Peer effect analysis with latent processes

Vincent Starck*
LMU München

November 4, 2024

Abstract

I propose a framework to analyze peer effects by modeling a latent sequence of decisions in continuous time. The method avoids linear-in-means regression or regression on conditional expectations – and thus reflection type problems (Manski, 1993) – by modeling the (unobserved) direction of causality, whose probability can be identified. I propose a parsimonious parametric specification to incorporate covariates and define a peer effect parameter meant to capture the causal peer influence of first-movers. The parameters are shown to be consistently estimated by maximum of likelihood methods and lends itself to standard inference under repeated network asymptotics.

Keywords: Peer effects, Continuous time, Causal Inference, Networks.

1 Introduction

Peer effects have traditionally been investigated via estimation of models of the form $\mathbb{E}[y|x,z] = f(\mathbb{E}[y|x],z)$, where x are characteristics used to define peer group and z are covariates directly affecting the outcome. Nevertheless, this type of model – even in nonlinear forms – is tautological as it suffers from the reflection problem (Manski, 1993), which threatens its well-definedness and leads to identification issues.

Spatial Autoregressive (SAR; Lee (2004)) or linear-in-means models bypass some of these concerns. Reflection issues disappear when the conditional expectation is

*I gratefully acknowledge financial support from the European Research Council (Starting Grant No. 852332)

replaced by a weighted average of outcomes if these are not meant to estimate the conditional expectation, but rather to characterize a system's equilibrium relationship. Relatedly, the use of non-overlapping peer groups alleviates reflection concerns insofar as they allow for identification of peer effects (Bramoullé, Djebbari and Fortin, 2009; De Giorgi, Pellizzari and Redaelli, 2010).

Nevertheless, these models suffer from a few shortcomings in practice and may generate spurious peer effects (Angrist, 2014). In addition, while such models naturally represent simultaneous decisions or equilibrium results, they are less suited for sequential, irreversible decisions. They do not provide a framework to discuss causality questions, time dynamics, or identification of the (probability of the) sequence of moves. For some applications, it is more natural to model peer effects as sequentially triggered by over time: first movers exert social influence on their peers.

I propose a new framework to analyze peer effects and model the latent sequence of decisions in continuous time. Modeling the latent sequence of decisions provides a few advantages. It offers a natural framework to discuss causality, as it breaks the simultaneity by disentangling first-movers that potentially generate a causal reaction and the subjects of influence[1]. It also enables a discussion of the diffusion process and counterfactual analysis about, say, the expected time before some fraction of the population adopts some technology or the impact of policy interventions.

Finally, it provides some identifying power: it is possible to identify peer effects from overlapping peer groups such as classrooms. It can also be easier to separate endogenous and exogenous peer effects. Exogenous peer effects (regressors of the form WX, where W is the adjacency matrix) are treated as another regressor while endogenous peer effects are triggered over time. Partial knowledge of the identity of first-movers or adoption times can be incorporated as well, providing additional information.

I formalize the counterfactual framework using potential outcomes and I define a causal peer effect parameter that can be consistently estimated by maximum of likelihood methods. Decisions are explicitely made asynchronously, though their timing may not be observed. The order of arrival may generally be of interest, not only for its own sake or because it contributes to our understanding of peer effects and has

[1]If there is no information about timing of adoption, the first-movers are not identified. It is however possible to identify the probability of an order of adoption given individuals' characteristics.

potential policy-relevance, *e.g.*, for targeting and diffusion.

The paper applies to both spatial and network applications, and as such has also connections to diffusion problems in the network literature (*e.g.*, He and Song (2018)). It may also be useful to analyze staggered treatment adoption (Shaikh and Toulis, 2021; Athey and Imbens, 2022) by relaxing the standard assumption that treatment adoptions arise independently.

2 Causality and peer effects

The goal is to model and estimate peer effects, *i.e.* how decisions of peers alter the decision probability of an individual. I focus on irreversible[2] decisions with initial state 0 and the decision to opt-in (switch to 1) is irreversible. For instance, the outcome y might represent vaccination status, technology adoption, retirement, etc.

Observing someone opting in modifies the likelihood that others do so. This can be because of conformity, information transmission, or other types of social influence. This implies that there is a shift in the distribution of the adoption time of individual i, T_i, upon observing the adoption of a peer.

Modeling the sequence of latent decisions offers a few advantages here. Besides some identifying power and the possibility to discuss dynamics over time, it will allow for a parametric specification where covariates are trivially incorporated and a single causal peer effect parameter can be identified.

Adapting the potential outcome notation (Neyman, 1923; Rubin, 1974) to the current setup[3], the adoption time is represented by $T_i(\tau)$ for $\tau \in \mathbb{R}^{n-1}$: the adoption time of individual i depends on the adoption times of other individuals. This dependence carries over to the outcomes $\mathbb{1}_{T_i<t}$, which are observed for $t = S > 0$.

[2]This can be because the action cannot be undone (*e.g.*, vaccination) or because the focus is on first-time events.

[3]Potential outcomes have been used to discuss causal peer effects in different contexts. The literature on interference ("exogeneous peer effects" or sometimes spillovers), where one's outcome depends on neighbors' treatment, addresses (versions of) the issue and identifies direct and indirect effects under various relaxations of SUTVA (Toulis and Kao, 2013; Sofrygin and van der Laan, 2016; Aronow and Samii, 2017; Arpino, Benedictis and Mattei, 2017; Liu et al., 2019; Forastiere, Airoldi and Mealli, 2020; Jackson, Lin and Yu, 2020; Sánchez-Becerra, 2021; Huber and Steinmayr, 2021). Potential outcome that depends on peer's outcome (*e.g.*, in Egami and Tchetgen Tchetgen (2024)) have been the object of less discussion. To my knowledge, the discussion of peer effects through the time dimension is new.

Another parameter of interest is the analog of average treatment effect: the mean change in the outcome upon a change in τ:

$$\delta(\tilde{\tau}, \tau) \stackrel{\text{def}}{=} \mathbb{E}[Y_i(\tilde{\tau}) - Y_i(\tau)] = \mathbb{P}[Y_i(\tilde{\tau}) = 1] - \mathbb{P}[Y_i(\tau) = 1] \tag{1}$$

For instance, one can be interested in $\tilde{\tau} = \vec{0} \in \mathbb{R}^{n-1}$ and $\tau = \infty e_i$[4] or $\tau = \vec{\infty}$ so that the parameter describes the change in adoption probability induced by the initial adoption of one or all peers compared to them never adopting.

We could also be interested in counterfactual effects such as the expected adoption time:

$$\text{EAT} = \mathbb{E}[T_i] \tag{2}$$

or related quantities such as conditional versions or the expected time before a fraction of the population adopts the technology.

The tools of causal inference suggests a few strategies to identify parameters of interest. Specifically, identification appears possible with repeated networks under randomization on times T_{-i}. Because of continuous time, however, one will also need either homogeneity-continuity or the ability to assign exact times. Identification would then require conditions such as

Assumption 2.1 (Unconfoundedness of timings)**.** $T_{-i} \perp\!\!\!\perp \{T_i(\tau), \tau\} | X, W$

where T_{-i} is the set of times of all individuals but i.

This assumption is often strong with because of the endogeneity of the network to individual characteristics, often in the form of homophily bias (Shalizi and Thomas, 2011). Observational data thus requires controlling for possible confounders. Modeling network formation can offer avenues for improvement, see **??**; large networks may offer further opportunities to identify latent variables (**??**). The assumption also holds trivially with randomization of assigned times, independently of covariates.

This assumption typically has to be supplemented by some structure on time dynamics, either because one needs to extrapolate results and discuss future diffusion or because randomizing imposes practical limitations, *e.g.*, limited range of feasible values such as $\tau = \vec{0}$.

[4] e_i is a vector whose only nonzero entry is a 1 in place i. I adopt the convention $0 \cdot \infty = 0$.

In the next section, I formalize a latent process that enables maximum of likelihood estimation.

3 Model and Likelihood

3.1 Illustration and identification with 2 individuals

Consider two individuals ($i = 1, 2$) who can make an irreversible decision at any point during a time frame represented by the interval $[0; S]$, where 0 is the start of eligibility and S is the time to observation, and indicate adoption by $y_i = 1$. Let the distribution of the time before decision be exponential with rates λ_1, λ_2. The decision of an individual may affect the probability that the other opts in, modifying the rate of time to adoption to $\lambda_1^{+_2}$ or $\lambda_2^{+_1}$ from that point on. Here, the subscript on the $+$ indicates whose move triggered a change in rate. Since there are only two people in this example, the subscript can and will be dropped for the remaining part of the subsection. The resulting change in the distribution of outcomes can be interpreted as a causal peer effect.

The probabilities for the four possible outcomes follow from algebra:

$$
\begin{aligned}
p_{00} &\stackrel{\text{def}}{=} \mathbb{P}[y_1 = y_2 = 0] = e^{-(\lambda_1+\lambda_2)S} \\
p_{10} &\stackrel{\text{def}}{=} \mathbb{P}[y_1 = 1, y_2 = 0] = \lambda_1 e^{-\lambda_2^+ S} g(\lambda_1 + \lambda_2 - \lambda_2^+) \\
p_{01} &\stackrel{\text{def}}{=} \mathbb{P}[y_1 = 0, y_2 = 1] = \lambda_2 e^{-\lambda_1^+ S} g(\lambda_1 + \lambda_2 - \lambda_1^+) \\
p_{11} &\stackrel{\text{def}}{=} \mathbb{P}[y_1 = y_2 = 1] = 1 - p_{00} - p_{10} - p_{01}
\end{aligned}
$$

where $g(\lambda) = \frac{1-e^{-\lambda S}}{\lambda}$ if $\lambda \neq 0$ and $g(0) = S$.

If $\lambda_1 = \lambda_2$[5], the family of rates can be obtained given the probabilities. In this case, $\lambda_1 = \lambda_2 = -\frac{\ln(p_{00})}{2S}$ and λ_2^+ can be recovered from $\frac{\lambda_1 S(p_{00} - e^{-\lambda_2^+ S})}{\ln(p_{00}) + \lambda_2^+ S} = p_{10}$, where the left-hand side is strictly decreasing. λ_1^+ can be obtained analogously.

Although it is not possible to identify the identity of the first mover – the individual who may have exerted a peer effect on the other individual – when $y_1 = y_2 = 1$, it is

[5]One can generally consider $\lambda_i = f(x_i; \theta)$ for a function f, observed covariates x_i, and finite dimensional identifiable parameter θ. $\lambda_1 = \lambda_2$ then corresponds to the intercept case with $x_1 = x_2$. This is the condition that would be imposed to identify rates from independent exponential draws.

possible to (i) estimate the peer effect strength and (ii) determine the probability of an individual moving first whenever the λ's are identified.

Note that in the absence of peer effects – $\lambda_i^+ = \lambda_i$ for $i = 1$ and $i = 2$ – the probabilities reduce to a standard exponential race with independent draws. The change in the exponential rates is thus a measure of dependence and social interaction. Peer effects can thus be described by the relative change in rate λ_i^+/λ_i.

Moving beyond this simple example, one might attempt to extend the analysis to arbitrary group sizes and to obtain an estimator of the family of rates. In the following subsections, I formalize the underlying stochastic process and show that, given data on $y_i, i = 1, \cdots, n$ the probability of the sample $P(y_1, \ldots, y_n|\{\lambda_i^{+_{k_1},\ldots,+_{k_m}})\})$ has a closed-form solution. This suggests a maximum of likelihood estimator whose asymptotic properties are investigated in Section 3. Consistency of the family of λ's follows upon necessary restrictions for identification.

3.2 General framework and definition of peer effects

Individual adoption is modeled with the following continuous time stochastic process: Let T_i^1 be an Exponential distribution with rate λ_i for all i. When a first individual, say k, adopts ($T_k^1 = \min(T_i^1)$), the rates for adoption evolve to $\lambda_i^{+_k}$ for all of k's neighbors. The process then repeats with the altered rates, and so on. The time to adoption is thus $T_i = \sum_{t=1}^{\tau_i} T_{k_t}^t$, where k_t is the index of the 'winner' at stage t, and we observe $y_i = \mathbb{1}_{T_i \leq S}$ at time S. S is a timeline during which each individual had the opportunity to adopt. S can be the time up to a deadline, a school year, or generally the time elapsed from availability/eligibility to observation by the researcher.

This process is quite general in terms of the dynamics it allows, though it imposes some time structure and distributional assumptions. I focus on Exponential distributions for two reasons. First, exponential waiting times arise automatically under the assumption of a constant probability per unit of time, a natural point of departure[6]. Second, this process offers computational niceties because of the properties of the ex-

[6]For the analysis of peer effect, the continuous time process is about separating people and assigning probabilities to different orderings. The exponentials imply that the probability of i moving next at any step is given its current rate divided by the sum over other rates. The functional form has little impact beyond this fact and will not be important for the identification or estimation of peer effects. When the time analysis is of special interest (*e.g.*, diffusion), the functional form may be more relevant and some applications may warrant different specifications.

ponential. When k adopts, the clock should be restarted for all of his or her friends, but not for anyone else. However, due to the memorylessness property, the clock may be re-started for everyone, with the rate of non-friend left unchanged. This can make both theoretical analysis and simulations simpler.

The model is viewed as causal with two main departures from a classical treatment effect framework. These stem from the nature of the problem: one's outcome induces another's treatment and this direction is often unobserved because information about the order of adoptions is typically limited, and there are multiple possible treatments that correspond to all possible dynamic selection of peers.

Peer effect are defined at the individual level through the changes in λ_i induced by neighboring adoptions. The model posits the existence (ex post) of a direction of causality, though the realized direction may not be observed by the researcher. This puts peer effects on stronger theoretical grounds as it defines a structural peer effect parameter and avoids ill-defined peer groups, which tend to be subject to the reflection problem (Manski, 1993).

3.3 Likelihood

Assume without loss that individuals who opted in are $1, \ldots, G$. Since the sequence of arrivals is unknown, the probability of the sample corresponds to

$$\begin{aligned}
&\mathbb{P}[y_1, \ldots, y_G \leq S; y_{G+1}, \ldots, y_N > S] \\
&= \sum_{p \in \mathcal{P}} \mathbb{P}[y_1, \ldots, y_G \leq S; y_{G+1}, \cdots, y_N > S; T_{p_1} < \ldots < T_{p_G}]
\end{aligned}$$

where the sum is over all permutations (with generic element $p = (p_1, \ldots, p_G)$) of the G first arrivals.

Consider the representative term in the sum with individual i being the i-th to

adopt:

$$\int_0^S \int_{t_1^1}^{\infty} \cdots \int_{t_1^1}^{\infty} \prod_{i=1}^{N} \lambda_i e^{-\lambda_i t_i^1} \quad \int_0^{S-t_1^1} \cdots \int_{t_2^2}^{\infty} \prod_{i=2}^{N} \lambda_1^{+_1} e^{-\lambda_i^{+_1} t_i^2} \cdots$$
$$\int_0^{S-\sum_{g=1}^{G-1} t_g^g} \cdots \int_{t_G^G}^{\infty} \prod_{i=G}^{N} \lambda_i^{+_1 \cdots +_{G-1}} e^{-\lambda_i^{+_1 \cdots +_{G-1}} t_i^G}$$
$$\int_{S-\sum_{g=1}^{G} t_g^g}^{\infty} \cdots \int_{S-\sum_{g=1}^{G} t_g^g}^{\infty} \prod_{i=G+1}^{N} \lambda_i^{+_1 \cdots +_G} e^{-\lambda_i^{+_1 \cdots +_G} t_i^{G+1}}$$
$$dt_1^1 \cdots dt_N^1 dt_2^2 \cdots dt_N^2 \cdots dt_G^G \cdots dt_N^G \cdots dt_1^{G+1} \cdots dt_N^{G+1}$$

which, after some algebra, reduces to

$$e^{-\sum_{i=G+1}^{N} \lambda_i^{+_1,\cdots,+_G} S} \left(\prod_{i=1}^{G} \lambda_i^{+_1,\cdots,+_{i-1}} \right) I_{\{c_i, i=1,\cdots,G\}} \tag{3}$$

with $I_{\{h_i, i=1,\cdots,H\}} \stackrel{\text{def}}{=} \frac{1}{\prod_{g=1}^{G} c_g} + (-1)^G \sum_{g=1}^{G} \frac{1}{\prod_{h \neq g} (c_g - c_h)} \frac{e^{-c_g S}}{c_g}$ and $c_g \stackrel{\text{def}}{=} \sum_{i=g}^{N} \lambda_i^{+_1,\cdots,+_{g-1}} - \sum_{i=G+1}^{N} \lambda_i^{+_1,\cdots,+_G}$.

Introducing $c_{G+1} = 0$ and then $\dot{c}_g \stackrel{\text{def}}{=} c_g + \sum_{i=G+1}^{N} \lambda_i^{+_1,\cdots,+_G} = \sum_{i=g}^{N} \lambda_i^{+_1,\cdots,+_{g-1}} \geq 0$ (with equality only if $g = G+1$ and $N = G$), the likelihood simplifies to[7]

$$\prod_{i=1}^{G} \lambda_i^{+_1,\cdots,+_{i-1}} \sum_{g=1}^{G+1} \frac{e^{-\dot{c}_g S}}{\prod_{h \neq g} \dot{c}_h - \dot{c}_g} \tag{4}$$

Remark Suppose there is no heterogeneity nor peer effects: $\lambda_i^{\{+\}} = \lambda$ for all i and collection of $+$. Then $\dot{c}_g = \lambda(N-(g-1))$ and the likelihood of permutation p becomes

[7] As in the 2 people example, it can be checked from the integral computation that the limit for the denominator converging to 0 gives the correct probability at the points where this function is undefined.

$$\prod_{i=1}^{G} \lambda_i^{+_1,\cdots,+_{i-1}} \sum_{g=1}^{G+1} \frac{e^{-\dot{c}_g S}}{\prod_{h\neq g} \dot{c}_h - \dot{c}_g} = \lambda^G \sum_{g=1}^{G+1} \frac{e^{-\lambda(N-(g-1))S}}{\prod_{h\neq g} \lambda(N-(h-1)) - \lambda(N-(g-1))}$$
$$= \sum_{g=1}^{G+1} \frac{e^{\lambda(g-1)S}}{\prod_{h\neq g} h - g}$$
$$= \sum_{g=1}^{G+1} \frac{e^{\lambda(g-1)S}}{(g-1)!\,(G+1-g)!}(-1)^{G+1-g}$$
$$= \frac{e^{-\lambda N S}(e^{\lambda S}-1)^G}{G!}$$
$$= \frac{e^{-\lambda(N-G)S}(1-e^{-\lambda S})^G}{G!}$$

Summing over all permutations yields $e^{-\lambda(N-G)S}(1-e^{-\lambda S})^G$, which is the likelihood from iid exponential random variables.

In its general form, the model handles considerable heterogeneity, albeit with too many parameters to be identified as such. One may restrict the heterogeneity in various ways, depending on empirical concerns and the goal of the analysis (for instance, whether heterogeneity in initial rates, in peer effects, in the identity of first-mover, etc. are more relevant).

A leading case of interest naturally arises from the desire to define a simple peer effect parameter while accounting for individual heterogeneity through observed covariates x_i. A way to make the model parsimonious and to incorporate covariates is to specify $\lambda_i^{+_{k_1}\cdots+_{k_L}} = g(x_i'\beta + N_i(k_1,\ldots,k_L)/d_i\ \delta)$ for some link function g, where $N_i(k_1,\ldots,k_L)$ is the number of neighbors for i among $k_1,\ldots,k_L$ and $d_i = N_i(1,\ldots,N) = \sum_j W(i,j)$ is the degree of node i. The exponential is a natural link function that ensures positivity of the rates.

Although summing over all permutations can lead to an impractical computational burden, the cost is significantly smaller in practice. First, the likelihood factorizes based on the components of W, whose size can be much small (classrooms, villages, connected component of friends, etc.). Second, the permutations depend on the number of adopters, not the size of the network. This means that if the share of adopters is low or there is some extra information – such as partial knowledge of the order of adoptions – the complexity of permutations can be substantially reduced.

Finally, approximations can further cut the number of permutations. In particular, maximizing the log-likelihood amounts to maximizing

$$\ln\left(\frac{1}{G!}\sum_{p\in\mathcal{P}}\left(\prod_{i=1}^{G}\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}\right)\sum_{g=1}^{G+1}\frac{e^{-\dot{c}_{p_g}S}}{\prod_{h\neq g}\dot{c}_{p_h}-\dot{c}_{p_g}}\right) \tag{5}$$

where the average over all permutations can be estimated by a random sample[8] of permutations by the law of large numbers.

Note finally that when individuals are broadly similar or when the network exhibits symmetry, it may be possible to group terms or simplify intermediate computations.

4 Asymptotic Theory

4.1 Identification

Consider a sample of (binary) variable y_i, covariates x_i, and an adjacency matrix W ($W_{ij} = 1$ if and only if i and j are "neighbors") that determines individuals' peer group.

Discussions about identification involves two main components. First, some functional form restriction has to be imposed as to limit the number of parameters, which scales with powers of N in the absence of restrictions (there are already N parameters λ_i, then all modified rates have to be considered).

A natural way to obtain a parsimonious model is to specify $\lambda_i = g(W, x_i, \theta)$ for a vector of covariates x_i and finite-dimensional vector of parameters θ. The modified rate can then be obtained upon adjusting with some scaling factor. For instance, a simple specification is

$$\lambda_i^{+_{k_1}\cdots+_{k_m}} = e^{x_i'\beta+{d_i}^{-1}\sum_{j=1}^{m}W(i,k_j)\delta} \tag{6}$$

where $d_i = \sum_j W(i,j)$ is the degree of i. This reduces the dimensionality of the

[8]Instead of sampling uniformly over the set of permutations, one can also generate sequences of arrival according to the stochastic process using a given family of rates. For a sample $\mathcal{P}_s$ of such draws, one can recover the likelihood from $\frac{1}{|\mathcal{P}_s|}\sum_{p\in\mathcal{P}_s}1/\mathbb{P}[p]\to^p\frac{G!}{\sum_{p\in\mathcal{P}}\mathbb{P}[p]}$ using that $\frac{\mathbb{P}[p]}{\sum_{p\in\mathcal{P}}\mathbb{P}[p]}$ is the probability of sampling permutation p under this sampling scheme. The score can be obtained similarly.

problem to $(\dim(x_i)+1)$ and lets the peer effects be described by a single parameter δ, which reflects the scaling of the rates induced by the peer group opting-in.

To the first-order in δ, we also have

$$\mathbb{E}[Y_i(\vec{0}) - Y_i(\vec{\infty})|X_i] \approx \delta S e^{x_i'\beta} e^{-e^{x_i'\beta}S} \tag{7}$$

so that δ is a scaling factor in the change in the probability of adoption induced by the peer group adopting at the onset.

The next concern is to recover enough observation to ensure the parameters are identified. Different information types can be considered; the following are sufficient for identification:

- Network with a block structure and a large number of blocks
- Sparse network and information about adoption times of first and second-movers

In the first case, identification of the parameters can be deduced from the probabilities for a block. For instance, β is identified from $\mathbb{P}[Y_b = 0|W_b, x_i]$ (as a function of the observed x_i) and then δ is identified from $\mathbb{P}[Y_b = e_1|W_b]$[9], where b is a block.

In the latter case, a sufficiently sparse network ensures that the number of unconnected first-movers grow. This identifies β directly under knowledge of their adoption times as in a standard exponential sampling without dependence. Then, the second-movers deliver the necessary information about δ.

The average degree of an individual is often viewed as constant or increasing only slowly as $n \to \infty$. Thus, sparsity is often a plausible restriction as one then models the probability of forming a link to be of order about n^{-1}.

Remark: Identification of the parameters ensures that the family of rates are identified. A consequence is that the probability of any sequence of adoption is also identified. Hence, while it is not possible to recover the identity of the first-movers, it is possible to assign a probability to that event.

[9]For simplicity, I'll assume that there are no isolated individuals. In general, δ can be identified as long as some peer effects actually take place.

4.2 Asymptotic properties

I consider a sequence of networks with blocks or components of size N_b, $b = 1, \cdots, B$. With independent blocks, the likelihood factorizes as $l \stackrel{\text{def}}{=} \ln(\mathbb{P}[Y = y]) = \sum_{b=1}^{B} l_b$, where l_b is the log-likelihood of block b. The log-likelihood of each block is given by the formula established in the previous section. The score and Hessian are easily computed and are available in closed form. The details are provided in the Appendix.

The estimator inherits the usual properties of maximum likelihood estimators: it is consistent and asymptotically normal under regularity conditions. Formally,

Theorem 4.1 (Consistency)**.** *Suppose the data generating process is given by the stochastic process described in Section 3.2 and that the family of rates is identified and belongs to the interior of a compact set. Then, if blocks are drawn independently with $W_b \sim \mathcal{D}_W$ such that $N_b < \overline{N}$, the maximum likelihood estimator is consistent with $\hat{\lambda}_i^* \to^p \lambda_i^*$, where $*$ represents any collection of $+$, as $B \to \infty$.*

Moreover,

Theorem 4.2 (Asymptotic Normality)**.** *Under the assumptions of Theorem 3.1, the maximum likelihood estimator is asymptotically normal as $B \to \infty$ with*

$$\sqrt{B}(\hat{\lambda} - \lambda) \to^d \mathcal{N}(0, \mathbb{E}[\partial l (\partial l)']) \tag{8}$$

These theorems allow for standard inference about the rates.

Remark: Spatial dependence Spatial dependence is determined by the value of δ in addition to the network structure. In particular, observations are independent in the absence of peer effects ($\delta = 0$). This suggests that blocks with weak ties (relative to δ for given structure and initial rates) could sometimes be treated as approximately independent.

Remark: Discrete outcomes One can generalize the process to deal with discrete outcomes. For instance, one could have $\lambda_i = \sum_k \lambda_{ik}$ and if $T_i = t$ then choice k occurs with probability λ_{ik}/λ_i. 0 is then a baseline opt-out choice, the default.

Remark: Continuous outcomes Continuous random variables can be handled by adding an optimization over unobserved binary adoption status. For instance, let

Y be the binary adoption status and y be the final (continuous) outcome given by $y = Z + \mathbb{1}_{Y \leq S}\mu$, where for instance $Z|X \sim \mathcal{N}(X'\gamma, \sigma^2)$ For any candidate vector $Y \in \{0,1\}^N$, this leads to a sum over all possible adoption patterns and the resulting likelihood can be optimized over to obtain a maximum of likelihood estimator.

Remark: Counterfactual analysis It is often of interest to assess the impact of policies, for instance targeting influential individuals in a network. The peer effect parameter allows a direct evaluation of the direct effect of imposing $y_i = 1$ at a given point. It is also easy to simulate the evolution of a network with and without imposing $y_i = 1$ for a group of selected individuals at time 0 and to assess the change in probabilities that $y_k = 1$ for $k \neq i$.

Remark: Exogenous peer effects One may add a "WX term" to the argument of the exponential to account for so-called exogenous peer effects, which affect one's outcome through the characteristics of peers.

5 Simulations

I now assess the performance of the estimator in simulations. I first consider correctly specified models in which all relevant covariates are available. In the second subsection, I investigate the robustness of the estimator to omitted variables, measurement errors, and group heterogeneities.

5.1 Simulations with block and homophilic network formation

I simulate the stochastic process described in Section 2.2 with an underlying network structure of either 'classrooms' or homophilic matching type, both of which are common in empirical studies.

First, I construct a network with 1000 individuals and a 'block' structure ($(W = I \otimes \iota\iota')$) with groups of size 5, 10, and 20. In the previous section's notation, this means $N = 1000$, $n_b = 5\ \forall b$, and $B = 1000/n_b$ and individuals are connected to all individuals within the same block. I set the family of rates to obey $\lambda_i^{+_{k_1}\cdots+_{k_m}} = e^{x_i'\beta + \frac{\sum_{j=1}^m W(i,k_j)}{d_i}\delta}$ with two covariates (uniformly on $[-1;1]$ and (standard) normally

distributed, respectively), various levels of peer effect strength δ (-0.5, 0, and 0.5), and $\beta = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$.

There is no information about the order of adoptions so all permutations *a priori* matter. I make use of the random sampling over permutations mentioned in Section 2 to alleviate the computational burden whenever the number of adopted people in a group exceeds 8.

Estimates are compared to SAR estimates from a simple (endogenous) regression on x_i and $W_i y$ and to the SAR maximum likelihood estimator[10] These are frequent estimates of peer effects which can be hoped to capture whether there is a peer influence, but cannot be expected to be consistent given incorrect specification; they would not capture the true nature of the peer effects, but could detect their sign and presence. Coefficients on covariates, which are also reported, have similarly no direct counterparts in a linear model and are not expected to be consistently estimated by regression or SAR MLE.

The results are reported in Table 1. The maximum of likelihood estimator described in the previous section performs well in all instances and exhibits very low bias. A standard regression is usually able to pick up the correct sign of peer effects in this specific setup, but cannot recover the structural coefficient. The SAR MLE broadly follows the same lines.

[10]The weighting matrix is row-normalized since the process suggests peer effects depend on the average number of adopted friends. Notice, however, that the model using sums has an equivalent representation using averages when groups have the same size: it amounts to scaling δ by group size.

Table 1: Simulations with 'classrooms' network structure

			Bias			Standard deviation			RMSE		
n_b	δ		Reg	SAR	Exp	Reg	SAR	Exp	Reg	SAR	Exp
5	-0.5	$\hat{\delta}$	0.39	0.43	0.03	0.04	0.03	0.12	0.39	0.43	0.13
		$\hat{\beta}_1$	−0.70	−0.37	0.00	0.02	0.03	0.09	0.70	0.38	0.09
		$\hat{\beta}_2$	−0.36	−0.19	0.00	0.01	0.02	0.05	0.36	0.20	0.05
	0	$\hat{\delta}$	−0.02	−0.01	0.00	0.08	0.04	0.10	0.08	0.04	0.10
		$\hat{\beta}_1$	−0.70	−0.37	0.00	0.02	0.03	0.09	0.70	0.37	0.09
		$\hat{\beta}_2$	−0.36	−0.20	0.00	0.01	0.02	0.05	0.36	0.20	0.05
	0.5	$\hat{\delta}$	−0.35	−0.41	−0.01	0.07	0.04	0.09	0.36	0.42	0.09
		$\hat{\beta}_1$	−0.71	−0.37	0.00	0.02	0.03	0.09	0.71	0.37	0.09
		$\hat{\beta}_2$	−0.36	−0.21	0.00	0.01	0.02	0.05	0.36	0.21	0.05
10	-0.5	$\hat{\delta}$	0.32	0.41	0.00	0.14	0.07	0.13	0.35	0.41	0.13
		$\hat{\beta}_1$	−0.69	−0.37	0.00	0.02	0.04	0.09	0.69	0.38	0.09
		$\hat{\beta}_2$	−0.35	−0.19	0.00	0.01	0.02	0.05	0.35	0.20	0.05
	0	$\hat{\delta}$	−0.01	−0.01	0.00	0.12	0.07	0.11	0.12	0.07	0.11
		$\hat{\beta}_1$	−0.70	−0.37	0.00	0.02	0.04	0.09	0.70	0.37	0.09
		$\hat{\beta}_2$	−0.36	−0.20	0.00	0.01	0.02	0.05	0.36	0.20	0.05
	0.5	$\hat{\delta}$	−0.37	−0.42	0.00	0.10	0.06	0.10	0.38	0.43	0.10
		$\hat{\beta}_1$	−0.71	−0.37	0.01	0.02	0.04	0.08	0.71	0.37	0.08
		$\hat{\beta}_2$	−0.36	−0.21	0.01	0.01	0.02	0.06	0.36	0.21	0.06
20	-0.5	$\hat{\delta}$	0.30	0.40	0.00	0.21	0.10	0.13	0.36	0.41	0.13
		$\hat{\beta}_1$	−0.70	−0.37	0.01	0.02	0.06	0.08	0.70	0.37	0.09
		$\hat{\beta}_2$	−0.36	−0.20	0.01	0.01	0.02	0.05	0.36	0.20	0.05
	0	$\hat{\delta}$	−0.06	−0.03	−0.01	0.18	0.10	0.11	0.19	0.10	0.11
		$\hat{\beta}_1$	−0.70	−0.36	0.00	0.02	0.06	0.08	0.70	0.36	0.08
		$\hat{\beta}_2$	−0.36	−0.20	0.00	0.01	0.02	0.05	0.36	0.20	0.05
	0.5	$\hat{\delta}$	−0.38	−0.42	−0.02	0.15	0.09	0.10	0.40	0.43	0.11
		$\hat{\beta}_1$	−0.71	−0.36	0.07	0.02	0.06	0.18	0.71	0.37	0.19
		$\hat{\beta}_2$	−0.36	−0.21	0.03	0.01	0.02	0.08	0.36	0.21	0.09

I now consider another network structure, in which individuals within groups decide whether to make a connection based on their characteristics. Specifically, I consider a homophilic link formation process in which individual match according to their similarities: $W_{ij} = 1$ iff $\frac{\|X_{1i}-X_{1j}\|+\|X_{2i}-X_{2j}\|}{2} < \eta_{ij}$), where the collection of

$\eta_{ij} = \eta_{ji}$ forms an array of independent uniform random variables. Group sizes are $5, 20$, or a larger group of 100 and the sample size is gain $N = 1000$.

The results are displayed in the next table.

Table 2: Homophilic

			Bias			Standard deviation			RMSE		
n_b	δ		Reg	SAR	Exp	Reg	SAR	Exp	Reg	SAR	Exp
5	-0.5	$\hat{\delta}$	0.39	0.42	0.03	0.04	0.03	0.13	0.39	0.43	0.13
		$\hat{\beta}_1$	−0.70	−0.38	−0.01	0.02	0.02	0.09	0.70	0.38	0.09
		$\hat{\beta}_2$	−0.36	−0.20	0.00	0.01	0.02	0.05	0.36	0.20	0.05
	0	$\hat{\delta}$	0.00	0.00	−0.02	0.04	0.03	0.11	0.04	0.03	0.11
		$\hat{\beta}_1$	−0.69	−0.38	0.02	0.02	0.02	0.09	0.69	0.38	0.09
		$\hat{\beta}_2$	−0.36	−0.19	0.01	0.01	0.02	0.05	0.36	0.20	0.05
	0.5	$\hat{\delta}$	−0.38	−0.41	−0.03	0.03	0.02	0.10	0.38	0.41	0.11
		$\hat{\beta}_1$	−0.71	−0.38	0.02	0.02	0.02	0.08	0.71	0.38	0.08
		$\hat{\beta}_2$	−0.36	−0.20	0.00	0.01	0.02	0.05	0.36	0.21	0.05
10	-0.5	$\hat{\delta}$	0.41	0.44	0.03	0.05	0.03	0.12	0.41	0.44	0.12
		$\hat{\beta}_1$	−0.69	−0.39	0.00	0.02	0.02	0.09	0.70	0.39	0.09
		$\hat{\beta}_2$	−0.35	−0.20	0.00	0.01	0.02	0.05	0.35	0.20	0.05
	0	$\hat{\delta}$	0.00	0.00	−0.02	0.04	0.03	0.10	0.04	0.03	0.10
		$\hat{\beta}_1$	−0.70	−0.38	0.00	0.02	0.02	0.08	0.70	0.38	0.08
		$\hat{\beta}_2$	−0.36	−0.20	0.00	0.01	0.02	0.05	0.36	0.20	0.05
	0.5	$\hat{\delta}$	−0.39	−0.42	−0.05	0.05	0.03	0.10	0.39	0.43	0.11
		$\hat{\beta}_1$	−0.71	−0.37	0.01	0.02	0.02	0.08	0.71	0.37	0.08
		$\hat{\beta}_2$	−0.36	−0.21	0.00	0.01	0.02	0.05	0.36	0.21	0.05

Although the performance of OLS or SAR-MLE in terms of bias and RMSE in the absence of peer effects ($\delta = 0$) suggests that these estimators may successfully detect the absence of social influence, notice that estimates are generally attenuated compared to the structural parameter and that decisions will eventually be based on tests or confidence intervals. As a result, the coverage performance of the confidence intervals may be a more relevant benchmark and will be analyzed in the next subsection.

Interestingly, OLS and SAR-MLE feature attenuation bias with respect to the structural parameter. As a result, they may seem to perform better in terms of

RMSE in the absence of peer effect. In practice, however, what matters is the test for the presence of peer effect or, equivalently, the resulting confidence intervals. In the next subsection, I explore the coverage performance of the three estimators to assess their ability to (correctly) not reject a null hypothesis of no peer effects in both correctly specified and misspecified models.

5.2 Mispecifications type of results

Peer effect studies are often subject to criticism due to modeling (Manski, 1993; Angrist, 2014) and empirical (Angrist, 2014) concerns. While it is hoped that the framework developed in this paper alleviates modeling concerns - in particular, by avoiding reflection problems -, it is of interest to evaluate the behavior of the estimator under frequent empirical difficulties: missing or omitted covariates, group level heterogeneity, or measurement error.

I focus here on the peer effect parameter δ, which will typically the parameter of interest.

Because OLS and SAR cannot identify the structural coefficient but could still detect the presence of peer effects, it is of interest to look at the coverage performance. I look at the frequency at which a 95% confidence interval contains 0, indicating the absence of peer effects, under the generating process in which peer effects are indeed absent ($\delta = 0$).

Table 3 reports the coverage of a 95% confidence interval under the homophilic network structure when the researcher (i) observes both covariates, (ii) observes only the first covariate, (iii) observes a mismeasured (with $(\mathcal{N}(0; 0.25))$ error) first covariate, and (iv)/(v)/(vi) there is (uniform on $[-1; 0]$) group heterogeneity (added to the argument of the exponential) in the (i)/(ii)/(iii) scenario.

Table 3: Coverage analysis with potential misspecification

			Coverage		
n_b	δ		Reg	SAR	Exp
5	0	Size	0.90	0.79	0.95
		Size	0.72	0.62	0.90
		Size	0.69	0.55	0.90
		Size	0.83	0.70	0.89
		Size	0.62	0.47	0.71
		Size	0.56	0.42	0.67

Table 4: Coverage performance of a 95% confidence interval from OLS with clustered standard errors, SAR-MLE, and maximum of likelihood on latent exponential processes.

The coverage performance of the estimator developed in the paper is far better

than that of OLS and SAR-MLE. Although the most serious issues (lack of covariate and measurement error combined with heterogeneity issues) can lead to severe size distortions, spurious peer effects are unlikely under more standard scenarios. The test for the presence of peer effect is adequately sized in the case of correct specification and is moderately distorted under measurement error or group heterogeneity.

Both OLS and SAR-MLE have a tendency to spuriously detect peer effects at a rate higher than the pre-specified level, even with homogeneous groups and adequate covariates. Any empirical difficulty such as measurement error, unobserved covariate, or heterogeneity leads by itself to a high risk of unwarranted rejection of the null of no peer effects, echoing critiques in Angrist (2014).

References

Angrist, Joshua D. 2014. "The perils of peer effects." *Labour Economics*, 30: 98–108.

Aronow, Peter M, and Cyrus Samii. 2017. "Estimating average causal effects under general interference, with application to a social network experiment." *The Annals of Applied Statistics*, 11(4): 1912–1947.

Arpino, Bruno, Luca De Benedictis, and Alessandra Mattei. 2017. "Implementing propensity score matching with network data: the effect of the General Agreement on Tariffs and Trade on bilateral trade." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3): 537–554.

Athey, Susan, and Guido W Imbens. 2022. "Design-based analysis in difference-in-differences settings with staggered adoption." *Journal of Econometrics*, 226(1): 62–79.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of peer effects through social networks." *Journal of econometrics*, 150(1): 41–55.

De Giorgi, Giacomo, Michele Pellizzari, and Silvia Redaelli. 2010. "Identification of social interactions through partially overlapping peer groups." *American Economic Journal: Applied Economics*, 2(2): 241–75.

Egami, Naoki, and Eric J Tchetgen Tchetgen. 2024. "Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2): 487–511.

Forastiere, Laura, Edoardo M Airoldi, and Fabrizia Mealli. 2020. "Identification and estimation of treatment and interference effects in observational studies on networks." *Journal of the American Statistical Association*, 1–18.

He, Xiaoqi, and Kyungchul Song. 2018. "Measuring diffusion over a large network." *arXiv preprint arXiv:1812.04195*.

Huber, Martin, and Andreas Steinmayr. 2021. "A framework for separating individual-level treatment effects from spillover effects." *Journal of Business & Economic Statistics*, 39(2): 422–436.

Jackson, Matthew O, Zhongjian Lin, and Ning Neil Yu. 2020. "Adjusting for peer-influence in propensity scoring when estimating treatment effects." *Available at SSRN 3522256*.

Lee, Lung-Fei. 2004. "Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models." *Econometrica*, 72(6): 1899–1925.

Liu, Lan, Michael G Hudgens, Bradley Saul, John D Clemens, Mohammad Ali, and Michael E Emch. 2019. "Doubly robust estimation in observational studies with partial interference." *Stat*, 8(1): e214.

Manski, Charles F. 1993. "Identification of endogenous social effects: The reflection problem." *The review of economic studies*, 60(3): 531–542.

Newey, Whitney K, and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of econometrics*, 4: 2111–2245.

Neyman, Jerzy. 1923. "On the application of probability theory to agricultural experiments. Essay on principles." *Ann. Agricultural Sciences*, 1–51.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66(5): 688.

Sánchez-Becerra, Alejandro. 2021. "Spillovers, Homophily, and Selection into Treatment: The Network Propensity Score." *Job market paper*.

Shaikh, Azeem M, and Panos Toulis. 2021. "Randomization tests in observational studies with staggered adoption of treatment." *Journal of the American Statistical Association*, 116(536): 1835–1848.

Shalizi, Cosma Rohilla, and Andrew C Thomas. 2011. "Homophily and contagion are generically confounded in observational social network studies." *Sociological methods & research*, 40(2): 211–239.

Sofrygin, Oleg, and Mark J van der Laan. 2016. "Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population." *Journal of causal inference*, 5(1).

Toulis, Panos, and Edward Kao. 2013. "Estimation of causal peer influence effects." 1489–1497, PMLR.

Appendix A: Proofs and Further results

5.3 Derivatives of the likelihood

The log-likelihood reads

$$l \stackrel{\text{def}}{=} \ln\left(\frac{1}{G!}\sum_{p\in\mathcal{P}}\sum_{g=1}^{G+1}\left(\prod_{i=1}^{G}\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}\right)\frac{e^{-\dot{c}_{p_g}S}}{\prod_{h\neq g}\dot{c}_{p_h}-\ddot{c}_{p_g}}\right) \tag{9}$$

Let

$$a_{pg} \stackrel{\text{def}}{=} \left(\prod_{i=1}^{G}\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}\right)\frac{e^{-\dot{c}_{p_g}S}}{\prod_{h\neq g}\dot{c}_{p_h}-\dot{c}_{p_g}} \tag{10}$$

and $\theta \stackrel{\text{def}}{=} (\beta', \delta)'$. Then, the score is given by

$$\frac{\partial l}{\partial \theta} = \sum_{p\in\mathcal{P}}\sum_{g=1}^{G+1}\frac{a_{pg}}{\sum_p\sum_g a_{pg}}\left(\sum_{i=1}^{G}\frac{\partial_\theta\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}}{\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}} - \partial_\theta\dot{c}_{p_g}S - \sum_{h\neq g}\frac{\partial_\theta(\dot{c}_{p_h}-\dot{c}_{p_g})}{\dot{c}_{p_h}-\dot{c}_{p_g}}\right) \tag{11}$$

and the Hessian is given by

$$H \stackrel{\text{def}}{=} \frac{\partial l}{\partial\theta\partial\theta'} = \sum_{p\in\mathcal{P}}\sum_{g=1}^{G+1}\partial_{\theta'}\left(\frac{a_{pg}}{\sum_p\sum_g a_{pg}}\right)\left(\sum_{i=1}^{G}\frac{\partial_\theta\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}}{\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}} - \partial_\theta\dot{c}_{p_g}S - \sum_{h\neq g}\frac{\partial_\theta(\dot{c}_{p_h}-\dot{c}_{p_g})}{\dot{c}_{p_h}-\dot{c}_{p_g}}\right)$$
$$+\sum_{p\in\mathcal{P}}\sum_{g=1}^{G+1}\frac{a_{pg}}{\sum_p\sum_g a_{pg}}\left(\sum_{i=1}^{G}\partial_{\theta'}\left(\frac{\partial_\theta\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}}{\lambda_{p_i}^{+_{p_1},\cdots,+_{p_{i-1}}}}\right) - \partial_\theta\dot{c}_{p_g}S - \sum_{h\neq g}\partial_{\theta'}\left(\frac{\partial_{\theta\theta'}(\dot{c}_{p_h}-\dot{c}_{p_g})}{\dot{c}_{p_h}-\dot{c}_{p_g}}\right)\right) \tag{12}$$

When $\lambda_i^{+_{k_1}\cdots+_{k_m}} = e^{x_i'\beta+{d_i}^{-1}\sum_{j=1}^m W(i,k_j)\delta}$, the derivatives read

$$\partial_\theta\lambda_i^{+_{k_1}\cdots+_{k_m}} = \lambda_i^{+_{k_1}\cdots+_{k_m}}\begin{pmatrix} x_i \\ {d_i}^{-1}\sum_{j=1}^m W(i,k_j)\end{pmatrix} \tag{13}$$

and

$$\partial_{\theta\theta'}\lambda_i^{+_{k_1}\cdots+_{k_m}} = \lambda_i^{+_{k_1}\cdots+_{k_m}}\begin{pmatrix} x_ix_i' & x_i{d_i}^{-1}\sum_{j=1}^m W(i,k_j) \\ x_i{d_i}^{-1}\sum_{j=1}^m W(i,k_j) & ({d_i}^{-1}\sum_{j=1}^m W(i,k_j))^2\end{pmatrix} \tag{14}$$

5.4 Proof of theorems

The proofs verify Newey and McFadden (1994)'s sufficient conditions for consistency and asymptotic normality of extremum estimators.

The parameter space is compact by assumption. The log-likelihood of block b is $l_b = \sum_{y\in 0,1^{N_b}}\mathbb{1}[Y_b = y]\ln(\mathbb{P}[Y_b = y])$, where the elements of the sum are bounded from below using that $N_b \leq \overline{N}$ and compactness of the rates and from above by 0. Since in addition $\mathbb{P}[Y_b = y]$ is continuous, uniform laws of large numbers implies (i) $\sup_{\{\lambda\}}|B^{-1}\sum_b l_b - \mathbb{E}[l_b]|\to 0$ and (ii) $\mathbb{E}[l_b]$ is continuous.

By continuity of the Hessian matrix and compactness of the parameter space, it follows that

$$\sup_{\{\lambda\}}|H_n - H|\to 0 \tag{15}$$

where H is continuous. Asymptotic normality of the score follows from laws of large numbers for triangular arrays, noting that $\mathbb{V}[l_b]\to\sigma^2$

By assumption, the parameters are identified and live in a compact set. The objective function converges uniformly to $\mathbb{E}[l_b]$, which is continuous, by uniforms law

of large numbers (*e.g.*, Lemma 2.4 in Newey and McFadden (1994)). Indeed, l_b is continuous and

$$\min_{y,p,\lambda} \mathbb{P}[p(\lambda)] \leq e^{l_b} \leq 1, \tag{16}$$

where the minimization is over a finite set for y, p since $N_b \leq \overline{N}$ and the minimization over λ is over a compact set for a continuous function, so that $\sup|B^{-1}\sum_{b=1}^{B} l_b - \mathbb{E}[l_b]| \to 0$.

$$\frac{1}{\sqrt{B}} \sum_{b=1}^{B} s_b \to^d \mathcal{N}(0; \mathbb{E}[s_b s_b']) \tag{17}$$

noting that the score has mean zero and

$$\begin{aligned}
\mathbb{V}[s_b] &= \mathbb{E}[s_b s_b'] \\
&\leq C\mathbb{E}[\partial L_b \partial L_b'] \\
&= C \sum_{W_b} \mathbb{P}[W_b] \mathbb{E}[\partial L_b \partial L_b' | W_b] \\
&= C \sum_{W_b} \mathbb{P}[W_b] \sum_{\tilde{y}} \mathbb{P}[Y = \tilde{y}] \partial L_b(\tilde{y}) \partial L_b'(\tilde{y}) \\
&\leq C \max_{\tilde{y}} \partial L_b(\tilde{y}) \partial L_b'(\tilde{y})
\end{aligned} \tag{18}$$