

On Evolution-Based Models for Experimentation Under Interference

Sadegh Shirani and Mohsen Bayati
 Graduate School of Business, Stanford University

Causal effect estimation in networked systems is central to data-driven decision making. In such settings, interventions on one unit can spill over to others, and in complex physical or social systems, the interaction pathways driving these interference structures remain largely unobserved. We argue that for identifying population-level causal effects, it is not necessary to recover the exact network structure; instead, it suffices to characterize how those interactions contribute to the evolution of outcomes. Building on this principle, we study an evolution-based approach that investigates how outcomes change across observation rounds in response to interventions, hence compensating for missing network information. Using an exposure-mapping perspective, we give an axiomatic characterization of when the empirical distribution of outcomes follows a low-dimensional recursive equation, and identify minimal structural conditions under which such evolution mappings exist. We frame this as a distributional counterpart to difference-in-differences. Rather than assuming parallel paths for individual units, it exploits parallel evolution patterns across treatment scenarios to estimate counterfactual trajectories. A key insight is that treatment randomization plays a role beyond eliminating latent confounding; it induces an *implicit sampling from hidden interference channels*, enabling consistent learning about heterogeneous spillover effects. We highlight causal message passing as an instantiation of this method in dense networks while extending to more general interference structures, including influencer networks where a small set of units drives most spillovers. Finally, we discuss the limits of this approach, showing that strong temporal trends or endogenous interference can undermine identification.

Key words: Randomized experiments, network interference, exposure mapping, implicit sampling

1. Introduction

Analyzing network data is central to scientific fields, with broad applications spanning operations research, economics, and social sciences (Jackson et al. 2008, Johari et al. 2022). These analyses are inherently challenging due to complex dependencies shaped by latent and observed network structures (Kolaczyk and Csárdi 2014). Estimating causal effects in networked systems presents a particularly challenging case (Hudgens and Halloran 2012, Ugander et al. 2013, Eckles et al. 2016). This complexity arises because, in addition to intricate network dependencies, researchers must face the fact that we can observe outcomes under only a single realized scenario (Holland 1986).

Interference induced by underlying network structures violates the stable unit treatment value assumption (SUTVA), necessitating methods that can disentangle causal relationships in interconnected settings (Imbens and Rubin 2015). Within the growing literature on experimental design under interference, one line of research investigates the use of outcome observations collected over time. Specifically, Li and Wager (2022a) highlight the utility of additional longitudinal data for inference and the complications introduced by temporal interference (Glynn et al. 2020).

Recent contributions have exploited temporal structures and observations to enhance estimation under interference (Farias et al. 2022, Hu and Wager 2022, Bojinov et al. 2023, Li et al. 2023, Ni et al. 2023, Boyarsky et al. 2023, Han et al. 2024, Mukaigawara et al. 2025, Jia et al. 2025). Collectively, these studies demonstrate that temporal dynamics can be utilized effectively to mitigate bias and improve statistical inference in the presence of interference, but they typically retain a unit-level or network-level description of interference.

This temporal approach has also been used in broader network problems to detect connections and structural properties. Systematic perturbations followed by observing the resulting outcome trajectories can provide information about network topology (Timme and Casadiego 2014). For example, observing transient responses to targeted or random perturbations enables the identification of causal relations among network nodes (Casadiego et al. 2017, Nitzan et al. 2017, Stepaniants et al. 2020).

Recently, Shirani and Bayati (2024) introduced causal message passing (CMP) for estimating treatment effects in networked populations. Building on the approximate message passing methodology (Donoho et al. 2009, Bayati and Montanari 2011), CMP introduces the concept of **experimental state evolution (ESE)**. ESE captures how outcome distributions evolve over time, with this temporal evolution characterized through mathematical functions called **ESE mappings**.

The derivation of ESE mappings and their application to treatment effect estimation have been studied under certain structural assumptions on outcome dynamics and interference patterns (Shirani and Bayati 2024, Bayati et al. 2024, Shirani et al. 2025). These papers derive ESE as a consequence of particular unit-level outcome models and random network assumptions. However, it remains unclear when such evolution-based descriptions exist, how they relate to classical exposure mappings for interference (Manski 2013, Aronow and Samii 2017), and under what structural conditions they support identification of population-level counterfactual trajectories without reconstructing the underlying network. In this paper, we reverse the perspective. Instead of starting from a detailed outcome and network model and characterizing limiting distributions, we take ESE itself as the primitive object and ask:

1. Under what assumptions on potential outcomes and exposure mappings does an ESE representation exist at the population level?
2. When such a representation exists, when does it suffice for identification of counterfactual evolutions, and how does it relate to familiar assumptions such as parallel trends?
3. How can partial structural knowledge, such as the presence of a small set of influencers or known clusters, be incorporated without requiring full network reconstruction?
4. In which environments do evolution-based approaches fail, regardless of the estimator?

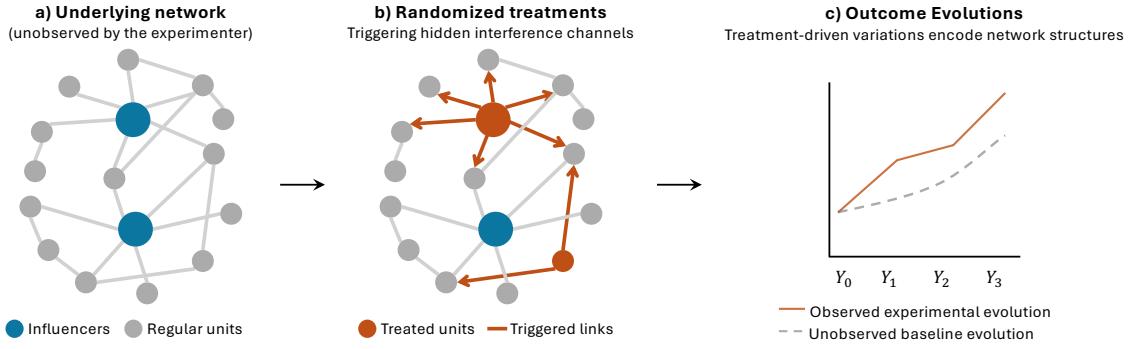


Figure 1 Randomized treatments implicitly sample hidden interference channels in a network with two influencers.

Our contributions are as follows. First, we provide an axiomatic treatment of the ESE framework. Using a potential-outcome evolution model and time-varying exposure mappings, we derive minimal regularity conditions under which population outcome distributions follow an ESE recursion driven by treatment assignments. Second, we show how ESE underpins an evolution-based estimation strategy and clarify its connection to difference-in-differences: under a distributional parallel evolution condition, counterfactual trajectories can be constructed by repeatedly propagating from common pre-treatment baselines, providing a distributional analogue of the parallel-trends assumption. This generalizes the evolution-based viewpoint beyond the dense network regimes considered by CMP and moves beyond the limitations of these methods. As concrete illustrations, we demonstrate how the approach extends to core-periphery settings where a small but known set of “influencers” exert disproportionate spillover effects. Finally, we show intrinsic limitations of evolution-based methods: we show that strong time trends or treatment-dependent exposure mechanisms violate the stability conditions required for ESE, and document empirically how such violations degrade evolution-based estimators.

A key insight is that treatment randomization addresses two separate challenges. The first concerns *unobserved characteristics*: random assignment makes treated and control units comparable on average, thereby removing confounding. The second, newly emphasized here, concerns *unobserved network structure*. Randomization injects structured variation into the hidden interference channels by triggering a representative subset of the links through which spillovers operate. By tracking how outcomes evolve across these perturbations, we effectively sample the latent pathways that transmit treatment effects. This *implicit sampling* mechanism enables identification of population-level causal effects without reconstructing the underlying network (Figure 1).

The remainder of the paper proceeds as follows. Section 2 introduces experimental setup along with an illustrative example that motivates the foundation of our approach. Section 3 establishes a set of minimal assumptions for the existence of ESE mappings. Section 4 analyzes ESE mappings through the lens of exposure mappings. Section 5 explains when outcome evolutions can be used for estimation. Finally, Section 6 discusses limitations and delineates the scope of the framework.

2. Problem Statement

We consider a set of N units indexed by $i = 1, \dots, N$, and suppose that all units can be observed across T rounds indexed by $t = 1, \dots, T$. For simplicity, we focus on binary treatment interventions and denote by W_t^i the treatment applied to unit i in round t . We assume that $W_t^i \sim \text{Bernoulli}(p_t)$ for some $0 \leq p_t \leq 1$, where $W_t^i = 1$ indicates that unit i receives the treatment in round t and $W_t^i = 0$ otherwise. We denote the treatment assignments across all units in round t by $\vec{W}_t := (W_t^1, \dots, W_t^N)^\top$, and define $\mathbf{W}_t := [\vec{W}_1 | \dots | \vec{W}_t]$ as the collective treatment assignment up to round t . We then use w_t^i , \vec{w}_t , and \mathbf{w}_t to denote generic realizations of W_t^i , \vec{W}_t , and \mathbf{W}_t , respectively.

Letting Y_0^i denote the initial outcome of unit i prior to any treatment administration, we follow the Neyman-Rubin causal framework (Rubin 1978, Imbens and Rubin 2015). Precisely, we assume that for any feasible treatment assignment $\mathbf{w}_t \in \{0, 1\}^{N \times t}$, the potential outcome of unit i in round t exists and denote it by $Y_t^i(\mathbf{w}_t) \in \mathbb{R}$. Then, we denote the vector of potential outcomes in round t by $\vec{Y}_t(\mathbf{w}_t) := (Y_t^1(\mathbf{w}_t), \dots, Y_t^N(\mathbf{w}_t))^\top$ and the panel of outcomes by $\mathbf{Y}(\mathbf{w}_T) := [\vec{Y}_0 | \vec{Y}_1(\mathbf{w}_1) | \dots | \vec{Y}_T(\mathbf{w}_T)]$. The experimentation procedure consists of assigning treatments according to \mathbf{W}_T and recording the outcome panel $\mathbf{Y}(\mathbf{W}_T)$. Then, motivated by population-level causal estimands,¹ we focus on aggregated outcomes: *how changes in \mathbf{W}_T impact the distribution of observed outcomes $\mathbf{Y}(\mathbf{W}_T)$* .

We now provide a formal definition of the sequence of ESE mappings f_1, \dots, f_T . For each round t , we define random variables $Y_{t-1}(\mathbf{w}_{t-1})$, $Y_t(\mathbf{w}_t) \in \mathbb{R}$ that follow the same distributions as the empirical distribution of the elements of the outcome vectors $\vec{Y}_{t-1}(\mathbf{w}_{t-1})$, $\vec{Y}_t(\mathbf{w}_t) \in \mathbb{R}^N$, respectively. The ESE mapping f_t then establishes a functional relationship from $Y_{t-1}(\mathbf{w}_{t-1})$ to $Y_t(\mathbf{w}_t)$. This sequential evolution process is illustrated in Figure 2. Indeed, f_t captures the mechanism by which past outcomes $Y_{t-1}(\mathbf{w}_{t-1})$ evolve in response to current treatments \vec{w}_t to generate new outcomes $Y_t(\mathbf{w}_t)$. The following example illustrates how the network structure contributes to this evolution.

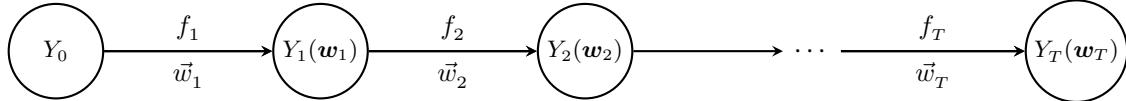


Figure 2 Experimental State Evolution: sequential transformation of outcome distributions via ESE mappings.

An Illustrative Example: Outcomes Evolution Encode Network Structure!

Consider a community of seven individuals, some of whom interact regularly, as shown on the left side of Figure 3. We aim to share new information about the benefits of regular exercise and study its impact on individuals' behavior by measuring each person's daily workout time. In this context, the initial outcome vector \vec{Y}_0 denotes the activity levels of individuals before the experiment begins.

¹ A well-studied example of population-level estimands is the total (or global) treatment effect, which contrasts the sample mean of outcomes under universal treatment with those under no treatment (Yu et al. 2022).

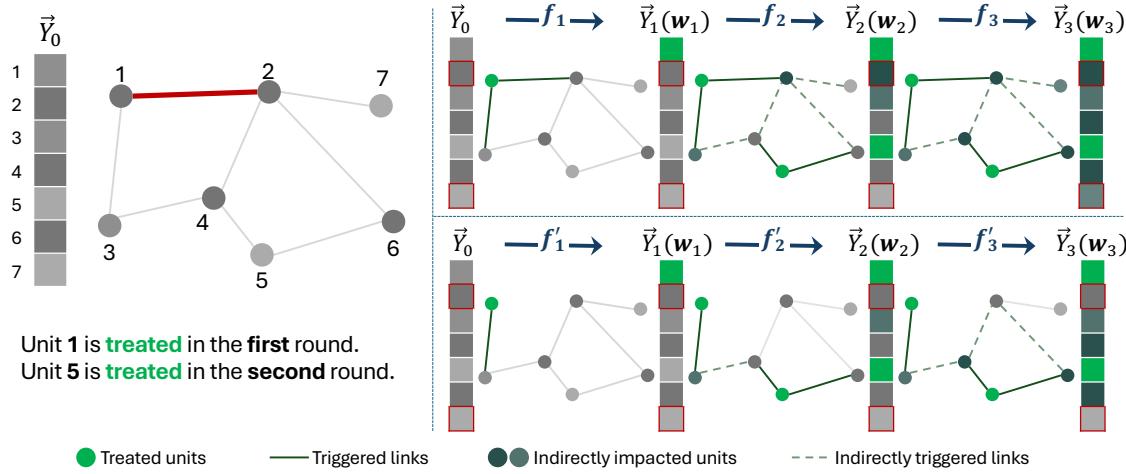


Figure 3 Outcomes evolution: treatments trigger interference channels both directly and indirectly, shaping the next column of outcomes. The columns show outcomes with (top) and without (bottom) a direct link between units 1 and 2. Even without observing the network, the outcome evolutions (for units 2 and 7) reveal differences in interference structures, implying distinct ESE mappings across the two cases.

To conduct the experiment, we share the information with a randomly selected group of individuals. These interventions not only impact the workout time of those treated units but can also alter the content of their regular conversations with their friends. For example, two friends may already speak daily about unrelated topics. Once one of them receives the new information, their conversation may shift to include the benefits of exercise. This shift marks their link as **triggered**: the connection does not change structurally, but its functionality may change relative to the baseline, allowing the treatment-related information to flow through it.

In subsequent days, some of those newly informed friends may pass the information to their own contacts, generating additional rounds of triggered interactions. Although we do not observe these triggered links or the conversations that carry them, we do observe their aggregate effect in how the outcome vector changes from one day to the next. By comparing each column of outcomes with the previous one, we see which parts of the population respond more strongly, revealing the footprint of the underlying interference pathways. This is the key insight: the evolution of outcomes encodes the network effect; in fact, *the treatment-driven variations in outcome columns provide indirect but observable evidence of the unobserved interference structure*.

Figure 3 illustrates this idea by comparing two otherwise identical seven-person networks: one in which Individuals 1 and 2 are friends (top panel) and one in which they are not (bottom panel). In both cases, Individuals 1 and 5 are treated in the first and second rounds, respectively, and outcomes are shown as columns, with color intensity representing workout activity. When 1 and 2 are friends, the influence of Individual 1 spreads through their connection, leading to stronger responses in subsequent rounds. When they are not connected, this channel is absent, and the

spillover pattern weakens (see the outcomes of Individuals 2 and 7). The differences in these evolving outcome patterns reveal the presence or absence of hidden connections.

In real-world experiments, networks are large and heterogeneous, and their structure is rarely observed. Randomization makes it possible to handle this complexity through an **implicit sampling mechanism**. When we randomly choose whom to inform about the benefits of exercise, we also randomly *trigger* the social links connected to those individuals by shifting the content of their interactions. Each treated person activates a distinct local neighborhood, and taken together these activations form a representative sample of the hidden connections in the community. The resulting randomized triggers shape how outcomes evolve from one round to the next, allowing us to infer properties of unobserved interference from observed dynamics. Because it is infeasible to track all such variations one by one, we instead characterize their aggregate through the evolution of outcome distribution. ESE mappings formalize this idea by modeling how outcome distributions change as treatments trigger interaction links, translating temporal patterns in the data into information about the underlying interference structure.

3. Experimental State Evolution

We begin by introducing two new notations. In addition to treatments and outcomes, each unit i also has a covariate vector $\vec{C}_t^i \in \mathbb{R}^{n_c}$ in each round t , for some positive integer n_c . This vector can incorporate unit-specific characteristics, such as age and gender for individuals, as well as time-dependent indicators, including the round number t . These covariates serve to characterize the heterogeneity across units and over time throughout the experimentation period.

Furthermore, to account for potential interference between experimental units, we rely on the concept of exposure mappings (Manski 2013, Aronow and Samii 2017). Specifically, we let $\vec{\mathcal{E}}_t^i(\mathbf{w}_t)$ be the *exposure* vector taking values in \mathbb{R}^{n_e} , for some positive integer n_e . That means, for a treatment allocation \mathbf{w}_t , the elements of $\vec{\mathcal{E}}_t^i(\mathbf{w}_t)$ encompass the effective influence received by unit i in round t from all other experimental units. The following assumption formalizes this.

ASSUMPTION 1 (Exposure vectors). *For any treatment assignment \mathbf{w}_t , there exist exposure vectors $\vec{\mathcal{E}}_t^1(\mathbf{w}_t), \dots, \vec{\mathcal{E}}_t^N(\mathbf{w}_t)$ that summarize the interference effects received by each unit in round t .*

We proceed by characterizing the unit-level outcome evolution. For simplicity, we assume that the outcome of unit i in round t (denoted by $Y_t^i(\mathbf{w}_t)$) depends only on its most recent outcome $Y_{t-1}^i(\mathbf{w}_{t-1})$, the current treatment w_t^i , as well as the covariates \vec{C}_t^i and exposure vectors $\vec{\mathcal{E}}_t^i(\mathbf{w}_t)$. This assumption defines a minimal memory model that facilitates tractable analysis. Our framework extends to more complex dynamics that involve more memory terms or full treatment trajectories, without requiring structural changes to the analytical setup. We formalize our potential outcome specification through the following assumption.

ASSUMPTION 2 (Potential outcome evolution). *We assume that all unit outcomes evolve following a similar mathematical rule and define the generic unit-level evolution through a function φ such that given initial pretreatment outcomes \vec{Y}_0 , we have*

$$\varphi : \{0, 1\} \times \mathbb{R} \times \mathbb{R}^{n_c} \times \mathbb{R}^{n_e} \mapsto \mathbb{R}, \quad Y_t^i(\mathbf{w}_t) := \varphi(w_t^i, Y_{t-1}^i(\mathbf{w}_{t-1}), \vec{C}_t^i; \vec{\mathcal{E}}_t^i(\mathbf{w}_t)). \quad (1)$$

Equation (1) presents the unit-level outcome dynamics. Importantly, evolution functions are assumed to be homogeneous across units and any heterogeneity arises through covariates C_t^i , and exposure vectors $\vec{\mathcal{E}}_t^i(\mathbf{w}_t)$. Consequently, two units with identical covariates and exposures will follow the same evolution path if initialized identically and share the exact same treatment trajectory.

Note that Eq. (1) characterizes the outcome evolution process, capturing how outcomes in each round depend on the previous round by incorporating $Y_{t-1}^i(\mathbf{w}_{t-1})$ on the right-hand side. This differs from the outcome generating process that maps treatments to potential outcomes without incorporating $Y_{t-1}^i(\mathbf{w}_{t-1})$ (Abadie et al. 2025). Our goal is to demonstrate how aggregate outcomes evolve according to these unit-level dynamics. To this end, we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with Ω representing the sample space, \mathcal{F} the sigma-algebra of events, and \mathbb{P} the probability measure.

ASSUMPTION 3 (Convergent potential outcomes). *For any treatment allocation \mathbf{w}_T , the potential outcome panel $\mathbf{Y}(\mathbf{w}_T)$ exists and satisfies the following condition.² For each round t , let $\hat{p}_{t;\mathbf{w}_t}^N$ denote the empirical distributions of $\{(w_t^i, Y_{t-1}^i(\mathbf{w}_{t-1}), \vec{C}_t^i, \vec{\mathcal{E}}_t^i(\mathbf{w}_t), Y_t^i(\mathbf{w}_t))\}_{i=1}^N$ that converges weakly, as $N \rightarrow \infty$, to a probability distribution $p_{t;\mathbf{w}_t}$.*

This assumption establishes the regularity conditions necessary for asymptotic analysis in our experimental setting with evolving outcomes. It requires that as the sample size increases, the joint distribution of key variables (current treatments and outcomes, recent outcomes, covariates, and exposure vectors) converges to a well-defined limiting distribution. This condition guarantees that the experimental environment exhibits required stability in large-sample regimes.

Notation convention. We fix a generic feasible treatment assignment \mathbf{w}_T , and omit explicit references when the context is clear. We also employ the following notations for any t :

$$(W_t^N, Y_{t-1}^N, \vec{C}_t^N, \vec{\mathcal{E}}_t^N, Y_t^N) \sim \hat{p}_{t;\mathbf{w}_t}^N \quad \text{and} \quad (W_t, Y_{t-1}, \vec{C}_t, \vec{\mathcal{E}}_t, Y_t) \sim p_{t;\mathbf{w}_t}, \quad (2)$$

without explicitly restating the detailed conditions specified in Assumption 3.

Assumptions 2 and 3 together with the definition of joint empirical distribution immediately imply that in finite samples where $N < \infty$, Y_t^N is equal to $\varphi(W_t^N, Y_{t-1}^N, \vec{C}_t^N; \vec{\mathcal{E}}_t^N)$. This equality is straightforward since each $(W_t^N, Y_{t-1}^N, \vec{C}_t^N, \vec{\mathcal{E}}_t^N, Y_t^N)$ take values with equal probability over the

² For (x^1, \dots, x^ℓ) , the empirical distribution is $\frac{1}{\ell} \sum_{j=1}^\ell \delta_{x^j}$, where δ_x is the Dirac measure at x . Weak convergence of probability measures μ_n to μ means $\int \psi d\mu_n \rightarrow \int \psi d\mu$ for all bounded continuous functions ψ .

corresponding values of all units. Indeed, each $(W_t^i, Y_{t-1}^i, \vec{C}_t^i, \vec{\mathcal{E}}_t^i, Y_t^i)$ will be observed under \hat{p}_t^N with a probability of $\frac{1}{N}$. However, to analyze the limiting behavior as $N \rightarrow \infty$, we require additional theoretical considerations, as established in the following theorem.

THEOREM 1 (ESE-I). *Under Assumptions 1-3, if the function φ is continuous, we have:*

$$Y_t \stackrel{\text{a.s.}}{=} \varphi(W_t, Y_{t-1}, \vec{C}_t; \vec{\mathcal{E}}_t) \quad (3)$$

where the recursion initiates from pretreatment outcome Y_0 and $\stackrel{\text{a.s.}}{=}$ denotes almost sure equality.

The result of Theorem 1 establishes the foundation for our subsequent analysis. The main objective of Theorem 1 is to rigorously characterize the assumptions required to ensure we can transition from unit-level evolution specified by Eq. (1) to the aggregate-level evolution outlined in Eq. (3). However, directly approaching Eq. (3) presents significant challenges as the exposure vectors $\vec{\mathcal{E}}_t$, which capture the interference structure, are essentially black boxes. Below, we provide a concise proof for Theorem 1, and in the subsequent section, we conduct a detailed analysis of the elements of Eq. (3).

Proof of Theorem 1. Let G_φ be the graph of the function φ :

$$G_\varphi := \left\{ (x, y) : x \in (\{0, 1\} \times \mathbb{R} \times \mathbb{R}^{n_c} \times \mathbb{R}^{n_e}), y \in \mathbb{R}, y = \varphi(x) \right\}.$$

Consider Assumptions 2 and 3, as well as the definition of empirical distributions, we know that

$$\hat{p}_{t; \mathbf{w}_t}^N(G_\varphi) = 1.$$

Since φ is a continuous function, the graph G_φ defines a closed set, and we can apply the Portmanteau theorem (e.g., Theorem 2.1 of Billingsley (2013)) to get

$$1 = \limsup_{N \rightarrow \infty} \hat{p}_{t; \mathbf{w}_t}^N(G_\varphi) \leq p_{t; \mathbf{w}_t}(G_\varphi),$$

which concludes the proof. \square

4. Interference Structure and Exposure Vectors

In this section, we examine the exposure vector concept as specified by Assumption 1. Basically, this assumption asserts that there exists a fixed-dimension vector that captures all relevant information from other units necessary to determine the potential outcome of each unit i . For example, the exposure vector $\vec{\mathcal{E}}_t^i$ may take the form of neighborhood averages or network centrality measures, any of which provides a finite-length summary sufficient to model interference effects (Toulis and Kao 2013, Basse et al. 2019, Forastiere et al. 2021, 2022, Leung 2022).

However, in realistic settings, we typically face a partially known interference structure. Our goal is to revise the ESE equation (Eq. (3)) with this perspective. We first explain the exposure mapping approach and decompose its concept into two distinct components: *network structure* (how units are connected) and *exposure mechanisms* (how those connected units influence each other).

4.1. Exposure Mapping

We proceed by adapting the exposure mapping definition from [Aronow and Samii \(2017\)](#) to our setting by incorporating the temporal dimension.

DEFINITION 1 (EXPOSURE MAPPING). Fix the treatment allocation \mathbf{w}_T . The exposure mapping of round t is a function e_t such that for all units i , we have

$$e_t : \{0, 1\}^N \times \mathbb{R}^N \times \mathbb{R}^{n_c} \mapsto \mathbb{R}^{n_e}, \quad \vec{\mathcal{E}}_t^i(\mathbf{w}_t) := e_t(\vec{w}_t, \vec{Y}_{t-1}(\mathbf{w}_{t-1}), \vec{C}_t^i). \quad (4)$$

Note that Definition 1 specifies the exposure vectors while explicitly differentiates between immediate spillover effects (arising immediately from treatment assignment \vec{w}_t) and peer effects that propagate through intermediate changes in units' outcomes $\vec{Y}_{t-1}(\mathbf{w}_{t-1})$.

The primary purpose of the exposure mapping framework is to simplify subsequent analysis by reducing model dimensionality (ideally achieving $n_e \ll N$). These mappings fundamentally encompass two distinct yet closely interrelated components of interference patterns:

- **Network structure:** This component specifies the connectivity structure between units through either binary relationships (e.g., adjacency matrices indicating whether units are connected) or weighted relationships (e.g., weighted directed graphs quantifying the magnitude of unit interactions). In the framework of [Aronow and Samii \(2017\)](#), this structural information can be incorporated into Eq. (4) through the covariate vector \vec{C}_t^i . For instance, \vec{C}_t^i may include the i -th row of a network adjacency matrix, thereby embedding unit i 's connectivity pattern.
- **Exposure mechanism:** This component describes the influence rules that govern how connected units affect one another's outcomes, essentially encoding the functional form of effects. Within the framework of [Aronow and Samii \(2017\)](#), these mechanisms are represented by the mathematical structure of the functions e_t in Eq. (4). For example, [Cai et al. \(2015\)](#) employs a model where spillover effects are captured through the fraction of treated neighbors, implementing an averaging mechanism with respect to peers' treatment assignments.

To effectively leverage this framework, properly specified exposure mappings are essential ([Aronow and Samii 2017](#)). Recent works study misspecified exposure mappings by restricting the scope of potential misspecification ([Leung 2022](#)) and separating the role of exposure mappings in defining causal effects of interest from assumptions about interference structure ([Sävje 2024](#)). Nevertheless, misspecification risk remains a substantial methodological challenge both at the exposure mechanism level (as discussed in [Auerbach et al. \(2024\)](#)) and through unobserved network connections ([Egami 2021](#), [Weinstein and Nevo 2023](#)).³

³ For insightful discussions on this topic, see the exchanges surrounding [Sävje \(2024\)](#) in [Auerbach et al. \(2024\)](#), [Leung \(2024\)](#), [Sävje \(2024\)](#).

While domain knowledge can often guide the specification of exposure mechanisms, identifying network structures presents significant challenges. Many settings in public health, marketplace experimentation, and social science involve large populations of interacting units. Tracking network connections in such contexts would require close observation of each experimental unit throughout the entire experiment. In what follows, we demonstrate how causal message-passing provides a robust framework that bypasses the need for explicit network structure knowledge; thereby, we gain the required insight for handling the exposure vector $\vec{\mathcal{E}}_t^i$ in our setting.

4.2. Causal Message Passing (CMP)

We begin by presenting a simplified version of the CMP outcome evolution (Shirani et al. 2025). Let g and h be two real-valued functions defined on $\{0, 1\} \times \mathbb{R} \times \mathbb{R}^{n_c}$. Consider A^{ij} and B_t^{ij} as unknown weights that quantify the *fixed* and *time-varying* influence of unit j on unit i , respectively. CMP considers the following family of outcome evolutions:

$$Y_t^i(\mathbf{w}_t) := h(w_t^i, Y_{t-1}^i(\mathbf{w}_{t-1}), \vec{C}_t^i) + \sum_{j=1}^N (A^{ij} + B_t^{ij}) g(w_t^j, Y_{t-1}^j(\mathbf{w}_{t-1}), \vec{C}_t^j). \quad (5)$$

Considering the outcome evolution specified in Eq. (1), CMP posits an additive structure. Specifically, Eq. (5) assumes that each unit's potential outcome evolves as the sum of two components: a unit-specific term (characterized by the h -part) and a weighted aggregation of all experimental units' status (characterized by A^{ij} , B_t^{ij} , and the g -part). From the exposure mapping perspective, CMP employs a one-dimensional exposure vector defined as

$$\mathcal{E}_t^i(\mathbf{w}_t) := \sum_{j=1}^N (A^{ij} + B_t^{ij}) g(w_t^j, Y_{t-1}^j(\mathbf{w}_{t-1}), \vec{C}_t^j). \quad (6)$$

In view of Eq. (4), CMP indeed separates the two components of exposure mapping under an additive assumption. Specifically, in Eq. (6), the weights $A^{ij} + B_t^{ij}$ capture the network structure, while the function g represents the underlying exposure mechanism.

Considering Eq. (6), CMP explicitly models how treatments trigger links. Note that the term $(A^{ij} + B_t^{ij})g(w_t^j, Y_{t-1}^j(\mathbf{w}_{t-1}), \vec{C}_t^j)$ describes the contribution of unit j to the outcome of unit i in round t . When unit j is treated in round t , the functionality of the link is shifted by replacing the baseline contribution $(A^{ij} + B_t^{ij})g(0, Y_{t-1}^j(\mathbf{w}_{t-1}), \vec{C}_t^j)$ with the treatment-induced contribution $(A^{ij} + B_t^{ij})g(1, Y_{t-1}^j(\mathbf{w}_{t-1}), \vec{C}_t^j)$. If this shift is non-zero, indicating that unit j impacts unit i , the resulting effect is incorporated into the outcome of unit i in the subsequent round.

CMP specializes to dense networks by modeling both first- and second-order interactions through independent Gaussian weights: $A^{ij} \sim \mathcal{N}(\mu^{ij}/N, \sigma^2/N)$ and $B_t^{ij} \sim \mathcal{N}(\mu_t^{ij}/N, \sigma_t^2/N)$. CMP then characterizes the limiting distribution of the interference effects. Specifically, let $\mathcal{E}_t(\mathbf{w}_t)$ follow the same

distribution as the empirical distribution of $\mathcal{E}_t^1(\mathbf{w}_t), \dots, \mathcal{E}_t^N(\mathbf{w}_t)$ as $N \rightarrow \infty$ in Eq. (6). Under certain regularity conditions, CMP establishes the following dynamics for the exposure effects:

$$\mathcal{E}_t(\mathbf{w}_t) \stackrel{\text{a.s.}}{=} \tilde{g}_t(W_t, Y_{t-1}, \vec{C}_t, Z_t), \quad (7)$$

where \tilde{g}_t is a function that depends on g (see Eq. (5)) and the distributions of the first-order weights $\mu^{ij} + \mu_t^{ij}$. Furthermore, Z_t is a Gaussian random variable that reflects the impact of the second-order Gaussian weights and W_t, Y_{t-1}, \vec{C}_t are as defined in Eq. (2).

Although this result allows CMP to bypass the observation of network structure by taking μ^{ij} , μ_t^{ij} , σ , and σ_t as unknowns, it still requires a rough characterization of the exposure mechanism. Specifically, Shirani and Bayati (2024) build their estimation by considering simple exposure mechanisms where the average of previous round outcomes, the average of current treatments, and their interaction term serve as proxies for interference patterns across units. Shirani et al. (2025) extend this result and allow consideration of a set of candidate exposure mechanisms. They then present a counterfactual cross-validation method to enable automatic selection of exposure mechanisms.

4.3. Partially Known Interference Structure

Our goal is to revise the ESE mappings in Eq. (3) by modeling the known and unknown components of the interference structure while ensuring the tractability of subsequent analysis. For this purpose, we consider the elements of the exposure vector for unit i in round t , as specified by Assumption 1:

$$\vec{\mathcal{E}}_t^i(\mathbf{w}_t) := (\mathcal{E}_t^{i(1)}(\mathbf{w}_t), \mathcal{E}_t^{i(2)}(\mathbf{w}_t), \dots, \mathcal{E}_t^{i(n_e)}(\mathbf{w}_t))^{\top}. \quad (8)$$

In view of Definition 1, we consider a sequence of “pseudo-functions” $e_t^1, \dots, e_t^{n_e}$ such that

$$e_t^l : \{0, 1\}^N \times \mathbb{R}^N \times \mathbb{R}^{n_c} \mapsto \mathbb{R}, \quad \mathcal{E}_t^{i(l)}(\mathbf{w}_t) := e_t^l(\vec{w}_t, \vec{Y}_{t-1}(\mathbf{w}_{t-1}), \vec{C}_t^i), \quad l = 1, \dots, n_e. \quad (9)$$

In Eq. (9), each e_t^l defines an operator from current treatments and previous outcomes to a scalar exposure measure, incorporating unit covariates. We use the term pseudo-function to emphasize that each e_t^l must remain well-defined when operating on infinite-dimensional vectors as $N \rightarrow \infty$.

Now, we can reflect the available information on interference structure through a careful design of pseudo-functions e_t^l . For example, suppose all we know is that the interaction network consists of K distinct clusters, each representing a cohesive subgroup with stronger internal connections than external ones. Then, we let $n_e = K$ and define e_t^l to capture how units belonging to cluster l impact unit i , with heterogeneity arising from unit-specific covariates \vec{C}_t^i . In the next step, inspired by CMP, we incorporate the unknown components in each direction of the exposure vector through the distribution of relevant treatments. In the clustered setting, we can assume that e_t^l reflects the influence of cluster l on unit i through the distribution of treatments of units belonging to cluster l .

Considering the random vector $\vec{\mathcal{E}}_t^i(\mathbf{w}_t)$ that follows the limiting empirical distribution of exposure vectors (as per Assumption 3), we formalize the necessary conditions in the following assumption.

ASSUMPTION 4 (Stable decomposition of interference). *For each round t and any treatment allocation \mathbf{w}_t , there exists a function f_t^e and random objects $\mathcal{I}_t(\mathbf{w}_t)$ and \mathcal{X}_t such that*

$$\vec{\mathcal{E}}_t(\mathbf{w}_t) \stackrel{\text{a.s.}}{=} f_t^e(\mathcal{I}_t(\mathbf{w}_t), \mathcal{X}_t),$$

where the functions f_t^e and the random object \mathcal{X}_t remain invariant across treatment allocations \mathbf{w}_t .

Assumption 4 is critical to our setting by decomposing the aggregate interference effects into three components: treatment-dependent factors $\mathcal{I}_t(\mathbf{w}_t)$, latent interference components \mathcal{X}_t , and the functional relationship f_t^e (which depends on pseudo-functions e_t^l). Specifically, the invariance condition (that both f_t^e and \mathcal{X}_t remain unchanged across treatment assignments) ensures that all variations in interference effects can be explained by random objects $\mathcal{I}_t(\mathbf{w}_t)$.

4.3.1. Illustrative examples We present several examples to clarify the conditions of Assumption 4 and explicitly characterize the treatment-endogenous object $\mathcal{I}_t(\mathbf{w}_t)$ in each case.

Causal message passing. Considering Eq. (7), CMP provides a concrete example where Assumption 4 holds under dense interference structures, with $\mathcal{I}_t(\mathbf{w}_t) = \{W_t, Y_{t-1}\}$, reflecting the distribution of current treatments and past outcomes. Notably, the additive model specification and second-order i.i.d. random weights enable a clean characterization of the exogenous component as $\mathcal{X}_t = \{\vec{C}_t, Z_t\}$. While the covariate component \vec{C}_t may be partially observable, Z_t can be interpreted as unit-level noise in large-sample regimes (Shirani et al. 2025).

Clustered network structure. Reconsider the clustered setting with K clusters, and let $W_t^{(l)}$ denote a random variable distributed as the treatments assigned to units in cluster l , for $l = 1, \dots, K$. In this case, $\mathcal{I}_t(\mathbf{w}_t) = \{W_t^{(1)}, \dots, W_t^{(K)}\}$. Intuitively, the treatment-endogenous object reduces to the collection of cluster-level treatment distributions. Combined with the unknown operators e_t^l , whose structure is captured by f_t^e in Assumption 4, this highlights the fact that each unit may be affected differently by each cluster of units.

Mean-field interference mechanisms. Another setting where Assumption 4 holds is when interference is mediated by mean-field quantities. For instance, Munro et al. (2021) analyze a market where spillovers arise through equilibrium prices that match supply and demand. In general, we may consider a mean-field quantity $M_t(\mathbf{w}_t)$ that mediates interference effects for each unit under treatment scenario \mathbf{w}_t . In this case, the treatment-endogenous object is $\mathcal{I}_t(\mathbf{w}_t) := \{M_t(\mathbf{w}_t)\}$.

Note that Assumption 4 can also be viewed through a mean-field perspective. Indeed, the treatment-driven object $\mathcal{I}_t(\mathbf{w}_t)$ serves as the mean-field quantity that summarizes *aggregate* interference effects. This is more general than standard mean-field models, which typically assume that *unit-level* interference is mediated by mean-field quantities. A deeper investigation of this connection is left for future work, as its relevance is context-dependent and in some applications unit-level mean-fields might be sufficient (Wager and Xu 2021, Johari et al. 2022, Munro et al. 2021).

Social influencers. Consider an experiment on a social media platform involving a small set of influencers and a large population of regular users. Influencers are users with disproportionately high reach and engagement who can affect the behavior of many others. Because of their outsized impact, we track each influencer’s treatment status individually while summarizing the remaining users through an aggregate measure. Without loss of generality, index units so that the influencers correspond to 1 through $n_e - 1$. We then define $\mathcal{I}_t(\mathbf{w}_t) = \{W_t^1, \dots, W_t^{n_e-1}, W_t\}$, where the first $n_e - 1$ components capture individual influencer treatments, and the final component W_t represents the treatment distribution across all remaining users.

4.4. Revised Experimental State Evolution

Finally, we revise the result of Theorem 1; considering the conditions of Assumption 4 on aggregate interference patterns, the following result is immediate.

THEOREM 2 (ESE-II). *Under Assumptions 1-4, if the function φ is continuous, there exist random functions f_t , $t = 1, \dots, T$, such that*

$$Y_t \xrightarrow{\text{a.s.}} f_t(W_t, Y_{t-1}; \mathcal{I}_t(\mathbf{w}_t)), \quad (10)$$

where

$$f_t(W_t, Y_{t-1}; \mathcal{I}_t(\mathbf{w}_t)) := \varphi\left(W_t, Y_{t-1}, \vec{C}_t; f_t^e(\mathcal{I}_t(\mathbf{w}_t), \mathcal{X}_t)\right) \quad (11)$$

and the recursion initiates from pretreatment initial outcomes Y_0 .

By the definition in Eq. (11), for each t , the ESE mapping f_t depends on unit-level evolutions captured through φ , interference functional structure reflected by f_t^e , and limiting distributions of covariates and exogenous interference components (captured by \vec{C}_t and \mathcal{X}_t , respectively).

Theorem 2 characterizes three distinct components of the experimental system. First, the initial outcomes collected before any treatment intervention, whose limiting empirical distribution is captured by Y_0 . Second, the ESE mappings f_t that encode the evolution mechanisms governing the response of the experimental system to the treatment assignment. Crucially, these first two components remain unchanged regardless of treatment allocation. Third, the variable inputs to the ESE mappings that change across different treatment scenarios.

This decomposition reveals a key insight: *ESE mappings characterize parallel evolution patterns across treatment scenarios, where aggregate outcomes evolve under identical mechanisms but with scenario-specific inputs.* We explore this parallel structure in the next section.

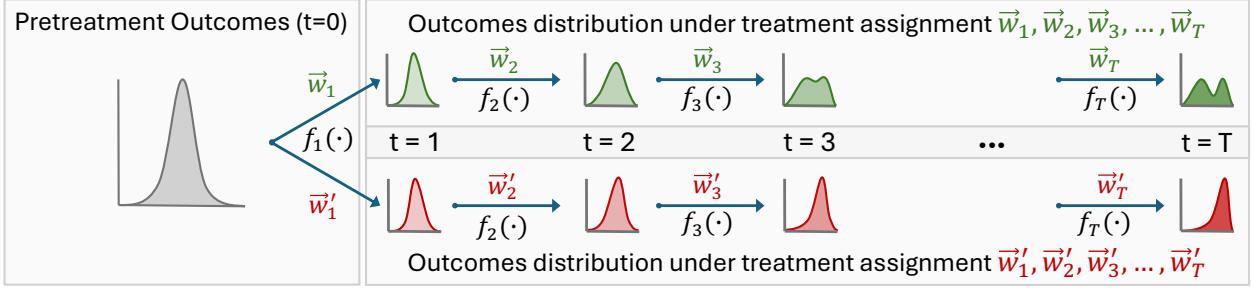


Figure 4 Functional parallelism: both scenarios evolve under same ESE mappings despite distinct distributions.

5. Estimation Strategy: Distributional Parallel Propagations

This section explains the parallel structures identified by ESE mappings across *different treatment scenarios* and explains when they enable counterfactual estimation. To illustrate, consider two distinct treatment assignments \mathbf{w}_T and \mathbf{w}'_T and their corresponding panels of potential outcomes:

$$\mathbf{Y}(\mathbf{w}_T) := [\vec{Y}_0 | \vec{Y}_1(\mathbf{w}_1) | \dots | \vec{Y}_T(\mathbf{w}_T)] \quad \mathbf{Y}(\mathbf{w}'_T) := [\vec{Y}_0 | \vec{Y}_1(\mathbf{w}'_1) | \dots | \vec{Y}_T(\mathbf{w}'_T)]. \quad (12)$$

From Theorem 2 and Eq. (12), each outcome panel characterizes a distinct trajectory with specific structural properties. First, both trajectories originate from a *common baseline*: the pretreatment observations \vec{Y}_0 . Beginning in round $t = 1$, these trajectories diverge into two distinct outcome vectors $\vec{Y}_1(\mathbf{w}_1)$ and $\vec{Y}_1(\mathbf{w}'_1)$. Crucially, while these outcomes differ, their distributions emerge from \vec{Y}_0 according to the same mathematical rule (the ESE mapping $f_1(\cdot)$) applied to different inputs \mathbf{w}_1 and \mathbf{w}'_1 , respectively. This parallel evolutionary mechanism persists throughout the experiment timeline, generating two distinct but structurally parallel potential evolutions (Figure 4).

Now, suppose we observe outcomes under \mathbf{w}_T and aim to estimate the counterfactuals under \mathbf{w}'_T . In the first stage, we use the observed outcomes $\mathbf{Y}(\mathbf{W}_T = \mathbf{w}_T)$ to estimate the ESE mappings f_t . In the second stage, we exploit the parallel structure across treatment scenarios to construct counterfactual trajectories recursively. Starting from the common baseline of pretreatment outcomes \vec{Y}_0 , we apply the estimated ESE mappings sequentially under the desired treatment assignment \mathbf{w}'_T , which yields the aggregate counterfactual outcomes for the entire experimental population.

Similar evolution-based ideas appear in the difference-in-differences (DID) (Angrist and Pischke 2009) and synthetic control methods (SCM) (Abadie et al. 2010). DID relies on the *parallel trends* assumption: the unobserved counterfactual trajectory of the treated group is assumed to evolve in parallel with the observed trajectory of the control group, starting from the pretreatment outcomes. SCM also uses pretreatment outcomes and constructs the counterfactual evolution of the treated group as a weighted combination of control group trajectories.

In general, parallel structures in the evolution of potential outcomes allow us to propagate forward from a common baseline and construct counterfactual trajectories through a sequential process. Under SUTVA, such parallelisms can be studied at the unit level, as in DID and SCM. Recent work extends this idea to networked settings by introducing modified parallel trends assumptions that account for interference (Xu 2023, Jetsupphasuk et al. 2025). However, our framework shifts the perspective to the aggregate level, where we model and characterize a conceptually similar form of functional parallelism at the distributional level.

Estimating the unknown ESE mappings f_t is still challenging in practice. With only a single realization of the outcome panel, the information available for learning these mappings is limited. Focusing on treatment-induced variations provides new opportunities for estimation. In many applications, it is not necessary to recover the exact functional form of the ESE mappings, and suitable approximations may be sufficient. To formalize this idea, we consider two distinct treatment assignments: the baseline scenario with no treatment and \mathbf{w}_T representing an alternative assignment. Under Assumptions 3 and 4, and using the notation established in Eq. (2), we adopt the following conventions for any t :

$$\begin{aligned} \left(0, Y_{t-1}^0, \vec{C}_t^0, \vec{\mathcal{E}}_t^0, Y_t\right) &\sim p_{t;0}, & \vec{\mathcal{E}}_t^0 &\stackrel{\text{a.s.}}{=} f_t^e(\mathcal{I}_t^0, \mathcal{X}_t), \\ \left(W_t, Y_{t-1}, \vec{C}_t, \vec{\mathcal{E}}_t, Y_t\right) &\sim p_{t;\mathbf{w}_t}, & \vec{\mathcal{E}}_t &\stackrel{\text{a.s.}}{=} f_t^e(\mathcal{I}_t, \mathcal{X}_t). \end{aligned} \quad (13)$$

Here, the two sequences of random variables, denoted by Y_t^0 and Y_t , represent the outcome distributions under their corresponding treatment scenarios.

THEOREM 3. *Suppose that the conditions of Theorem 2 hold and the ESE mappings f_t are three times continuously differentiable for all t . Then,*

$$Y_t = \alpha_t W_t + \beta_t Y_{t-1} + \gamma_t \mathcal{I}_t + \eta_t W_t Y_{t-1} + \theta_t W_t \mathcal{I}_t + \delta_t + R_t, \quad (14)$$

where $\alpha_t, \beta_t, \gamma_t, \eta_t, \theta_t$, and δ_t are random coefficients depending on the baseline scenario (with no treatment) and independent of W_t . Additionally, the term R_t collects the unretained second-order terms containing $(Y_{t-1} - Y_{t-1}^0)^2$, $(\mathcal{I}_t - \mathcal{I}_t^0)^2$, or $(Y_{t-1} - Y_{t-1}^0)(\mathcal{I}_t - \mathcal{I}_t^0)$, together with the third-order derivatives of f_t evaluated at an intermediate point between the baseline and the perturbed inputs.

The relation in Eq. (14) captures how the treatment intervention, together with network effects, shifts the outcomes relative to the baseline scenario. When treatment effects are small in magnitude, the quantities Y_{t-1} and Y_{t-1}^0 (as well as \mathcal{I}_t and \mathcal{I}_t^0) are expected to remain sufficiently close. In this case, the remainder term R_t in Eq. (14) can be ignored. The smoothness assumption in Theorem 3 is also standard in the literature, see for example Li and Wager (2022b).

Proof of Theorem 3. By the result of Theorem 2 and Assumption 4, we know that

$$Y_t^0 \stackrel{\text{a.s.}}{=} f_t(0, Y_{t-1}^0; \mathcal{I}_t^0), \quad Y_t \stackrel{\text{a.s.}}{=} f_t(W_t, Y_{t-1}; \mathcal{I}_t), \quad (15)$$

where both recursions start from the same pre-treatment outcome Y_0 . Note that Eq. (15) involves two sequences of almost sure events. Since the union of a countable sequence of measure-zero sets is a measure-zero set, there exists $\mathcal{S} \subseteq \Omega$ with $\mathbb{P}(\mathcal{S}) = 1$ such that for all $\omega \in \mathcal{S}$, the equalities in Eq. (15) hold for all t (the dependence of random variables on ω is omitted for clarity of presentation). Applying Taylor's expansion pathwise for every $\omega \in \mathcal{S}$, we get the following:

$$\begin{aligned} Y_t &= Y_t^0 + W_t \partial_x f_t + (Y_{t-1} - Y_{t-1}^0) \partial_y f_t + (\mathcal{I}_t - \mathcal{I}_t^0) \partial_z f_t \\ &\quad + \frac{1}{2} (W_t)^2 \partial_x^2 f_t + W_t (Y_{t-1} - Y_{t-1}^0) \partial_{xy} f_t + W_t (\mathcal{I}_t - \mathcal{I}_t^0) \partial_{xz} f_t + R_t, \end{aligned} \quad (16)$$

where $\partial_x f_t$, $\partial_y f_t$, and $\partial_z f_t$ denote the partial derivatives of $f_t(0, Y_{t-1}^0, \mathcal{I}_t^0)$ with respect to its first, second, and third arguments respectively; higher-order derivatives follow the same notation convention. In Eq. (16), the term R_t collects the unretained second-order terms together with the third-order derivatives of f_t as the remainder of our Taylor expansion. Considering W_t is a 0-1 variable, we know that $(W_t)^2 = W_t$ and we can rewrite Eq. (16) as given by Eq. (14) with

$$\begin{aligned} \alpha_t &:= \partial_x f_t + \frac{1}{2} \partial_x^2 f_t - Y_{t-1}^0 \partial_{xy} f_t - \mathcal{I}_t^0 \partial_{xz} f_t, \\ \beta_t &:= \partial_y f_t, \quad \gamma_t := \partial_z f_t, \quad \eta_t := \partial_{xy} f_t, \quad \theta_t := \partial_{xz} f_t, \\ \delta_t &:= Y_t^0 - Y_{t-1}^0 \partial_y f_t - \mathcal{I}_t^0 \partial_z f_t. \end{aligned}$$

This concludes the proof. \square

The proof of Theorem 3 reveals that randomized treatment assignment plays a fundamental role. Because treatments are assigned independently of all other factors, including the initial pre-treatment outcomes and so the baseline outcomes, we can clearly isolate treatment-induced variations (whether direct or network-mediated) in aggregate outcomes from the underlying model parameters. This separation enables simplifying the estimation of unknown ESE mappings to the estimation of the unknown coefficients in Eq. (14). Once these coefficients are estimated, we can substitute W_t in Eq. (14) with treatment assignments from alternative scenarios and, starting from the same pre-treatment outcomes Y_0 , recursively construct other counterfactual trajectories.

We emphasize that, for the coefficients in Eq. (14) to be identifiable, additional structural assumptions are needed. For example, in the CMP framework, where $\mathcal{I}_t = \{W_t, Y_{t-1}\}$, the right-hand side of Eq. (14) admits a simpler form. If we further assume that the coefficients are time-invariant, this formulation reduces to the model of Shirani and Bayati (2024), for which consistency results are available. For the other examples in § 4.3.1, Eq. (14) should be reformulated to incorporate the specific structural features of the interference mechanisms in each setting and to ensure that the resulting model remains identifiable. A full examination of this problem lies beyond the scope of this work and should be tailored to each setting based on the experimental context.

6. Discussion and Conclusion

This work adopts an evolution-based perspective on counterfactual estimation under interference. Instead of modeling outcomes directly as functions of treatments, we focus on evolutionary models that capture how outcomes transition from one observation to the next, characterizing the treatment-driven variations. Related approaches, such as Markovian interference models (Farias et al. 2022) and switchback experiments (Hu and Wager 2022, Bojinov et al. 2023), rely *directly* on temporal outcomes to improve estimation and design. Our framework instead centers on the evolution rules: the mechanism that operates *between* observations and captures how treatment-triggered interference drives aggregate changes in outcome dynamics.

Building on the causal message passing framework, which offers a concrete instantiation of evolution-based estimation (Shirani and Bayati 2024, Shirani et al. 2025), we take a broader view. We investigate the minimal structural conditions under which evolution-based analysis is possible, albeit imposing stronger regularity assumptions on potential outcomes. We also highlight the notion of distributional parallel propagation as the functional parallelism across counterfactual scenarios. These results clarify how evolution-based approaches should be interpreted.

The core operating assumption requires a stable decomposition of aggregate interference effects into two parts: an endogenous component that varies across treatment scenarios and an exogenous component that remains invariant to treatment assignment. This decomposition enables isolating treatment-driven variations as outcomes evolve during the experiment. However, in certain settings, this invariance property is unrealistic. Exposure mechanisms may themselves shift with treatment. For example, in systems with threshold dynamics where peer influence activates only after a critical mass of the treatment adoption, the required decomposition does not hold.

Relying on treatment-driven variations over time also constrains the applicability of evolution-based frameworks. To illustrate this, we apply two versions of the CMP algorithm to the election-simulator data of Shirani and Bayati (2025), which replicates the voter mobilization experiment of Bond et al. (2012). The simulator generates a synthetic social network and records voting intentions for 20,000 users beginning 40 days before election day.

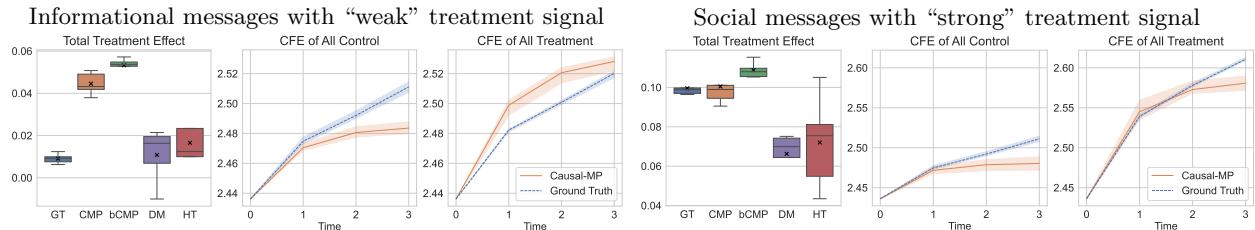


Figure 5 The performance of evolution-based estimators under strong time trends: CMP succeeds when the treatment signal is strong (social message) but struggles when it is weak (informational message).

The experiment includes two treatment scenarios: an informational message that provides factual voting details and a social message that additionally displays peer behavior. Each user is independently assigned treatment with probability 0% in the first 10 days and with probabilities 20%, 40%, and 80% in the second, third, and fourth 10-day blocks, respectively. Because users are not online every day, the raw data contain missing outcomes. We therefore aggregate each 10-day block into a single round, producing a panel with four rounds. Since the simulator is fully controlled, both scenarios provide ground-truth outcomes, enabling evaluation of the estimation method.

Figure 5 reports the results, comparing ground-truth (GT) values with estimates from basic CMP (bCMP) (Shirani and Bayati 2024), the full CMP method (Shirani et al. 2025), the difference-in-means estimator (DM), and the Horvitz–Thompson estimator (HT)⁴ (Sävje et al. 2021). For each scenario, we also plot the counterfactual evolution of the sample-mean outcomes (CFE) under both all-control and all-treatment conditions, using the GT values and the CMP estimates.

In both scenarios, the middle panels show a clear upward trend in the all-control case, reflecting the natural increase in voting intention as election day approaches. In the informational-message scenario (left panels), where the treatment signal is weak, both CMP methods struggle to separate treatment-induced variation from this strong time trend and therefore overestimate the treatment effect. In contrast, in the social-message scenario (right panels), where the treatment signal is substantially stronger, both CMP methods succeed in distinguishing treatment-driven changes from the underlying time trend and produce improved treatment-effect estimates.

In conclusion, in settings where network interference is widespread, it may not be practical to attempt direct observation of the underlying network. Evolution-based approaches are particularly useful in such cases, but their validity depends on the interference structures remaining stable across treatment allocations. When this stability holds, methods such as causal message passing algorithms offer a principled and cost-effective strategy for estimating treatment effects.

⁴ Difference-in-means (DM) and Horvitz-Thompson (HT) are expressed as:

$$\hat{\tau}_{\text{DM}} := \frac{\sum_{i=1}^N Y_T^i W_T^i}{\sum_{i=1}^N W_T^i} - \frac{\sum_{i=1}^N Y_T^i (1 - W_T^i)}{\sum_{i=1}^N (1 - W_T^i)}, \quad \hat{\tau}_{\text{HT}} := \frac{1}{N} \sum_{i=1}^N \left(\frac{Y_T^i W_T^i}{\mathbb{E}[W_T^i]} - \frac{Y_T^i (1 - W_T^i)}{\mathbb{E}[1 - W_T^i]} \right).$$

References

- Abadie, A., Agarwal, A., and Shah, D. (2025). A causal inference framework for data rich environments. *arXiv preprint arXiv:2504.01702*.
- Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Aronow, P. M. and Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912 – 1947.
- Auerbach, E., Auerbach, J., and Tabord-Meehan, M. (2024). Discussion of ‘causal inference with misspecified exposure mappings: separating definitions and assumptions’. *Biometrika*, 111(1):21–24.
- Basse, G. W., Feller, A., and Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494.
- Bayati, M., Luo, Y., Overman, W., Shirani Faradonbeh, M. S., and Xiong, R. (2024). Higher-order causal message passing for experimentation with complex interference. *Advances in Neural Information Processing Systems*, 37:81836–81856.
- Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.
- Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.
- Bojinov, I., Simchi-Levi, D., and Zhao, J. (2023). Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777.
- Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.
- Boyarsky, A., Namkoong, H., and Pouget-Abadie, J. (2023). Modeling interference using experiment roll-out. *arXiv preprint arXiv:2305.10728*.
- Cai, J., Janvry, A. D., and Sadoulet, E. (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108.
- Casadiego, J., Nitzan, M., Hallerberg, S., and Timme, M. (2017). Model-free inference of direct network interactions from nonlinear collective dynamics. *Nature communications*, 8(1):2192.
- Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.
- Eckles, D., Karrer, B., and Ugander, J. (2016). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021.
- Egami, N. (2021). Spillover effects in the presence of unobserved networks. *Political Analysis*, 29(3):287–316.
- Farias, V., Li, A., Peng, T., and Zheng, A. (2022). Markovian interference in experiments. *Advances in Neural Information Processing Systems*, 35:535–549.

- Forastiere, L., Aioldi, E. M., and Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918.
- Forastiere, L., Mealli, F., Wu, A., and Aioldi, E. M. (2022). Estimating causal effects under network interference with bayesian generalized propensity scores. *Journal of Machine Learning Research*, 23(289):1–61.
- Glynn, P. W., Johari, R., and Rasouli, M. (2020). Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064.
- Han, K., Basse, G., and Bojinov, I. (2024). Population interference in panel experiments. *Journal of Econometrics*, 238(1):105565.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.
- Hu, Y. and Wager, S. (2022). Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*.
- Hudgens, M. G. and Halloran, M. E. (2012). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- Jackson, M. O. et al. (2008). *Social and economic networks*, volume 3. Princeton university press Princeton.
- Jetsupphasuk, M., Li, D., and Hudgens, M. G. (2025). Estimating causal effects using difference-in-differences under network dependency and interference. *arXiv preprint arXiv:2502.03414*.
- Jia, Z., Kallus, N., and Yu, Z. S. (2025). Clustered switchback designs for experimentation under spatio-temporal interference. *arXiv preprint arXiv:2312.15574*.
- Johari, R., Li, H., Liskovich, I., and Weintraub, G. Y. (2022). Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 68(10):7069–7089.
- Kolaczyk, E. D. and Csárdi, G. (2014). *Statistical analysis of network data with R*, volume 65. Springer.
- Leung, M. P. (2022). Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293.
- Leung, M. P. (2024). Discussion of ‘causal inference with misspecified exposure mappings: separating definitions and assumptions’. *Biometrika*, 111(1):17–20.
- Li, S., Johari, R., Kuang, X., and Wager, S. (2023). Experimenting under stochastic congestion. *arXiv preprint arXiv:2302.12093*.
- Li, S. and Wager, S. (2022a). Network interference in micro-randomized trials. *arXiv preprint arXiv:2202.05356*.
- Li, S. and Wager, S. (2022b). Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358.
- Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.
- Mukaigawara, M., Imai, K., Lyall, J., and Papadogeorgou, G. (2025). Spatiotemporal causal inference with arbitrary spillover and carryover effects. *arXiv preprint arXiv:2504.03464*.
- Munro, E., Wager, S., and Xu, K. (2021). Treatment effects in market equilibrium. *arXiv preprint arXiv:2109.11647*.

- Ni, T., Bojinov, I., and Zhao, J. (2023). Design of panel experiments with spatial and temporal interference. *Available at SSRN 4466598*.
- Nitzan, M., Casadiego, J., and Timme, M. (2017). Revealing physical interaction networks from statistics of collective dynamics. *Science advances*, 3(2):e1600396.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.
- Sävje, F. (2024). Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):1–15.
- Sävje, F., Aronow, P., and Hudgens, M. (2021). Average treatment effects in the presence of unknown interference. *Annals of statistics*, 49(2):673.
- Shirani, S. and Bayati, M. (2024). Causal message-passing for experiments with unknown and general network interference. *Proceedings of the National Academy of Sciences*, 121(40):e2322232121.
- Shirani, S. and Bayati, M. (2025). Simulating and experimenting with social media mobilization using llm agents. *arXiv preprint arXiv:2510.26494*.
- Shirani, S., Luo, Y., Overman, W., Xiong, R., and Bayati, M. (2025). Can we validate counterfactual estimations in the presence of general network interference? *arXiv preprint arXiv:2502.01106*.
- Stepaniants, G., Brunton, B. W., and Kutz, J. N. (2020). Inferring causal networks of dynamical systems through transient dynamics and perturbation. *Physical Review E*, 102(4):042309.
- Sävje, F. (2024). Rejoinder: Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):25–29.
- Timme, M. and Casadiego, J. (2014). Revealing networks from dynamics: an introduction. *Journal of Physics A: Mathematical and Theoretical*, 47(34):343001.
- Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR.
- Ugander, J., Karrer, B., Backstrom, L., and Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.
- Wager, S. and Xu, K. (2021). Experimenting in equilibrium. *Management Science*, 67(11):6694–6715.
- Weinstein, B. and Nevo, D. (2023). Causal inference with misspecified network interference structure. *arXiv preprint arXiv:2302.11322*.
- Xu, R. (2023). Difference-in-differences with interference. *arXiv preprint arXiv:2306.12003*.
- Yu, C. L., Airolди, E. M., Borgs, C., and Chayes, J. T. (2022). Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119.