

A stochastic gradient relational event additive model for modelling US patent citations from 1976 to 2022

Edoardo Filippi-Mazzola  and Ernst C. Wit 

Faculty of Informatics, Institute of Computing, Università della Svizzera italiana, Lugano Switzerland

Address for correspondence: Edoardo Filippi-Mazzola, Faculty of Informatics, Institute of Computing, Università della Svizzera italiana, Via la Santa 1, Lugano 6900, Switzerland. Email: edoardo.filippi-mazzola@usi.ch

Abstract

Until 2022, the US patent citation network contained almost 10 million patents and over 100 million citations, presenting a challenge in analysing such expansive, intricate networks. To overcome limitations in analysing 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 modelling the effect of each predictor non-linearly. Overall, our model identifies key factors driving patent citations and reveals insights in the citation process.

Keywords: B-splines, citation networks, patent analysis, relational event models, stochastic gradient descent

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 & 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 & 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). Analysing 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 a variety of 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 analysing large and complex datasets.

The relational event model (REM) (Butts, 2008; Perry & 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 behavioural 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 analysing citation networks of patents, as they allow us to

Received: April 24, 2023. Revised: February 26, 2024. Accepted: April 16, 2024

© The Royal Statistical Society 2024.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

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 (2020) 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. (2021) introduced non-linear effects to model non-linear structures. Nevertheless, as shown by Bauer et al. (2022), introducing non-linear effects when REMs 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 modelling approach, where each predictor can be represented by a smooth effect through B-splines (De Boor, 1972; Schoenberg, 1946, 1969). To address the estimation challenges in large networks, particularly when using smooth effects, STREAM employs the adaptive moment (ADAM) optimizer (Kingma & 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 are 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 paper, we start by describing the USPTO patent data in Section 2 on which this analysis is based. After developing the theoretical foundation in Section 3, we apply the framework to the patent citation network in Section 4. Although STREAM was specifically designed to work with citation networks, this modelling framework can easily be applied to model general relational event data.

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 fulfil the duty of disclosure. 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 USPTO. In the latter, the examiner committee has to integrate additional documents and patent citations. European Patent Office 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 & 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 patents issued only by the USPTO means losing track of those citations that go to patents outside the US jurisdictions. Second, 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 pre-processing procedures for recreating the dataset can be found at https://github.com/efm95/STREAM/tree/main/data_preprocessing.

Figure 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.

3 Stochastic gradient relational event additive model

Relational event models are a class of statistical models used to analyse 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 STREAM for the network of patent citations.

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

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

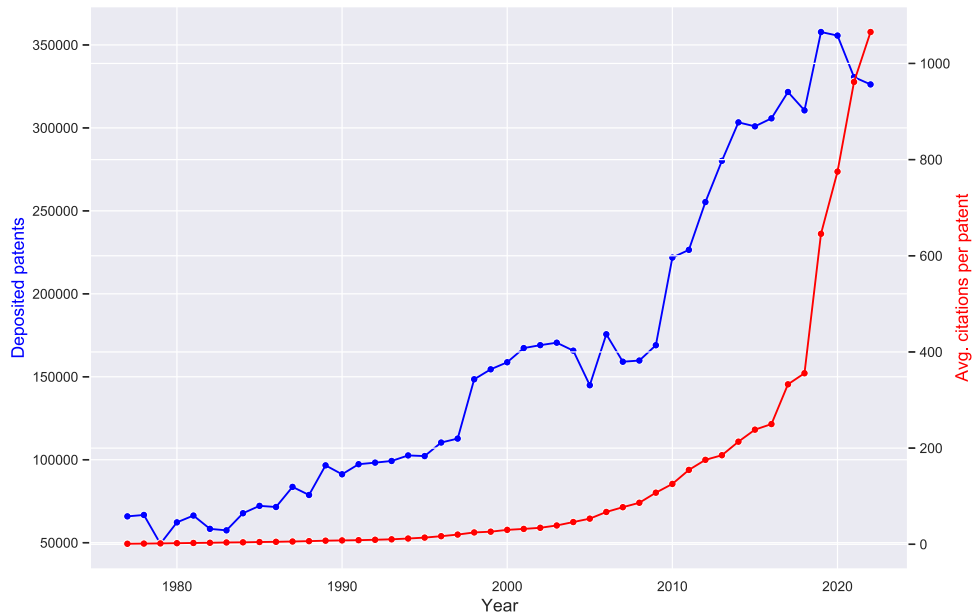


Figure 1. Number of deposited patents per year and the number of patent citations per patent per year since 1976.

The counting process $N_{sr}(t)$ is a local sub-martingale 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 λ_0 and an exponential function of q covariates $x_{sr}(t)$ with corresponding parameter β , i.e.

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

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 analysed 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 Eq. (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 $R(t \mid \mathcal{H}_{t^-})$ 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:

$$PL(\beta) = \prod_{i=1}^n \left(\frac{\exp\{\sum_{k=1}^q \beta_k x_{s_i r_i k}(t_i)\}}{\sum_{r \in R(t_i | \mathcal{H}_{t_i^-})} \exp\{\sum_{k=1}^q \beta_k x_{s_i r k}(t_i)\}} \right). \tag{2}$$

In its logarithmic form, Eq. (2) assumes a concave behaviour, allowing the coefficients to be estimated via a Newton approach. The partial likelihood in Eq. (2) is directly inspired by the model presented by Butts (2008) for temporal ordinal data. Indeed, Eq. (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.

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 analyse 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 (2020) suggests that estimates are reliable with just one control per case. With a single control, the denominator in Eq. (2) is the sum of the rates for the cited patent with covariates x_{s,r_i} and a randomly sampled non-cited patent with covariates x_{s,r_i^*} . Then the sampled case-control version of the partial likelihood (2) is given as

$$\tilde{PL}(\beta) = \prod_{i=1}^n \left(\frac{\exp\left\{\sum_{k=1}^q \beta_k \left(x_{s,r_i,k}(t_i) - x_{s,r_i^*,k}(t_i)\right)\right\}}{1 + \exp\left\{\sum_{k=1}^q \beta_k \left(x_{s,r_i,k}(t_i) - x_{s,r_i^*,k}(t_i)\right)\right\}} \right), \quad (3)$$

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

3.3 Basis expansion of covariates

The core assumption of relational event modelling 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 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 & 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 $x_{srk}(t)$, 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 online supplementary material S1), the B-spline effect associated to the k th covariate x_{srk} is then a linear combination of d coefficients with d basis functions, i.e.

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

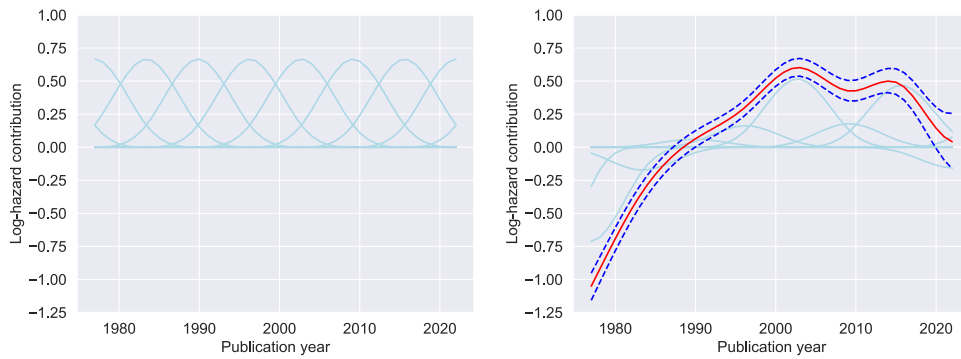


Figure 2. Splines associated to the *nodal effects*. Left: *receiver-publication year* uniform basis transformation. Right: *receiver-publication year* estimated effect, where the basis model matrix is multiplied with its respective vector of estimated coefficients.

Figure 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 Eq. (1), the full model for the intensity function becomes

$$\lambda_{sr}(t | H_{t^-}) = \lambda_0(t) e^{\sum_{k=1}^q \hat{f}_k(x_{srk}(t)) \mathbb{1}_{\{(s,r) \in R(t_i | \mathcal{H}_{t^-})\}}}, \tag{4}$$

where its partial likelihood approximation with one control is given as

$$\tilde{P}L(\theta) = \prod_{i=1}^n \left[1 + \exp \left\{ - \sum_{k=1}^q \sum_{j=1}^d \theta_{jk} \left[B_{j,p}^k(x_{s_i r_i k}(t_i)) - B_{j,p}^k(x_{s_i r_i^* k}(t_i)) \right] \right\} \right]^{-1}. \tag{5}$$

To smooth the estimated B-splines resulting from maximizing the partial likelihood in Eq. (5), various penalization terms can be added. One reliable option is the use of P-splines (Eilers & 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.

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

$$\hat{\Lambda}_0(t) = \sum_{t_i < t} \left[\exp \left\{ \sum_{k=1}^q \hat{f}_k(x_{s_i r_i k}(t_i)) \right\} + \exp \left\{ \sum_{k=1}^q \hat{f}_k(x_{s_i r_i^* k}(t_i)) \right\} \right]^{-1} \frac{2}{n(t_i)}, \tag{6}$$

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 point-wise baseline hazard estimate can be calculated by taking differences between subsequent events of the cumulative hazard, i.e.

$$\hat{\lambda}_0(t_i) = \frac{\hat{\Lambda}_0(t_i) - \hat{\Lambda}_0(t_{i-1})}{t_i - t_{i-1}}, \quad \text{for } i = 1, \dots, n. \quad (7)$$

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 (SGD) approach through the ADAM optimizer to fit the partial likelihood. Stochastic gradient descent has proved to be a reliable technique for estimating logistic regression models in large-scale scenarios (Bottou, 2010; Lin et al., 2007). 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. Among these, ADAM has gained in popularity in the machine-learning field for its scalability and its convergence reliability. 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 $\nabla \tilde{PL}(\theta)_b$ 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:

$$\begin{aligned} m_b &\leftarrow \zeta_1 m_{b-1} + (1 - \zeta_1) \nabla \tilde{PL}(\theta)_b \\ v_b &\leftarrow \zeta_2 v_{b-1} + (1 - \zeta_2) \nabla \tilde{PL}(\theta)_b^2, \end{aligned}$$

where m and v are the first- and second-moment gradients, respectively, and ζ_1 and ζ_2 are hyper-parameters 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:

$$\begin{aligned} \hat{m}_{b,s} &= \frac{m_b}{1 - \zeta_1^s} \\ \hat{v}_{b,s} &= \frac{v_b}{1 - \zeta_2^s}, \end{aligned}$$

where as s increases, ζ_1^s and ζ_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 ψ is the learning rate that determines the magnitude of each parameter update, while ϵ represents a small scalar added to prevent division by zero (usually 1e-8). For our application, ψ 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 Modelling 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.

4.1 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 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 [online supplementary material S1](#).

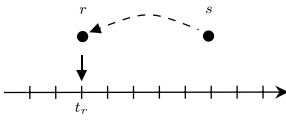
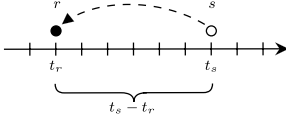
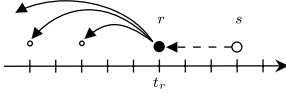
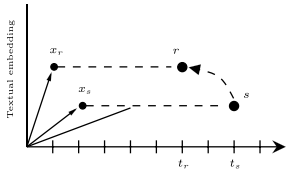
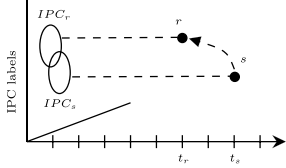
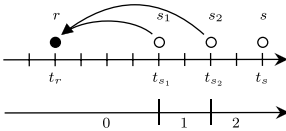
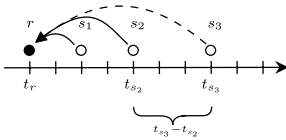
4.1.1 Node effects

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 difference between the issue dates of the citing patent and the cited patent can also be a factor in driving 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

Table 1. Effects and their corresponding mechanisms

Effect	Mechanism	Statistic
Receiver-publication year		t_r
Time lag		$t_s - t_r$
Receiver outdegree		$\sum_{r' \in R(t_r \mathcal{H}_{t_r}^r)} \mathbb{1}_{\{(r, r', t_r)\}}$
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		$\min_{t_i < t_s} \{(t_s - t_i) \mid \mathbb{1}_{\{(s, r, t_i)\}} = 1\}$

Note. r represents the cited patent, i.e. the receiver, s represents the citing patent, i.e. the sender. t_s and t_r , respectively, represent the issue date for the sender and for the receiver. \mathbf{x}_r and \mathbf{x}_s are the patent abstract embedding vectors, while IPC_r and IPC_s are vectors of IPC patent classes.

centrality. In this context, the outdegree of a patent represents the number of patents itself cites at the time of its publication.

4.1.2 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 (Filippi-Mazzola et al., 2023; Kuhn et al., 2020) 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 & Gurevych, 2019](#)). It uses a vectorized loop to compute pair-wise 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 & Kuhn, 2015](#)), 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 & 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 analysing patent classes.

4.1.3 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 $x_{s^*k}(t)$ 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 behaviour over time.

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. 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 (5), this involves $2q$ matrices for both cases and controls.

The model fitting problem will, therefore, be divided into two parts: (a) defining an efficient way to compute the basis function for millions of rows and (b) 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 (5) as our loss metric. The code for STREAM can be found at <https://github.com/efm95/STREAM>.

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 2–5 show the fitted splines with 10 degrees of freedom. Uncertainty estimates are provided via point-wise 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.

4.3.1 Node effects

One remarkable result can be seen in the *receiver-publication year* curve in Figure 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 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 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 3–7 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 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

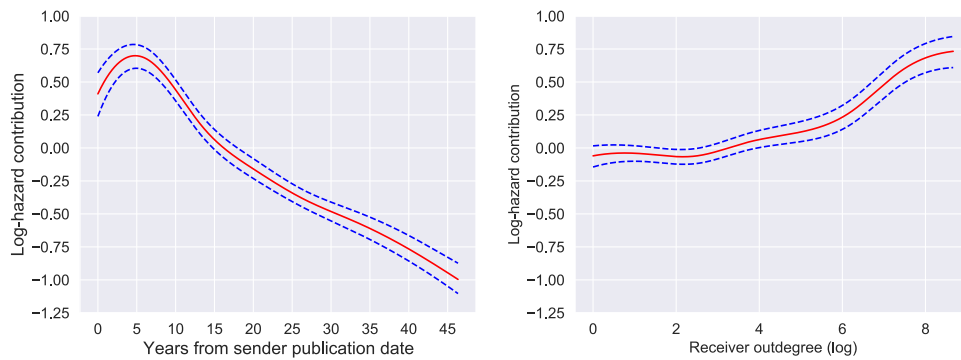


Figure 3. Splines associated to the *nodal effects*. Left: *time lag*. Right: *receiver outdegree* in log terms.

patents in one's patent application, as it makes a patent more visible and accessible to other inventors, increasing the likelihood of 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.

4.3.2 Similarity effects

The curves for both *textual similarity* and *IPC relatedness* shown in Figure 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 & 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 with greater technological similarity remain more likely to cite each other.

4.3.3 Time-varying effects

The two time-varying effects in Figure 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.

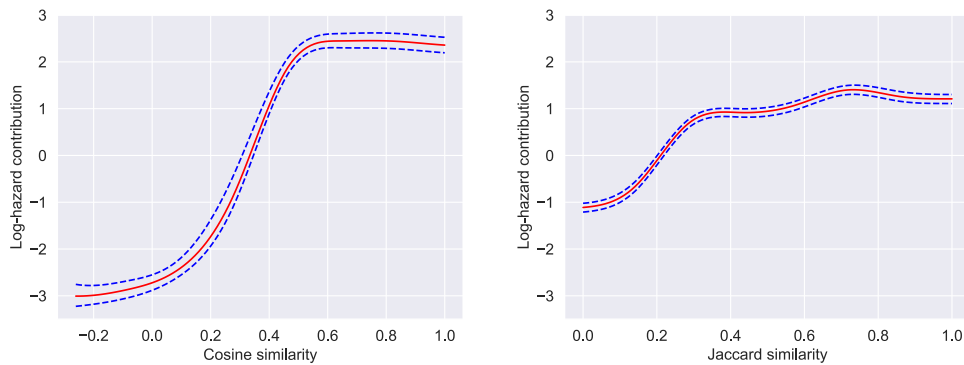


Figure 4. Splines associated to the *similarity effects*. Left: *textual similarity*. Right: *IPC relatedness*.

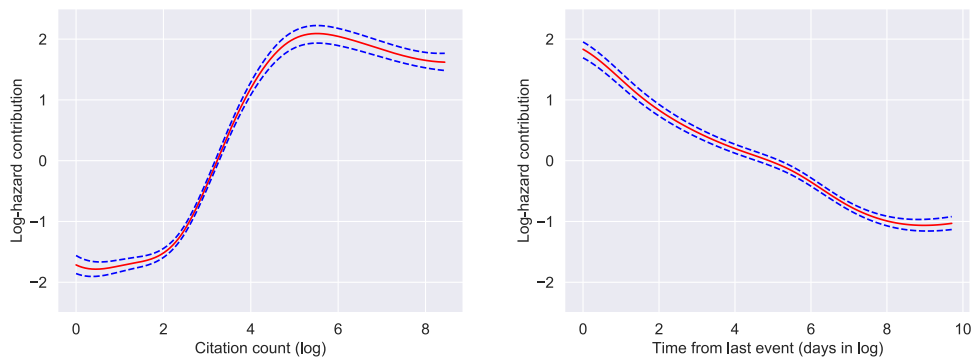


Figure 5. Splines associated with the *time-varying effects*. Left: *cumulative citations received in log scale*. Right: *time from last event in days in log scale*.

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 modelling this effect.

4.4 Estimated baseline hazard

To analyse 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 Eq. (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 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

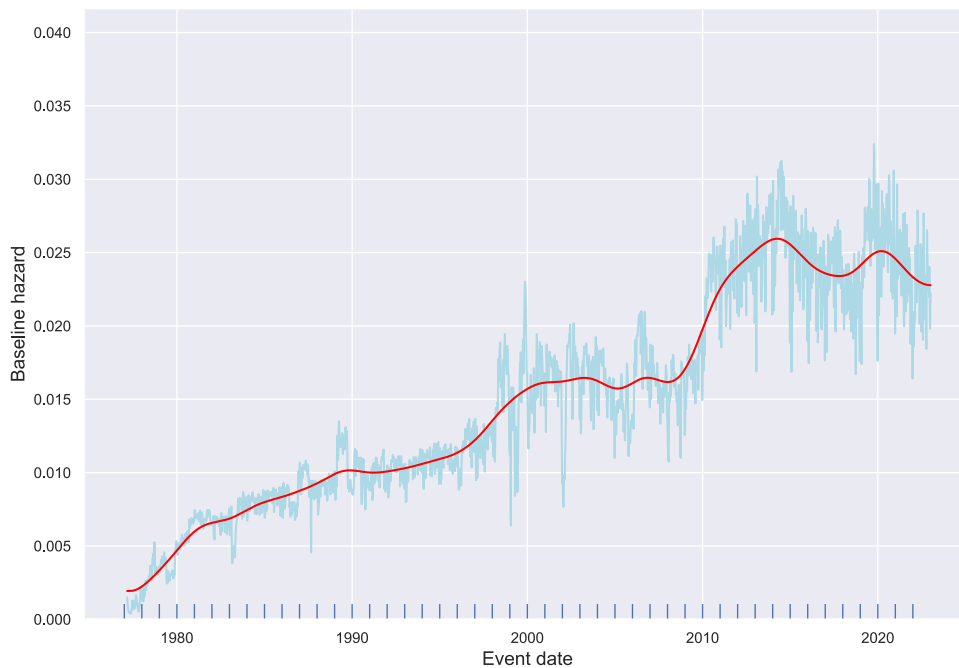


Figure 6. Baseline hazard estimated through the adapted estimator in Eq. (7) from [Borgan et al. \(1995\)](#). The smoother line is the application of a Gaussian filter to smooth the resulting estimate and capture the general trend of the baseline.

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 analysing 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.

5 Conclusions

Relational event models are a sophisticated and effective approach for analysing 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 modelling 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 STREAM. 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 behaviour of, e.g. the *textual similarity* curve in Figure 4 over the observed period, particularly in light of recent discussions on changes in the generative process of patent citations (Filippi-Mazzola et al., 2023; Kuhn et al., 2020). 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 (2022) could be further investigated to be applied to the STREAM to assess the temporal decay of predictors.

In this work, the citation dynamics are modelled 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 modelling polyadic interactions among patents. Indeed, the flexibility of the STREAM approach could be combined with the newly proposed relational hyper-event model (Lerner & Lomi, 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 modelling complex non-linear effects in large event networks.

Acknowledgments

The authors are thankful to the editor, the associate editor, and two anonymous reviewers for their valuable comments.

Conflict of interest: The authors declare that there are no conflicts.

Funding

This work was supported by funding from the Swiss National Science Foundation (grant 192549).

Data availability

Data are fetched directly from the USPTO (<https://bulkdata.uspto.gov/>) through fastpat (<https://github.com/iamlemec/fastpat>). Data from 1976 to 2022 are available as CSV files on Kaggle (<https://www.kaggle.com/datasets/filippimazz/patents-citations>). The pre-processing approach for this data is available on our repository (<https://github.com/efm95/STREAM>).

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

References

- Bacchiocchi E., & Montobbio F. (2010). International knowledge diffusion and home-bias effect: Do USPTO and EPO patent citations tell the same story?: International knowledge diffusion and home-bias effect. *Scandinavian Journal of Economics*, 112(3), 441–470. <https://doi.org/10.1111/j.1467-9442.2010.01614.x>
- Bauer V., Harhoff D., & Kauermann G. (2022). A smooth dynamic network model for patent collaboration data. *Advances in Statistical Analysis*, 106, 97–116. <https://doi.org/10.1007/s10182-021-00393-w>
- Bianchi F., Filippi-Mazzola E., Lomi A., & Wit E. C. (2024). Relational event modeling. *Annual Review of Statistics and Its Application*, 11(1), 297–319. <https://doi.org/10.1146/annurev-statistics-040722-060248>
- Bianchi F., Stivala A., & Lomi A. (2022). Multiple clocks in network evolution. *Methodological Innovations*, 15(1), 29–41. <https://doi.org/10.1177/20597991221077877>
- Borgan O., Goldstein L., & 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. <https://doi.org/10.1214/aos/1176324322>
- Bottou L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Physica-Verlag HD.
- Brandes U., Lerner J., & Snijders T. A. (2009). Networks evolving step By step: Statistical analysis of dyadic event data. In *2009 International Conference on Advances in Social Network Analysis and Mining* (pp. 200–205). IEEE.

- Butts C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1), 155–200. <https://doi.org/10.1111/j.1467-9531.2008.00203.x>
- Cox D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- Cox D. R. (1975). Partial likelihood. *Biometrika*, 62(2), 269–276. <https://doi.org/10.1093/biomet/62.2.269>
- De Boor C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1), 50–62. [https://doi.org/10.1016/0021-9045\(72\)90080-9](https://doi.org/10.1016/0021-9045(72)90080-9)
- Duchi J., Hazan E., & Singer Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2121–2159. <https://doi.org/10.5555/1953048.2021068>
- Eilers P. H. C., & Marx B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–121. <https://doi.org/10.1214/ss/1038425655>
- Ernst H. (2003). Patent information for strategic technology management. *World Patent Information*, 25(3), 233–242. [https://doi.org/10.1016/S0172-2190\(03\)00077-2](https://doi.org/10.1016/S0172-2190(03)00077-2)
- Filippi-Mazzola E., Bianchi F., & Wit E. C. (2023). Drivers of the decrease of patent similarities from 1976 to 2021. *PLoS One*, 18(3), 1–13. <https://doi.org/10.1371/journal.pone.0283247>
- Foucault Welles B., Vashevko A., Bennett N., & Contractor N. (2014). Dynamic models of communication in an online friendship network. *Communication Methods and Measures*, 8(4), 223–243. <https://doi.org/10.1080/19312458.2014.967843>
- Fritz C., Thurner P. W., & Kauermann G. (2021). Separable and semiparametric network-based counting processes applied to the international combat aircraft trades. *Network Science*, 9(3), 291–311. <https://doi.org/10.1017/nws.2021.9>
- Hastie T., & Tibshirani R. (1986). Generalized additive models. *Statistical Science*, 1(3), 297–310. <https://doi.org/10.1214/ss/1177013604>
- Juozaitienė R., Seebens H., Latombe G., Essl F., & Wit E. C. (2023). Analysing ecological dynamics with relational event models: The case of biological invasions. *Diversity and Distributions*, 29(10), 1208–1225. <https://doi.org/10.1111/ddi.13752>
- Juozaitienė R., & Wit E. C. (2022). Non-parametric estimation of reciprocity and triadic effects in relational event networks. *Social Networks*, 68, 296–305. <https://doi.org/10.1016/j.socnet.2021.08.004>
- Kuhn J., Younge K., & Marco A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, 51(1), 109–132. <https://doi.org/10.1111/1756-2171.12307>
- Kuhn J. M. (2010). Information overload at the US Patent and Trademark Office: Reframing the duty of disclosure in patent law as a search and filter problem. *Yale Journal of Law & Technology*, 13, 90–139. <http://hdl.handle.net/20.500.13051/7775>
- Lerner J. (1994). The importance of patent scope: An empirical analysis. *The RAND Journal of Economics*, 25(2), 319–333. <https://doi.org/10.2307/2555833>
- Lerner J., & Lomi A. (2020). 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. <https://doi.org/10.1017/nws.2019.57>
- Lerner J., & Lomi A. (2023). Relational hyperevent models for polyadic interaction networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(3), 577–600. <https://doi.org/10.1093/jrsssa/qnac012>
- Lin C.-J., Weng R. C., & Keerthi S. S. (2007). Trust region Newton methods for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 561–568). ACM.
- Park M., Leahey E., & Funk R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942), 138–144. <https://doi.org/10.1038/s41586-022-05543-x>
- 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., ... Chintala S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* 32 (pp. 8024–8035). Curran Associates, Inc.
- Perry P. O., & 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. <https://doi.org/10.1111/rssb.12013>
- 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. <https://doi.org/10.1007/s100510050359>
- Reimers N., & Gurevych I. (2019, November). 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.
- Schoenberg I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of oscillatory interpolation. A second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2), 112–141. <https://doi.org/10.1090/qam/1946-04-02>

- Schoenberg I. J. (1969). Cardinal interpolation and spline functions. *Journal of Approximation Theory*, 2(2), 167–206. [https://doi.org/10.1016/0021-9045\(69\)90040-9](https://doi.org/10.1016/0021-9045(69)90040-9)
- Scotchmer S. (1991). Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of Economic Perspectives*, 5(1), 29–41. <https://doi.org/10.1257/jep.5.1.29>
- Sharma P., & Tripathi R. C. (2017). Patent citation: A technique for measuring the knowledge flow of information and innovation. *World Patent Information*, 51, 31–42. <https://doi.org/10.1016/j.wpi.2017.11.002>
- Trajtenberg M., & Jaffe A. B. (2002). *Patents, citations, and innovations: A window on the knowledge economy*. The MIT Press.
- Tranmer M., Marcum C. S., Morton F. B., Croft D. P., & 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. <https://doi.org/10.1016/j.anbehav.2014.12.005>
- Uzzi B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42(1), 35–67. <https://doi.org/10.2307/2393808>
- Vu D., Lomi A., Mascia D., & Pallotti, F. (2017). Relational event models for longitudinal network data with an application to interhospital patient transfers. *Statistics in Medicine*, 36(14), 2265–2287. <https://doi.org/10.1002/sim.7247>
- Vu D., Pattison P., & Robins G. (2015). Relational event models for social learning in MOOCs. *Social Networks*, 43, 121–135. <https://doi.org/10.1016/j.socnet.2015.05.001>
- Vu, D. Q., Asuncion A. U., Hunter D. R., & Smyth P. (2011). Dynamic egocentric models for citation networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 857–864). Omnipress, 2600 Anderson St, Madison, WI, United States.
- Welles B. F., Vashevko A., Bennett N., & Contractor N. (2014). Dynamic models of communication in an online friendship network. *Communication Methods and Measures*, 8(4), 223–243. <https://doi.org/10.1080/19312458.2014.967843>
- Whalen R., Lungeanu A., DeChurch L., & Contractor N. (2020). Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17(3), 615–639. <https://doi.org/10.1111/jels.v17.3>
- 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. <https://doi.org/10.1111/j.1467-9868.2010.00749.x>
- Yan B., & Luo J. (2017). Measuring technological distance for patent mapping. *Journal of the Association for Information Science and Technology*, 68(2), 423–437. <https://doi.org/10.1002/asi.2017.68.issue-2>
- Younge K. A., & Kuhn J. M. (2015). Patent-to-patent similarity: A vector space model. *SSRN Electronic Journal*. <http://dx.doi.org/10.2139/ssrn.2709238>. <https://ssrn.com/abstract=2709238>.