leboeuf, K., Muscedere, R., Wu, H., and Ahmadi, M. (2009). Efficient hardware implementation of the hyperbolic tangent sigmoid function. *2009 IEEE International Symposium on Circuits and Systems*, pages 2117–2120.
- Narayan, S. (1997). The generalized sigmoid activation function: Competitive supervised learning. *Information Sciences*, 99(1):69–82.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.
- Niezink, N. M. and Campana, P. (2022). When things turn sour: A network event study of organized crime violence. *Journal of Quantitative Criminology*.
- Noel, M. M., L, A., Trivedi, A., and Dutta, P. (2023). Growing cosine unit: A novel oscillatory activation function that can speedup training and reduce parameters in convolutional neural networks. *arXiv preprint arXiv:2108.12943*.

- Oancea, B. and Dragoescu, R. M. (2014). Integrating r and hadoop for big data analysis. *arXiv preprint arXiv:1407.4908*.
- Pallotti, F., Weldon, S. M., and Alessandro, L. (2022). Lost in translation: Collecting and coding data on social relations from audio-visual recordings. *Social Networks*, 69:102–112.
- Park, H., Ree, J. J., and Kim, K. (2013). Identification of promising patents for technology transfers using triz evolution trends. *Expert systems with applications*, 40(2):736–743.
- Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Patison, K. P., Quintane, E., Swain, D. L., Robins, G., and Pattison, P. (2015). Time is of the essence: an application of a relational event model for animal social networks. *Behavioral Ecology and Sociobiology*, 69(5):841–855.
- Pattison, P. and Robins, G. (2002). Neighborhood-based models for social networks. *Sociological Methodology*, 32(1):301–337.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849.
- Pilny, A., Proulx, J. D., Dinh, L., and Bryan, A. L. (2017). An adapted structural framework for the emergence of communication networks. *Communication Studies*, 68(1):72–94.

- Pilny, A., Schechter, A., Poole, M. S., and Contractor, N. (2016). An illustration of the relational event model to analyze group interaction processes. *Group Dynamics: Theory, Research, and Practice*, 20(3):181–195.
- Pinheiro, J. and Bates, D. (2006). *Mixed-effects Models in S and S-PLUS*. Springer-Verlag, New York.
- Popper, K. R. (1965). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Basic Books, 2nd edition.
- Preis, T., Schneider, J. J., and Stanley, H. E. (2011). Switching processes in financial markets. *Proceedings of the National Academy of Sciences*, 108(19):7674–7678.
- Quintane, E. and Carnabuci, G. (2016). How do brokers broker? Tertius gaudens, tertius iungens, and the temporality of structural holes. *Organization Science*, 27(6):1343–1360.
- Quintane, E., Conaldi, G., Tonellato, M., and Lomi, A. (2014). Modeling relational events: A case study on an open source software project. *Organizational Research Methods*, 17(1):23–50.
- Quintane, E., Pattison, P., Robins, G., and Mol, J. M. (2013). Short-and long-term stability in organizational networks: Temporal structures of project teams. *Social Networks*, 35(4):528–540.
- Radicchi, F., Fortunato, S., and Vespignani, A. (2012). Citation networks. In *Models of Science Dynamics*, pages 233–257. Springer.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass.
- Rastelli, R. and Corneli, M. (2021). Continuous latent position models for instantaneous interactions. *arXiv preprint arXiv:2103.17146*.
- Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.
- Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134.

- Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Renshaw, S. L., Livas, S. M., Petrescu-Prahova, M. G., and Butts, C. T. (2023). Modeling complex interactions in a disrupted environment: Relational events in the wtc response. *Network Science*, 11(2):295–323.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge ; New York.
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph ( $p^*$ ) models for social networks. *Social Networks*, 29(2):173–191.
- Rothman, D. (2021). *Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More*. Packt Publishing: Birmingham, UK.
- Schaefer, D. R. and Marcum, C. S. (2017). Modeling network dynamics. In Light, R. and Moody, J., editors, *The Oxford Handbook of Social Networks*, chapter 14, pages 254–287. Oxford University Press, New York.
- Schecter, A., Pilny, A., Leung, A., Poole, M. S., and Contractor, N. (2018). Step by step: Capturing the dynamics of work team process through relational event sequences. *Journal of Organizational Behavior*, 39(9):1163–1181.
- Schecter, A. and Quintane, E. (2021). The power, accuracy, and precision of the relational event model. *Organizational Research Methods*, 24(4):802–829.
- Schoenberg, I. (1969). Cardinal interpolation and spline functions. *Journal of Approximation Theory*, 2(2):167–206.
- Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. part b. on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4:112–141.
- Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. R. (2020). Exponential-family models of random graphs: inference in finite, super and infinite population scenarios. *Statistical Science*, 34(4):627–662.

- Scotchmer, S. (1991). Standing on the shoulders of giants: cumulative research and the patent law. *Journal of economic perspectives*, 5(1):29–41.
- Sharma, P and Tripathi, R. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation. *World Patent Information*, 51:31–42.
- Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21(1):149–172.
- Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31(1):361–395.
- Snijders, T. A. B. (2005). Models for longitudinal network data. In *Models and Methods in Social Network Analysis*, volume 1, pages 215–247. Cambridge University Press.
- Snijders, T. A. B. (2017). Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, 4:343–363.
- Snijders, T. A. B., Koskinen, J., and Schweinberger, M. (2010a). Maximum likelihood estimation for social network dynamics. *The Annals of Applied Statistics*, 4(2):567–588.
- Snijders, T. A. B., Pattison, P, Robins, G., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153.
- Snijders, T. A. B., Van de Bunt, G. G., and Steglich, C. (2010b). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60.
- Sorenson, O., Rivkin, J. W., and Fleming, L. (2006). Complexity, networks and knowledge flow. *Research Policy*, 35(7):994–1017.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.
- Stadtfeld, C. and Block, P (2017). Interactions, actors, and time: Dynamic network actor models for relational events. *Sociological Science*, 4:318–352.

- Stadtfeld, C. and Geyer-Schulz, A. (2011). Analyzing event stream dynamics in two-mode networks: An exploratory analysis of private communication in a question and answer community. *Social Networks*, 33(4):258–272.
- Stadtfeld, C., Hollway, J., and Block, P. (2017). Dynamic network actor models: Investigating coordination ties through time. *Sociological Methodology*, 47(1):1–40.
- Stadtfeld, C., Vörös, A., Elmer, T., Boda, Z., and Raabe, I. J. (2019). Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences*, 116(3):792–797.
- Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., Van den Broeck, W., Régis, C., Lina, B., et al. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176.
- Temizsoy, A., Iori, G., and Montes-Rojas, G. (2015). The role of bank relationships in the interbank market. *Journal of Economic Dynamics and Control*, 59:118–141.
- Thomas, D. C. (1981). General relative-risk models for survival time and matched case-control analysis. *Biometrics*, 37(4):673–686.
- Tonellato, M., Tasselli, S., Conaldi, G., Lerner, J., and Lomi, A. (2023). A microstructural approach to self-organizing: The emergence of attention networks. *Organization Science*.
- Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, pages 172–187.
- Trajtenberg, M. and Jaffe, A. B. (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. The MIT Press, Cambridge, Massachusetts, United States.
- Tranmer, M., Marcum, C. S., Morton, F. B., Croft, D. P., and de Kort, S. R. (2015). Using the relational event model (rem) to investigate the temporal dynamics of animal social networks. *Animal Behaviour*, 101:99–105.
- Tsukagoshi, H., Sasano, R., and Takeda, K. (2021). DefSent: Sentence embeddings using definition sentences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

- Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–418. Association for Computational Linguistics.
- Tsukagoshi, H., Sasano, R., and Takeda, K. (2022). Comparison and Combination of Sentence Embeddings Derived from Different Supervision Signals. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 139–150.
- Tuma, N. B. and Hannan, M. T. (1984). *Social Dynamics Models and Methods*. Harcourt Brace Jovanovic Publishers, San Diego.
- Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
- Uzaheta, A., Amati, V., and Stadtfeld, C. (2023). Random effects in dynamic network actor models. *Network Science*, 11(2):249–266.
- Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42(1):35–67.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.
- Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, 45(3):707–723.
- Veugelers, R. and Wang, J. (2019). Scientific novelty and technological impact. *Research Policy*, 48(6):1362–1372.
- Vieira, F., Leenders, R. T. A. J., McFarland, D., and Mulder, J. (2022). Bayesian mixed-effect models for independent dynamic social network data. *arXiv preprint arXiv:2204.10676*.
- Vinciotti, V. and Wit, E. (2017). Preface to the themed issue on “Networks and Society”. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 66(3):451–453.
- Vu, D. Q., Asuncion, A. U., Hunter, D. R., and Smyth, P. (2011). Dynamic Ego-centric Models for Citation Networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 857–864.

- Vũ, D. Q., Lomi, A., Mascia, D., and Pallotti, F. (2017). Relational event models for longitudinal network data with an application to interhospital patient transfers. *Statistics in Medicine*, 36(14):2265–2287.
- Vũ, D. Q., Pattison, P., and Robins, G. (2015). Relational event models for social learning in moocs. *Social Networks*, 43:121–135.
- Wang, J. and Chen, Y.-J. (2019). A novelty detection patent mining approach for analyzing technological opportunities. *Advanced Engineering Informatics*, 42:100941.
- Wang, M.-Y., Chang, D.-S., and Kao, C.-H. (2010). Identifying technology trends for r&d planning using triz and text mining. *R&D Management*, 40(5):491–509.
- Wang, P., Robins, G., Pattison, P., and Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, 35(1):96–115.
- Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to markov graphs and  $p^*$ . *Psychometrika*, 61(3):401–425.
- Welles, B. F., Vashevko, A., Bennett, N., and Contractor, N. (2014). Dynamic models of communication in an online friendship network. *Communication Methods and Measures*, 8(4):223–243.
- Whalen, R., Lungeanu, A., DeChurch, L., and Contractor, N. (2020). Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17(3):615–639.
- White, H. C. (1963). *An Anatomy of Kinship: Mathematical Models for Structures of Cumulated Roles*. Prentice-Hall series in mathematical analysis of social behavior. Prentice-Hall, NJ.
- Wiemers, J. and Neyer, U. (2003). Why do we have an interbank money market? Technical report, IWH Discussion Papers.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.
- Wood, S. N. (2017). *Generalized Additive Models: an Introduction with R (2nd ed.)*. CRC Press, Boca Raton.



- Wu, L., Waber, B. N., Aral, S., Brynjolfsson, E., and Pentland, A. (2008). Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an IT configuration task. *Available at SSRN 1130251*.
- Yan, B. and Luo, J. (2017). Measuring technological distance for patent mapping. *Journal of the Association for Information Science and Technology*, 68(2):423–437.
- Yoon, J. and Kim, K. (2012). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*, 90(2):445–461.
- Younge, K. A. and Kuhn, J. M. (2016). Patent-to-Patent Similarity: A Vector Space Model. *SSRN Electronic Journal*.
- Zachrison, K. S., Amati, V., Schwamm, L. H., Yan, Z., Nielsen, V., Christie, A., Reeves, M. J., Sauser, J. P., Lomi, A., and Onnela, J.-P. (2022). Influence of hospital characteristics on hospital transfer destinations for patients with stroke. *Circulation: Cardiovascular Quality and Outcomes*, 15(5):e008269.
- Zappa, P. and Vu, D. Q. (2021). Markets as networks evolving step by step: Relational event models for the interbank market. *Physica A: Statistical Mechanics and its Applications*, 565(C).
- Zeiler, M. D. (2012). ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701*.