Direct Inference of Effect of Treatment (DIET) for a Cookieless World

Shiv Shankar UMass
Ritwik Sinha Adobe Research
Saayan Mitra Adobe Research
Moumita Sinha Adobe Inc
Madalina Fiterau UMass

Abstract

Brands use cookies and device identifiers to link different web visits to the same consumer. However, with increasing demands for privacy, these identifiers are about to be phased out, making identity fragmentation a permanent feature of the online world. Assessing treatment effects via randomized experiments (A/B testing) in such a scenario is challenging because identity fragmentation causes a) users to receive hybrid/mixed treatments, and b) hides the causal link between the historical treatments and the outcome. In this work, we address the problem of estimating treatment effects when a lack of identification leads to incomplete knowledge of historical treatments. This is a challenging problem which has not been addressed in literature yet. We develop a new method called DIET, which can adjust for users being exposed to mixed treatments without *the entire history of treatments being available*. Our method takes inspiration from the Cox model, and uses a proportional outcome approach under which we prove that one can obtain consistent estimates of treatment effects even under identity fragmentation. Our experiments, on one simulated and two real datasets, show that our method leads to up to 20% reduction in error and 25% reduction in bias over the naive estimate.

1 INTRODUCTION

Enterprises devote significant effort to optimize their customers' experience on websites and mobile applications. The standard approach to optimizing user experience is to conduct randomized experiments, commonly referred to as "A/B Tests". All randomized experiments require that we reliably identify the same user across their whole journey, so that we can show them a consistent experience. Unfortunately, the rapid evolution of the web ecosystem makes this assumption fraught. There are multiple reasons for this. First, the explosion in the number of devices means that most people are concurrently using multiple devices (e.g., a phone, a tablet, or different browsers on a laptop) to interact with the same brand (through apps or websites). Often, this interaction happens while the customer is not logged into their account. Collectively, this trend means that the customer's identity is fragmented and their experience may not be consistent, since one device may be served experience "A" while another sees "B". A second source of fragmentation comes from an increased focus on individual privacy, as embodied in laws like General Data Protection Regulation (GDPR)[1]. GDPR requires websites to explicitly require the visitor's permission before using any cookies that may be used to track the user across their web sessions. Aligning with this trend, all major browser and app ecosystems have taken or planned steps to discourage the use of cookies and device identifier for marketing and experience optimization (Seufert, 2020; Schiff, 2020; Bohn, 2020).

This phenomenon where a single user's activities get associated with mutually disconnected identifiers with each capturing only a fraction of the exposures and behaviors is known as *Identity fragmentation* (Saha Roy et al., 2015; Coey and Bailey, 2016; Lin and Misra, 2021). In the presence of fragmentation, treatment exposures cannot be connected to outcomes, thus making the task of reliably estimating causal effects challenging. For example, a user might see an ad for a product on their phone but purchase it on their laptop. Unless the marketer can connect the two sessions to the same user, the estimated effect of the ad (Sinha et al., 2014) is biased (known as *Fragmentation Bias*). A common way to overcome this is to connect sessions that belong to the same user; perhaps using features such as IP address, location, or device type. This idea is known as *session stitching* or *identity linking* (Fellegi and Sunter, 1969). However this method a) has no formal guarantees on linking accuracy, b) does not allow for quantifying the uncertainty of the estimate, and c) partially linked

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

[1] https://gdpr-info.eu/

data can be misleading. Moreover as we show, the bias of estimated effects under partial linkage can be difficult to quantify (See AppendixB.3).

Fragmentation, while common in digital marketing, is not limited to this area alone. It crops up in many areas such as health care (Jason, 2020), social sciences (Ruggles et al., 2018) and public administration (Churches et al., 2002). In medical studies, the related problem of non-compliance (patients takes a drug different from that assigned to them) has received some attention (Sagarin et al., 2014), but the scale of this problem (proportion of non-compliers) is still assumed to be a small fraction of the sample size of the experiment.

Most methods for estimating treatment effects require complete knowledge of historical treatments, which is not possible under identity fragmentation. In this work we address the challenging task of estimating treatment effects under fragmentation, by considering it as a form of interference. To the best of our knowledge, ours is the first work to address this problem. Inspired by the proportional hazard models (Cox, 1972), we propose an estimator based on multiplicative effects which we call Direct Inference of Effect of Treatment (*DIET*). We conduct experiments with DIET on observational logs of online sales from an e-tailer and patient health care records, and find that DIET can reduce error by upto 20%.

Contributions: Our paper is the first to address the estimation of sequential/cumulative treatment effect *under identity fragmentation*. Our proposed method goes beyond cookies as the unit of analysis in A/B testing. Finally, our theoretical results generalize the results of Coey and Bailey (2016) and Lin and Misra (2021) on attenuation due to fragmentation.

2 PROBLEM FORMULATION

2.1 Motivation

Consider the example of an online retailer which is looking to deploy a new recommendation system S in place of their existing system E. To assess the efficacy of the new system, the company performs an A/B test by delivering recommendation from S (which we denote as treatment 1) to a portion of its online visitors. Strategy S is considered superior to E if the outcome of interest (say purchase) is statistically significantly more for S compared to E. The users can visit the website from both their laptop and mobile without signing in. Without availability of a unique identifier the retailer is not able to consistently use S tofor giveing the user recommendations. For example, at time step 1, the user visited the website on a laptop and was assigned treatment 1. The next visit, if happening from the mobile device, means that the user may now get recommendation from E (treatment$=0$). As such, purchases from this user areis due to the mixed influence of both treatments. Since the final goal is to evaluated the benefits of theis regarding wide scale deployment of S, we would like ; ideally we want to compare the outcomes when the users always got treatment 1. However, considering the mixed treatment outcome for the aforementioned user can bias the treatment effect estimates. Moreover, we do not have any way of knowing whether the user received a mixed treatment or not.

2.2 Notation

We use the Neyman potential outcome framework for analysis (Neyman, 1923; Rubin, 1974). Suppose there are N units (individuals) interacting with an online experimentation system. Each of the units has associated covariates X which are fixed and observed. Each unit is observed for T consecutive periods. In each period t, we know the outcome Y_{it} and the treatment status $Z_{it} \in \{0, 1\}$ for each unit i [2].

We denote a random variable or its value in a particular period with subscript and their past history with superscript. For example, $\mathbf{Y}_i^{1:t} := (Y_{i1}, Y_{i2}, \ldots, Y_{it})$ and $\mathbf{Z}_i^{1:t} := (Z_{i1}, Z_{i2}, \ldots, Z_{it})$ are i's outcome values and treatment status up to period t, respectively; $\mathbf{Z}^{t:s} := (Z_{1t}, \ldots, Z_{Nt}, \ldots, Z_{1s}, \ldots, Z_{Ns})$ is the history of treatment assignment for all the units between period t and period s. Moreover, we use $\mathbf{Z}^{t:s} \setminus \mathbf{Z}_i^{t':s'}$ to denote the same history without unit i's treatment status between time t' and s'.

The outcome for each unit i at period t, Y_{it}, is jointly decided by the entire sequence of treatment assignments. This is a form of interference, however the interference is only temporal, and a unit's outcome only depends on their own treatments and not on any others, that is,

$$Y_{it} = Y_{it}(\mathbf{Z}_i^{1:T}) = Y_{it}(Z_{i1}, Z_{i2}, \ldots, Z_{iT}).$$

We are interested in the causal effect generated by the change of a treatment assignment history. For each unit i at time t, we define the individualistic treatment effect of history $\mathbf{z}^{1:T}$ relative to history $\tilde{\mathbf{z}}^{1:T}$ as:

$$\tau_{it}(\mathbf{z}^{1:T}, \tilde{\mathbf{z}}^{1:T}) := Y_{it}(\mathbf{z}^{1:T}) - Y_{it}(\tilde{\mathbf{z}}^{1:T}).$$

Two quantities which are of importance for this work are the 'instantaneous' and 'long-term' effects of a treatment allocation to a unit. Formally these are defined as:

- Contemporary direct effect: $\tau_{it}(1, 0; \mathbf{Z}^{1:T} \setminus Z_{it}) := Y_{it}(1, \mathbf{Z}^{1:T} \setminus Z_{it}) - Y_{it}(0, \mathbf{Z}^{1:T} \setminus Z_{it})$. It captures the

[2] We are assuming a simplified setting where there are no new users, and where we can identify a consistent time index t for the user's visits without actual linking of them. In practice, without cookies such indexing too is challenging

effect of observation (i,t)'s treatment, Z_{it}, on its current outcome, Y_{it}, with all other treatments remaining the same.

- Cumulative direct effect: $\tau_{it}(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t}; \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}) := Y_{it}(\mathbf{z}^{s:t}, \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}) - Y_{it}(\tilde{\mathbf{z}}^{s:t}, \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t})$, where $s \leq t$. This is the cumulative effect of unit i's history between period s and t on its own outcome in period t and induced by temporal interference. Once again any treatment not included in the period $s:t$ are kept identical.

These quantities are often unidentifiable from observational data. A common practice is to marginalize over both treatment assignments and over the N units to get Expected Average Treatment Effect (EATE) (Pearl, 2000). The corresponding effects are given by[3]:

$$\tau_t := \frac{1}{N}\sum_{i=1}^{N} \tau_{it} = \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:T} \setminus Z_{it}} \left[\tau_{it}(1, 0; \mathbf{Z}^{1:T} \setminus Z_{it}) \right],$$

$$\tau_t(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t}) := \frac{1}{N}\sum_{i=1}^{N} \tau_{it}(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t}) = \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}} \left[\tau_{it}(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t}; \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}) \right].$$

2.2.1 Problem Statement

In the absence of cookies, each visit by a unit has to be treated as a separate unit. This cookie-free unit has no history; so while the outcome observed is $Y_{it}(\mathbf{Z}_i^{1:T})$, only treatment Z_{it} is observed. The question of concern for this work is estimating the cumulative direct effect $\tau_t(\vec{\mathbf{1}}, \vec{\mathbf{0}})$ (which we will sometimes also call global treatment effect) of using treatment 1 over 0. This is the effect of the marketer deploying the treatment $Z_{i,t} = 1 \, \forall t$ against the treatment $Z_{i,t} = 0 \, \forall t$; which we shall denote as $\vec{\mathbf{1}}$ and $\vec{\mathbf{0}}$, respectively. If the history of treatments for each unit is known, it is in principle straightforward to estimate the cumulative direct effect. However the lack of identification means for each unit we can only observe the current treatment, and do not know the earlier treatments assigned to the unit. This makes most existing methods on estimating cumulative effect inapplicable to our problem. In the following section, we propose a method to infer cumulative effects from the observed data by estimating the relation between the contemporary and cumulative effects. Our method relies on some assumptions standard for treatment effect estimation: sequential ignorability (Imai et al., 2010), no contagion (Hudgens and Halloran, 2008) and positivity (Pearl, 2000). We refer the readers to the Appendix A for a detailed discussion of these assumptions.

[3]We write τ_{it} rather than $\tau_{it}(1,0)$ for simplicity and $\mathbb{E}_i$ represents averaging over units

3 METHOD

3.1 Intuition

To establish the intuition for our model, we describe a simple example in this section. Our general model is described in the next section. Consider the simple setting with 3 time periods, i.e., for a user i, we have 3 treatments ($\mathbf{Z}_i^{1:3} = Z_{i,1}, Z_{i,2}, Z_{i,3}$) and 3 outcomes ($\mathbf{Y}_i^{1:3} = Y_{i,1}, Y_{i,2}, Y_{i,3}$) over the time periods. Furthermore, let us assume that the outcome Y_{it} depends only on the current and previous treatment allocations of the corresponding unit, i.e., observed outcome is $Y_{it}(Z_{it}, Z_{i,t-1})$ and $Y_{it}(.,.) \perp\!\!\!\perp Z_{i,k\leq t-2}, Z_{j\neq i,.}$. For the purpose of this section, we will suppress the dependence on the user covariates X.

In the first two time periods the users retain cookies, and so we can observe – and, in an experimental setup, control– the treatments each user receives in the first two steps. Before the third visit, the cookie expires, and hence when the user visits the third time, we do not have any information about the treatment history. Moreover due to lack of information about the history, in the third period treatment allocation necessarily undergoes randomization. Since the periods 1 and 2 are standard sequential treatment problems, we can focus only on the treatment effect in the third period. Our goal is to estimate the treatment effect on the outcome in the third time-period $Y_{i,3}$ under a constant treatment, i.e., $\mathbb{E}[Y_{i,3}(1,1) - Y_{i,3}(0,0)]$.

Let p_1 denote the probability of assigning any user the treatment 1 in the second time period; and $\bar{Y}_i$ denote the observed outcome in the third period (i.e. $\bar{Y}_i = Y_{i,3}(Z_{i,3}, Z_{i,2})$). Observe that while $\bar{Y}_i$ depends on $Z_{i,2}$ we cannot observe $Z_{i,2}$. The naive treatment effect estimate applied on the post-cookie time period (which is measuring the direct effect of $Z_{i,3}$) gives:

$$\hat{\tau}_3 = \frac{\sum_{i=1}^{N} \bar{Y}_i \mathbb{I}[Z_{i,3} = 1]}{\sum_{i=1}^{N} \mathbb{I}[Z_{i,3} = 1]} - \frac{\sum_{i=1}^{N} \bar{Y}_i \mathbb{I}[Z_{i,3} = 0]}{\sum_{i=1}^{N} \mathbb{I}[Z_{i,3} = 0]}$$

Here $\mathbb{I}$ is the indicator function. The above expression is simply the difference in the average outcome between the treated ($Z_{.,3} = 1$) and untreated ($Z_{.,3} = 0$) group. If the observed outcome only depended on the recent treatment $Z_{.,3}$, then under sequential ignorability, the above estimate is unbiased for the treatment effect. However when the outcome depends on the history we get

$$\begin{aligned}
&\mathbb{E}[\hat{\tau}_3] = \mathbb{E}_{\mathbf{Z}^2}\mathbb{E}[Y_{i,3}(1, \mathbf{Z}^2)] - \mathbb{E}_{\mathbf{Z}^2}\mathbb{E}[Y_{i,3}(0, \mathbf{Z}^2)] \\
&= p_1\mathbb{E}[Y_3(1,1)] + p_0\mathbb{E}[Y_3(1,0)] - p_1\mathbb{E}[Y_3(0,1)] \\
&\quad - p_0\mathbb{E}[Y_3(0,0)] \\
&= p_1\mathbb{E}[(Y_3(1,1) - Y_3(0,1))] + p_0\mathbb{E}[(Y_3(1,0) - Y_3(0,0))] \\
&= 0.5\mathbb{E}[Y_3(1,1) - Y_3(0,0)] + 0.5\mathbb{E}[Y_3(1,0) - Y_3(0,1)] \quad (1)
\end{aligned}$$

where p_1, p_0 are the marginal probabilities of $Z_2 = 1, 0$, respectively; and we have suppressed i for notational convenience. Then in the last line, we have assumed perfect

randomization to set $p_1 = p_0 = 0.5$ On the other hand the desired estimand is $\mathbb{E}[(Y_3(1,1) - Y_3(0,0)]$.

Note that we have not used assumed anything specific about the form of the outcome and so the above expression is true without any other model assumption. A different outcome process can now lead to different treatment effects under the same observational data. For example, in the simple case that outcome for mixed treatments are the same i.e. $Y_3(1,0) = Y_3(0,1)$, then the above expression becomes $0.5(Y_3(1,1) - Y_3(0,0))$. One the other hand if the outcome depends only on the latest treatment i.e. $Y_3(Z_3, Z_2) = Y_3(Z_3)$, then the earlier estimate $(Y_3(1,1) - Y_3(0,0))$ equals the desired estimand. This shows that the cumulative treatment effect is unidentifiable from only the observed data in the case of temporal interference under identity fragmentation. This is not surprising given that different joint distributions can have the same marginal distributions (Koller and Friedman, 2009).

However, the above expression also suggests that if the outcomes under mixed/interfered treatments, i.e., $Y(1,0), Y(0,1)$ are related to $Y(\vec{1})$ and $Y(\vec{0})$ then we can in principle adjust the previous estimator τ_t to obtain the cumulative effect $\tau_t(\vec{1}, \vec{0})$ without any further experiments. For example if $Y_3(1,0) - Y_3(0,1) = \alpha(Y_3(1,1) - Y_3(0,0)$, then Equation 1 becomes

$$\mathbb{E}[\hat{\tau}_3] = 0.5(1+\alpha)[Y_3(1,1) - Y_3(0,0)] \tag{2}$$

More generally if we could connect the potential outcome on history Z to the potential outcomes $Y(\vec{1})$ and $Y(\vec{0})$, then the above expression can be rewritten to factor out the desired estimand $Y_3(1,1) - Y_3(0,0)$.

3.2 DIET

Building on the above insight, next we present the core assumption of our work which makes cumulative effect estimation in the current scenario feasible.

Diet Assumption 1 (Outcome Model)**.**

$$Y_{it}(\mathbf{Z}_i^{1:t}) - Y_{it}(\vec{\mathbf{0}}) = (Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}})) * c_i(\mathbf{Z}_i^{1:t}, X_i) \tag{3}$$

In this formulation, $c_i(\vec{\mathbf{0}}) = 0$ and $c_i(\vec{\mathbf{1}}) = 1$, for every unit. Note that c depends on the covariates X_i or be a random variable as well instead of a constant function. In that case, we assume that given the history of treatments (and any applicable covariates) $c_i(\mathbf{Z}_i^{1:t}, X_i) \perp\!\!\!\perp Y_{it}(\vec{\mathbf{0}}), Y_{it}(\vec{\mathbf{1}})$.

This model has some functional resemblance to the Cox Proportional Hazards model (Cox, 1972; Breslow, 1975; Clayton and Cuzick, 1985). However, unlike the Cox model, which is focused on survival analysis[4], we have a baseline individual treatment effect $(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))$ affected by the history of treatments by a relative-risk/proportionality factor c. Moreover, unlike the Cox model, the factor in our case can take any real value.

Theorem 1. *For the outcome model described in Assumption 1, cumulative treatment effects for any desired treatment history and marginal/direct treatment effects are related when conditioned on the covariates. Specifically,* $\forall \mathbf{z}, \mathbf{z}', \Lambda_t(\mathbf{z}, \mathbf{z}') := \dfrac{\tau_t(\mathbf{z}, \mathbf{z}')}{\tau_t}$ *is a function of only* X.

The key idea is that if the differences between the various potential outcomes are multiplicatively related, then so is the direct treatment effect to the cumulative treatment effect. More specifically similar to Equation 2, the expected cumulative treatment effect can be factored out from the expression of direct treatment effect. For a detailed proof of the theorem, we refer the reader to the AppendixB.1.

While Theorem 1 gives an existence result for the relation between the contemporary direct effect (which we can also call marginal treatment effect) and the global treatment effect, it may not be useful. More precisely, the exact value of the constant Λ appearing in Theorem 1 depends not only on the unknown risk coefficients c_i but also on the treatment allocation process which in turn depends on the cookie loss process. This creates a challenge in estimation of cumulative effects from direct effect without further assumptions.

However if one can estimate this dependence, consistent estimators of the cumulative effect can be constructed even in the presence of fragmentation. This is the key idea behind our method, to infer the cumulative effects from direct effect estimates. In Algorithm 1, we present our algorithm called DIET (Direct Inference of Effect of Treatment) for estimating the global treatment effect from fragmented data. With some restrictions on the data, our DIET algorithm can identify the aforementioned dependence between cumulative and direct effects from observational data. For this purpose, we make explicit our assumption about fragmentation and cookie-loss, which allow the aforementioned dependence to be identifiable.

Diet Assumption 2 (Limited Interference and Fragmentation)**.**

$$\begin{aligned}
&a)\ \exists H \qquad s.t. \qquad c_i(\mathbf{Z}_i^{1:t}) = c_i(\mathbf{Z}_i^{t-H:H}) \\
&b)\ \Pr\Big(\bigcap_{j=1}^{H+1} \Delta_{i,t+j} = 0 | \Delta_{i,t} = 1\Big) > \varepsilon > 0 \\
&c)\ \Delta_{i,t} \perp \mathbf{Z}_i^{1:t}, \mathbf{Y}_i^{1:T} | X_i
\end{aligned}$$

Here $\Delta_{i,t}$ represents the event that the cookie for unit i was lost just before time t.

[4]The Cox hazard model can be mathematically written as $\lambda(t|x) = \lambda_0(t)exp(\beta x)$ where λ_0 is the baseline hazard rate and $exp(\beta x)$ is the covariate dependent relative risk. Borrowing this term we would call our c function relative risk.

These assumptions means that a) the relative risk function c_i has dependence on only a finite number H of previous treatments, and not affected by an arbitrarily long history. This is natural in many settings such as offers and ads, where the effect of treatments decay with time. Furthermore, b) ensures that the fragmentation due to deletion/loss of cookies is limited so that some units can be tracked for more time steps than the limit H. This is not restrictive as some users are likely to accept cookies or use only one device. In other cases their attributes might be unique enough for stitching methods to be very accurate. Finally, c) means that the event of cookie loss is conditionally independent of treatments and outcomes. Both parts b) and c) are implicitly related to standard assumptions in causal inference literature (Hernan and Robins, 2013; Banerjee and Duflo, 2009; Pearl, 2000).

Algorithm 1 DIET procedure

Observed data $(X_i, Y_i^{1:T}, Z_i^{1:T})$, Horizon H
Estimate $\hat{\tau}_t(\vec{\mathbf{1}}, \vec{\mathbf{0}})$ of $\tau_t(\vec{\mathbf{1}}, \vec{\mathbf{0}})$
Procedure:

1. Estimate propensities $\pi = P(Z_{i,t}|\mathbf{Z}_i^{1:t-1}, \mathbf{Y}_i^{1:t-1}, X_i)$
2. Stratify observation based on covariates X_i
3. Use propensity model π and stratified observations to estimate conditional direct effect τ_t for users with history, i.e., $\hat{\tau}_t^D(X) = \hat{\mathbb{E}}_{X_i=X}\hat{\mathbb{E}}_{\mathbf{Z}^{t-H-1:t-1}\sim\pi}[\tau_{it}(1, 0; \mathbf{Z}^{t-H-1:t-1}|X)]$
4. Use propensity model π and stratified observations to compute conditional cumulative effect $\tau_t(\vec{\mathbf{1}}, \vec{\mathbf{0}})$ for users with history, i.e., $\hat{\tau}_t^C(X) = \hat{\mathbb{E}}_{X_i=X}\hat{\mathbb{E}}_{\mathbf{Z}^{t-H-1:t-1}\sim\pi}[\tau_{it}(\vec{\mathbf{1}}, \vec{\mathbf{0}}; \mathbf{Z}^{t-H-1:t-1}|X)]$
5. Compute estimates $\hat{\Lambda}_t(X) = \hat{\tau}_t^C(X)/\hat{\tau}_t^D(X)$
6. Learn function to estimate $\Lambda_t^f(X)$ from $\hat{\Lambda}_1(X), \hat{\Lambda}_2(X), \cdots, \hat{\Lambda}_t(X)$
7. Use propensity model π and stratified observations to estimate conditional marginal effects for users with history, i.e., $\hat{\tau}_t^g(X) = \hat{\mathbb{E}}_{X_i=X}\hat{\mathbb{E}}_{\mathbf{Z}^{t-H-1:t-1}\sim\pi}[\tau_{it}(1, 0; \mathbf{Z}^{t-H-1:t-1}|X)]$
8. Infer $\hat{\tau}_t(\vec{\mathbf{1}}, \vec{\mathbf{0}})$ as $\mathbb{E}_X[\hat{\tau}_t^g(X)\Lambda_t^f(X)]$

Theorem 2. *Under Assumptions [1-2], the procedure in Algorithm 1 gives a consistent estimator of cumulative treatment effects.*

We explain the idea behind the method and refer the readers to the AppendixB.2 for the full proof. With a high enough ε (or enough number of units), we will have some users with histories longer than H available. These can be then used to construct an estimate of the cumulative treatment effect (Step 4). Similarly, we can estimate the contemporary direct effect (marginal treatment effect) of the treatments from the data (Step 3). We can then put these two estimates together to learn the functional dependence between these two effects (Step 5 in Algorithm 1). These are all noisy estimates (primarily due to the noise of global effect estimates); however our procedure gives us access to multiple such estimates. These can then be fit together to learn the true Λ function (Step 6 in Algorithm 1). The learnt Λ can be used for inferring back the global effect. In the case of experimental designs where the treatment policy is already known to the marketer, Step 1 of the algorithm can be skipped. Lastly, our method is concerned with inferring the cumulative effect from the direct effect estimate. While we need a propensity model π as well as estimator for the effects τ^C, τ^D, our algorithm is agnostic to the choice of such estimators. In our experiments, we would use a vanilla Horvitz-Thompson estimator for measuring the effects and a logistic regression model for propensity estimation.

Special Case Lin and Misra (2021) assume a model with the outcome being linearly dependent on cumulative marketing expenditure. This situation is a special case of our model, where the relative function $c_i(\mathbf{Z}_i^{t-H:T}) = \frac{\sum_{t=1}^{H} Z_{it}}{H}$. Lin and Misra (2021) then show that the conducting analysis on cookie-level data leads to the treatment effect being attenuated by the "number of identities". Plugging these assumption about fragmentation and c_i into Equation 9 (Appendix), one can show that the direct treatment effect is a similarly attenuated value of the cumulative treatment effect (more discussion can be found in the Appendix). As a simple case, if we consider the example of Equation 1, and put $c_i(Z_{t-1}, Z_t) = \frac{Z_{t-1}+Z_t}{2}$, we can see that the observed direct treatment effect $= 0.5\times$(true cumulative effect); i.e., the effect is attenuated. However, unlike their model, we can handle an unknown and generic c function.

4 EXPERIMENTS

We perform three sets of experiments. First, we compare our proposal on a simulated dataset, and compare it to the true simulated treatment effect. Next, we compare our proposed method on a web analytics log dataset, by simulating different levels of fragmentation. Finally, we compare our method on a healthcare dataset with observational data.

Synthetic Simulation: For synthetic experiments, we generated data for $50,000$ users for 10 time steps following the outcome model in Equation equation 3. We set the history window $H = 3$. For the hazard function $c_i(Z)$ we chose for each user i a vector $w_c^i \in [0, 1]^3$, and set $c_i(Z) = f(w_c^i \cdot Z)$. The base outcomes $Y_{it}(\vec{\mathbf{1}}), Y_{it}(\vec{\mathbf{0}})$ were obtained from user covariates in a similar fashion, i.e., $Y_{it}(\vec{\mathbf{1}}) = w_1^i \cdot X_{it}$. We first compute the ratio Λ of the 'marginal' treatment effect and the treatment effect for history $\mathbf{z}$, i.e., $\mathbb{E}[Y_t(\mathbf{z}) - Y_t(\vec{\mathbf{0}})]/\mathbb{E}[Y_t(1) - Y_t(0)]$ for different histories and at each time step. We present results on

three types of treatment allocations labeled as *Late* (which corresponds to treatment at only the last time step and 0 before that, i.e., $\mathbf{Z}_i^{1:H} = [0, \ldots, 0, 1]$), *Alter* (which alternates between treatments, i.e., $\mathbf{Z}_i^{1:H} = [0, 1, 0, 1, .., 0, 1]$) and *Complete* (which is a treatment allocation of only treatment 1, i.e., $\mathbf{Z}_i^{1:H} = [1, 1, \ldots, 1]$). In Figure 1 we provide a bar graph of the point estimate and the standard deviation of $\Lambda(\mathbf{Z})$ for the three different histories $\mathbf{Z}$. The consistency of Λ is neither limited to specific monotonic c or positive effects; we experimented with different risk ratios (quadratic and sinusoidal) and found similar results. For example in Figure 1(c) we plot the value of $\Lambda(H)$ for a more complex sinusoidal relative risk function.

Equation 9 shows that if c_i do not depend on time, neither does Λ. We experimentally show this in the Appendix we experimentally show that this is true, where we plot the function Λ over time (Figure 7a). Quantitatively the deviation of the estimate is of order 0.01 against the estimate value of order 1, and fluctuating around its mean with no visible trend, demonstrating that time doesnt have any impact on Λ.

Online Sales: We evaluate our estimator on a cookie-less A/B testing scenario using observational data. For this purpose we use a historical click-stream data made available by YooChoose. This dataset (Ben-Shimon et al., 2015) is the click event sessions for a major European e-tailer collected over 6 months. The logs consist of a variable length history of the sessions and includes data about clicks and purchases, along with other information like offers, product type etc. For the outcome, we considered whether an event led to a purchase, and for the treatment we considered whether the event happened in context of a promotion or special offer. In our experiments we used two versions of DIET: Constant and Linear (based on the fitting model used in Step 6 of Algorithm 1)

We first analyse whether incorporating previous treatments affects the treatment effect estimate. If history has little or no effect on the outcomes, there is no problem in using the naive treatment estimate (i.e., the marginal effect). For this purpose, we experiment with different horizon lengths H (time steps), estimate the treatment effect knowing the entire history and compare it against both the marginal effect (MTE), and two versions of our DIET estimator. In Figure 2, we plot the bias of these estimates (along with the standard error) over different sub-samples of the data. We see that the marginal effect is sufficiently biased once the history is taken into account. Furthermore, since conversion rates in e-commerce is around 0.5-2%; the bias of the MTE is of the same order as the treatment effect. As such, in this application using the MTE instead of the correct estimator can lead to significant error ($\sim 100\%$). On the other hand our DIET estimator though having slightly higher variance reduces the bias considerably.

The previous experiment is in a setting where all history is available. Furthermore, it is a situation where DIET has the greatest advantage in terms of availability of long histories. Hence, we next analyse the impact of cookie deletion on effect estimation. We randomly select a set of users to retain history, while the other users delete their cookie every time. In these experiments we set the dependence horizon to 3. We vary the percentage of users for whom we allow access to histories and compare performance of different estimators. Since some fraction of users always retain cookies, a natural baseline is to use traditional full history based methods restricted to this smaller dataset of tracked users. Another solution relies on stitching to connect the fragmented identities and using the enhanced dataset with full history methods. We implement these as alternative baselines and use LTMLE (Stitelman et al., 2012; Schomaker et al., 2019) to compute treatment effects. LTMLE is G-computation (Robins and Hernán, 2009) based extension of the Targeted Maximum Likelihood Estimation/TMLE approach of van der Laan (2010) which is knwon to be (semi-parametrically) asymptotically efficient

We measure the mean squared error between estimated effect by different estimators and an oracle estimator with access to all the histories for all the users. Figu The results are plotted in Figure 4, where we can see that as fragmentation reduces the MSE of all predictors improve. LTMLE is known to asymptotoically efficient, and so it is not surprising that it can achieve lowest error. However, when the number of users with known history reduces, DIET starts outperforming LTMLE. This is because LTMLE cannot use fragmented records, whereas DIET can leverage these fragmented records. We also see a consistent improvement (upto 20%) over a wide range of fragmented identities by using DIET.

We also plot the bias of these baselines as the fraction of users with history change. This is plotted in Figure 3. While the bias of LTMLE as well as Linear DIET is small, the bias of stitching based baseline is both high and varies unpredictably as the percentage of fragmented identities changes. This is consistent with our earlier proposition that stitching based methods have uncontrollable bias. Furthermore while LTMLE remains unbiased, the reduction in history availability increases the variability of the estimate.

MIMIC 3: We perform experiments on a healthcare based application using the Medical Information Mart for Intensive Care (MIMIC-III) database (Johnson et al., 2016). The dataset consists of trajectories of clinical measurements (e.g., heart rate, respiratory rate), assigned treatments (vasodilators, antibiotics and so on) as well as other covariates of ICU patients. For our experiment we focus on the subset of patients in MIMIC whose primary diagnosis is sepsis. Since the are many potential treatment drugs (and their combinations), we categorized them by types and then

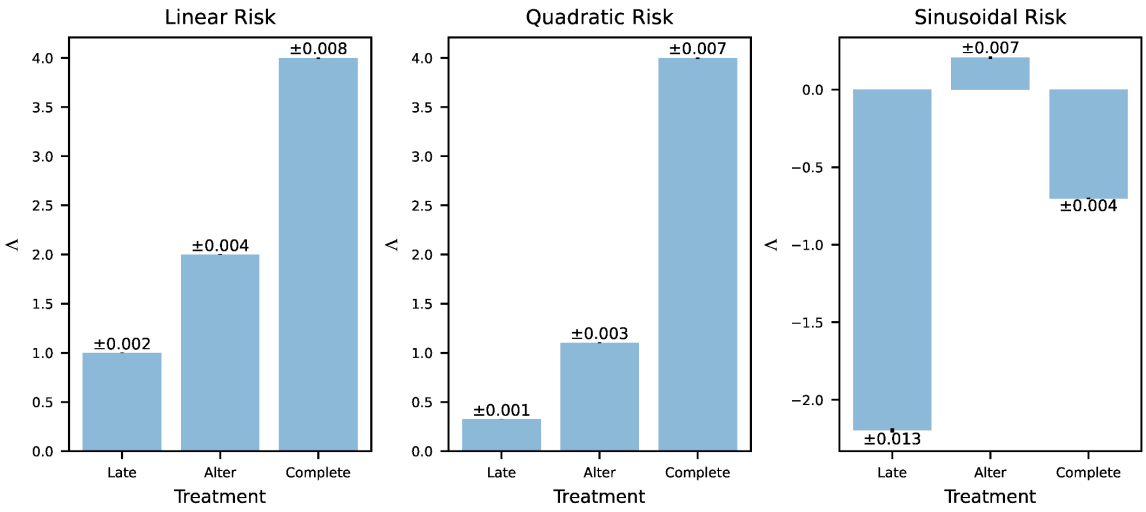


Figure 1: Bar chart of population Λ function (i.e. the ratio of cumulative effect to direct effect) for various treatment policies (Late,Alter, Complete). Note that this is the population level average function. Different risk coefficients/hazard models a) Linear, b) Quadratic and c) Sinusoidal have been plotted in different figures. The standard deviation has reported on top of bars.

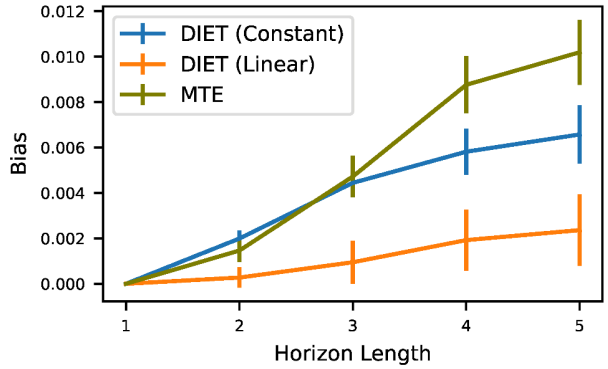


Figure 2: Bias (with standard error across trials) of different estimators as the time horizon of interference is increased on YooChoose click-stream logs.

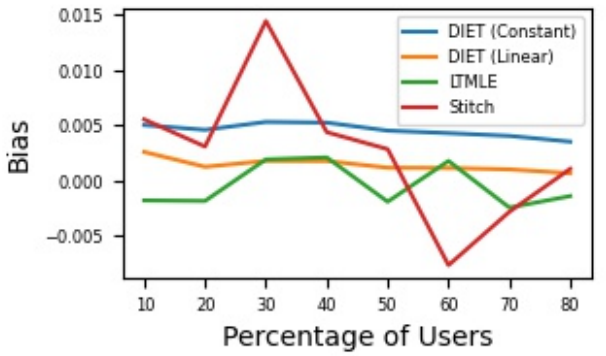


Figure 3: Bias of different estimators with H=3 as percentage of users with unfragmented identity varies on YooChoose click-stream logs. Stitching tends has an unpredictable bias while LTMLE is unbiased.

further binned them into binary labels. The goal once again is to assess the cumulative treatment effect of a regime providing treatment 1 over treatment 0. We asses various outcomes such as blood glucose level, oxygenation and white blood cell counts.

We repeat the experiments we performed for the YooChoose data. We first demonstrate temporal interference/history dependence on the outcomes. These results can be seen in Figure6 in the Appendix. In this case, the bias of marginal estimate is significantly higher than Linear DIET. We also observe smaller improvement over MTE with DIET constant. This is likely because the model assumptions do not hold in this case. For drug treatments, different dose regimens can induce unpredictable effects violating the outcome model assumption (Martinez et al., 2012). Moreover, as medical practitioners do not randomize prescriptions and base their decision on a variety of factors; observational medical datasets almost surely violate assumptions of positivity and lack of confounders. Next we assess the error of the estimator as the percentage of users with available history varies. We see greater variance and difference between the two different DIET estimators in these experiments. However the general trend of lower error compared to stitching and lesser variance compared to LTMLE still holds. We see that at low history availability DIET can reduce the error from LTMLE and stitching by over 30%. Moreover DIET dominates stitching almost everywhere and is competitive with LTMLE at higher data availability as well.

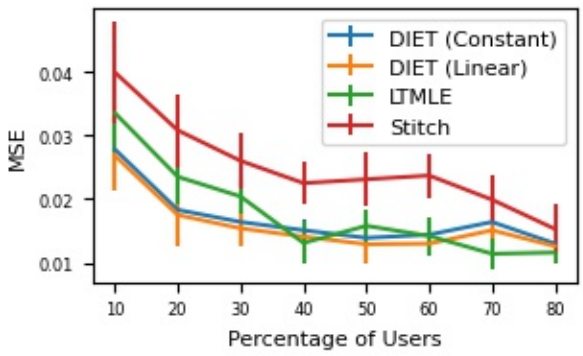


Figure 4: MSE (with standard error across trials) of estimators as percentage of users with unfragmented identity varies on YooChoose click-stream logs. Generally LTMLE has lower error then DIET with high history availability, but is worse than DIET when histories are unavailable. Stitching on the other hand tends to always have higher error. In both cases, with less history the variability of existing methods is higher than DIET.

5 RELATED WORKS

Stitching without Cookies: Considerable research has gone into stitching fragmented user behaviour (Saha Roy et al., 2015; Kim et al., 2017; Jin et al., 2019; De Smedt et al., 2021). But these strategies rely on using features of devices (e.g., IP Address) to predict which pairs of devices represent the same human user. In an atmosphere of increased privacy sensitivity, such strategies lead to further mistrust of organizations, for instance, the GDPR explicitly forbids the use of IP address as a feature for digital marketing. There are efforts within the industry to overcome the absence of cookies, such as an approach called "Topics" [5]. These are however focused on ad targeting and outreach towards users. On the other hand, our goal is to assess the impact of a given long-term experience or treatment.

Hazard models: The Cox model (Cox, 1972) and similar proportional hazards models (Chen et al., 2010; Cao et al., 2015) have a rich history in survival analysis. In the context of online marketing, the effect of marketing interventions on screen time (Barbieri et al., 2016), return times (Kapoor et al., 2014), time to opening emails (Sinha et al., 2018), and long-term crowd engagement (Chandar et al., 2022; Gu et al., 2022) have been explored with survival models. However, none of these address the question of estimating the treatment effect under identity fragmentation.

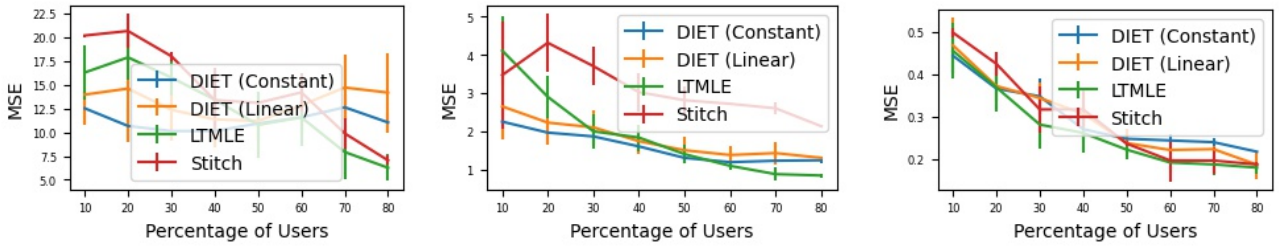
Fragmentation bias: Weaknesses of using cookie-level data against individual-level data is known in literature (Chatterjee et al., 2003; Bleier and Eisenbeiss, 2015; Hoban and Bucklin, 2015); but there is not a lot of work addressing these from a formal perspective. Some approaches used for parameter estimation from observational user logs include missing data imputation (Novak et al., 2015) and aggregation (Rutz et al., 2011). Taylor and Eckles (2018) have suggested to focus on an ITT like setting for assessing network influence. Koehler et al. (2013, 2016) uses a combination of server logs, publisher provided data (PPD), and public data to measure the reach of of online ad campaigns when enough tracking information is unavailable. However these methods are non-causal or user preference models and are inapplicable for treatment effect estimation.

Earlier work most related to the current work is by Coey and Bailey (2016) on bias caused by fragmentation when using cookie-level data. Lin and Misra (2021) characterize fragmentation bias in linear models. While our problem is also caused due to using cookie-level data, unlike these works our focus is on a temporal setting where historical treatments effect the current outcome. We also show that their results are a corollary of our results.

Interference: There have been many attempts to deal with interference in the literature (Hudgens and Halloran, 2008; Blackwell and Glynn, 2018). But these assume strong restrictions on the structure of spillover. For inter-unit interference spatial models (Beck et al., 2006; LeSage and Pace, 2009), and network models (Graham, 2008; Acemoglu et al., 2015; Leung, 2020) are commonly used. Recently some work has focused on how to account for general interference (Papadogeorgou et al., 2020; Zigler and Papadogeorgou, 2018; Ogburn et al., 2020) . Yet these works concentrate on creating experimental designs rather than the analysis of observational studies (Savje et al., 2018; Aronow et al., 2019; Chin, 2019). Furthermore these techniques are designed primarily for spatial interference. Recently, Shankar et al. (2022) devise a modified experimentation procedure to perform A/B testing when user identities are fragmented among different brands or related channels. However there approach does not consider temporal interference and is restricted to cooperating channels.

Dynamic and Sequential Treatments: Pioneering work by Robins (1986) leds to development of Structural Nested Models (SNM) and Marginal Structural Models (MSM) (Robins and Hernán, 2009; Robins et al., 2000; Hernán et al., 2000) which can be used to obtain unbiased estimates of the cumulative effects for sequential and dynamic treatments (Blackwell, 2013). Extensions of these ideas have been proposed for incorporating deep neural networks (Lim, 2018; Lin et al., 2021; Bica et al., 2020; Melnychuk et al., 2022; Frauen et al., 2022; Li et al., 2021). However none of these methods can work without knowledge of the entire history of treatments, and hence cannot be used in the presence of identity fragmentation. While difference in difference (DID) methods have also been proposed to do estimation for longitudinal studies (Goodman-Bacon, 2018; Callaway and Sant'Anna, 2020), they focus

[5] https://bit.ly/3w1g8ak



(a) Glucose (b) White Blood Cell (c) Haemoglobin

Figure 5: Mean Squared Error (with standard error of trials) against the percentage patients with available history on sepsis patients from MIMIC for different outcome measures. MSE is measured between prediction from different estimators and true effect estimated from data. Generally LTMLE has lower error then DIET with high history availability, but is worse than DIET when histories are unavailable. Stitching on the other hand tends to always have higher error and unpredictable bias. In both cases, with less history the variability of existing methods is higher than DIET.

on a single known treatment allocation with uncertain timing. Furthermore these methods are known to be biased under heterogenous treatment (Wang, 2021; Sun and Abraham, 2020; De Chaisemartin and d'Haultfoeuille, 2020). Our work instead focuses on multiple treatment without fragmented user idenitities. To the best of our knowledge we are the first to address treatment effect estimation under temporal fragmentation.

6 CONCLUSION

We consider treatment effect estimation under in presence of identity fragmentation for sequential treatments. This setting naturally arises in the online businesses due to 1) users using multiple devices or 2) depreciation of cookies. To the best of our knowledge we are the first to address this problem . We prove that under multiplicative model cumulative treatment effects are functionally related to direct treatment effect. We then provide a consistent estimator under such a model and verify these results by experimenting on simulated data. We also use our proposed DIET method on real data and find that it can lead to upto 20% reduction in error.

Our paper opens the door to more research in this area, which can have significant social impact as well. Estimates suggest that record fragmentation in healthcare systems imposes additional costs upto $1,950 per patient in the US (Jason, 2020). Hence research in this area also serves an important social good in terms of reducing healthcare costs.

References

Acemoglu, D., García-Jimeno, C., and Robinson, J. A. (2015). State capacity and economic development: A network approach. *American Economic Review*, 105(8):2364–2409.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Aronow, P. M., Samii, C., and Wang, Y. (2019). Design based inference for spatial experiments.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424.

Banerjee, A. V. and Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, 1:151–178.

Barbieri, N., Silvestri, F., and Lalmas, M. (2016). Improving post-click user engagement on native ads via survival analysis. In *Proceedings of the 25th International Conference on World Wide Web*, pages 761–770.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. In *Selected Works of Debabrata Basu*, pages 167–206. Springer.

Basu, D. (1980). Randomization and the analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):575–582.

Beck, N., Gleditsch, K. S., and Beardsley, K. (2006). Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly*, 50:27–44.

Ben-Shimon, D., Tsikinovsky, A., Friedmann, M., Shapira, B., Rokach, L., and Hoerle, J. (2015). Recsys challenge 2015 and the yoochoose dataset. RecSys '15.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. (2020). Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*.

Blackwell, M. (2013). A framework for dynamic causal inference in political science. *American Journal of Political Science*, 57(2):504–520.

Blackwell, M. and Glynn, A. N. (2018). How to make causal inferences with time-series cross-sectional data under selection on observables. *American Political Science Review*, 112(4):1067–1082.

Bleier, A. and Eisenbeiss, M. (2015). Personalized online advertising effectiveness: The interplay of what, when, and where. *Marketing Science*, 34(5):669–688.

Bohn, D. (2020). Google to 'phase out' third-party cookies in chrome.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57.

Callaway, B. and Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Cao, H., Churpek, M. M., Zeng, D., and Fine, J. P. (2015). Analysis of the proportional hazards model with sparse longitudinal covariates. *Journal of the American Statistical Association*, 110(511):1187–1196.

Chandar, P., St. Thomas, B., Maystre, L., Pappu, V., Sanchis-Ojeda, R., Wu, T., Carterette, B., Lalmas, M., and Jebara, T. (2022). Using survival models to estimate user engagement in online experiments. In *Proceedings of the ACM Web Conference 2022*, pages 3186–3195.

Chatterjee, P., Hoffman, D. L., and Novak, T. P. (2003). Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541.

Chen, K., Guo, S., Sun, L., and Wang, J.-L. (2010). Global partial likelihood for nonparametric proportional hazards models. *Journal of the American Statistical Association*, 105(490):750–760.

Chin, A. (2019). Central limit theorems via stein's method for randomized experiments under interference. *arXiv:1804.03105 [math.ST]*.

Churches, T., Christen, P., Lim, K., and Zhu, J. X. (2002). Preparation of name and address data for record linkage using hidden markov models. *BMC Medical Informatics and Decision Making*, 2(1):1–16.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A (General)*, 148(2):82–108.

Coey, D. and Bailey, M. (2016). People and cookies: Imperfect treatment assignment in online experiments. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1103–1111.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Davey Smith, G. and Ebrahim, S. (2004). Mendelian randomization: Prospects. *Potentials, and*.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

De Smedt, J., Lacka, E., Nita, S., Kohls, H.-H., and Paton, R. (2021). Session stitching using sequence fingerprinting for web page visits. *Decision Support Systems*, 150:113579.

Durbin, J. (1953). A note on regression when there is extraneous information about one of the coefficients. *Journal of the American Statistical Association*, 48(264):799–808.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Frauen, D., Hatt, T., Melnychuk, V., and Feuerriegel, S. (2022). Estimating average causal effects from patient trajectories. *arXiv preprint arXiv:2203.01228*.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica*, 76(3):643–660.

Gu, Z., Bapna, R., Chan, J., and Gupta, A. (2022). Measuring the impact of crowdsourcing features on mobile app user engagement and retention: A randomized field experiment. *Management Science*, 68(2):1297–1329.

Hajek, J. (1971). Comment on 'An essay on the logical foundations of survey sampling, Part one.'. In Godambe, V. and Sprott, D., editors, *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pages 561–570.

Hernan, M. A. and Robins, J. M. (2013). *Causal Inference*. Chapman and Hall/CRC, Boca Raton, FL.

Hoban, P. R. and Bucklin, R. E. (2015). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*, 52(3):375–393.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71.

Jason, C. (2020). 3 consequences of patient matching, health record issues.

Jin, D., Heimann, M., Rossi, R. A., and Koutra, D. (2019). Node2bits: Compact time-and attribute-aware node representations for user stitching. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 483–506. Springer.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kapoor, K., Sun, M., Srivastava, J., and Ye, T. (2014). A hazard based approach to user return time prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1719–1728.

Katz, D., Baptista, J., Azen, S., and Pike, M. (1978). Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics*, pages 469–474.

Kim, S., Kini, N., Pujara, J., Koh, E., and Getoor, L. (2017). Probabilistic visitor stitching on cross-device web logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1581–1589.

Koehler, J., Skvortsov, E., Ma, S., and Liu, S. (2016). Measuring cross-device online audiences.

Koehler, J., Skvortsov, E., and Vos, W. (2013). A method for measuring online audiences.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

Leung, M. P. (2020). Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2):368–380.

Li, R., Hu, S., Lu, M., Utsumi, Y., Chakraborty, P., Sow, D. M., Madan, P., Li, J., Ghalwash, M., Shahn, Z., et al. (2021). G-net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Machine Learning for Health*, pages 282–299. PMLR.

Lim, B. (2018). Forecasting treatment responses over time using recurrent marginal structural networks. *advances in neural information processing systems*, 31.

Lin, L., Sperrin, M., Jenkins, D. A., Martin, G. P., and Peek, N. (2021). A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagnostic and prognostic research*, 5(1):1–16.

Lin, T. and Misra, S. (2021). The identity fragmentation bias. *Available at SSRN 3507185*.

Martinez, M. N., Papich, M. G., and Drusano, G. L. (2012). Dosing regimen matters: the importance of early intervention and rapid attainment of the pharmacokinetic/pharmacodynamic target. *Antimicrobial agents and chemotherapy*, 56(6):2795–2805.

Melnychuk, V., Frauen, D., and Feuerriegel, S. (2022). Causal transformer for estimating counterfactual outcomes. *arXiv preprint arXiv:2204.07258*.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. *Statistical Science*, 5:465–80. Section 9 (translated in 1990).

Novak, J., Feit, E. M., Jensen, S., and Bradlow, E. (2015). Bayesian imputation for anonymous visits in crm data. *Available at SSRN 2700347*.

Ogburn, E. L., Sofrygin, O., Diaz, I., and van der Laan, M. J. (2020). Causal inference for social network data. *arXiv preprint arXiv:1705.08527*.

Ogburn, E. L. and VanderWeele, T. J. (2014). Causal diagrams for interference. *Statistical science*, 29(4):559–578.

Papadogeorgou, G., Imai, K., Lyall, J., and Li, F. (2020). Causal inference with spatio-temporal data: Estimating the effects of airstrikes on insurgent violence in iraq. *arXiv preprint arXiv:2003.13555*.

Pearl, J. (2009). Remarks on the method of propensity score.

Pearl, J. a. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period?application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J. M. and Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. *Longitudinal data analysis*, 553:599.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1990). Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3):279–292.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rubin, D. B. (2010). Propensity score methods. *American journal of ophthalmology*, 149(1):7–9.

Ruggles, S., Fitch, C. A., and Roberts, E. (2018). Historical census record linkage. *Annual review of sociology*, 44:19–37.

Rutz, O. J., Trusov, M., and Bucklin, R. E. (2011). Modeling indirect effects of paid search advertising: Which keywords lead to more future visits? *Marketing Science*, 30(4):646–665.

Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., and Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: statistical approaches and design issues. *Psychological methods*, 19(3):317.

Saha Roy, R., Sinha, R., Chhaya, N., and Saini, S. (2015). Probabilistic deduplication of anonymous web traffic. In *Proceedings of the 24th International Conference on World Wide Web*, pages 103–104.

Savje, F., Aronow, P. M., and Hudgens, M. G. (2018). Average treatment effects in the presence of unknown interference. *arXiv:1711.06399 [math.ST]*.

Schiff, A. (2020). Apple wwdc 2020: A version of intelligent tracking prevention is coming to the app world.

Schomaker, M., Luque-Fernandez, M. A., Leroy, V., and Davies, M.-A. (2019). Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Statistics in medicine*, 38(24):4888–4911.

Seufert, E. (2020). What does google's deprecation of gaid look like?

Shankar, S., Sinha, R., Mitra, S., Swaminathan, V., Mahadevan, S., and Sinha, M. (2022). Privacy aware experiments without cookies. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM 23)*.

Shiba, K. and Kawahara, T. (2021). Using propensity scores for causal inference: pitfalls and tips. *Journal of epidemiology*, page JE20210145.

Sinha, M., Vinay, V., and Singh, H. (2018). Modeling time to open of emails with a latent state for user engagement level. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 531–539.

Sinha, R., Saini, S., and Anadhavelu, N. (2014). Estimating the incremental effects of interactions for marketing attribution. In *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*, pages 1–6. IEEE.

Stitelman, O. M., De Gruttola, V., and van der Laan, M. J. (2012). A general implementation of tmle for longitudinal data applied to causal inference in survival analysis. *The international journal of biostatistics*, 8(1).

Sun, L. and Abraham, S. (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

Taylor, S. J. and Eckles, D. (2018). Randomized experiments to detect and estimate social influence in networks. *Complex spreading phenomena in social systems*, pages 289–322.

Theil, H. and Goldberger, A. S. (1992). On pure and mixed statistical estimation in economics. In *Henri Theil's Contributions to Economics and Econometrics*, pages 317–332. Springer.

van der Laan, M. J. (2010). Targeted maximum likelihood based causal inference: Part i. *The international journal of biostatistics*, 6(2).

Wang, Y. (2021). Causal inference under temporal and spatial interference. *arXiv preprint arXiv:2106.15074*.

Zigler, C. M. and Papadogeorgou, G. (2018). Bipartite causal inference with interference. *arXiv:1807.08660 [stat.ME]*.

(Supplementary Material)

A Assumptions

Before proceeding further to the outcome model which would allow us to estimate the desired treatment effects, we introduce some other critical assumptions. Note that these assumptions are either self-evidently true or common in causal inference literature (and not specific to our method).

Assumption 1 (No reverse causality)**.** *If* $\mathbf{Z}^{1:t} = \tilde{\mathbf{Z}}^{1:t}$*, then*

$$Y_{it}(\mathbf{Z}^{1:T}) = Y_{it}(\tilde{\mathbf{Z}}^{1:T}).$$

for any i *and* t*.*

This assumption requires that the potential outcome of any unit i at period t is not affected by treatments assigned in the future. While there is no physical mechanism which future to affect the past outcomes, if information about future treatments is available in advance units can adjust their behavior, leading to reverse causal influences. Another possible scenario in which this can happen is if observations are recorded by distributed or asynchronous systems, or simply measurement errors. In such a case the time stamps need not form a strict order and the logs might show reverse causality. Finally in application of biostatistics (Davey Smith and Ebrahim, 2004), this can happend due to violations of exchangeability(Hernan and Robins, 2013). Assumption 1 precludes any such possibilities and allows us to write $Y_{it}(\mathbf{Z}^{1:T})$ as $Y_{it}(\mathbf{Z}^{1:t})$[6].

Assumption 2 (Sequential ignorability)**.**

$$\mathbf{Z}_t \perp Y_{it}(\mathbf{Z}_t, \mathbf{Z}^{1:T} \setminus \mathbf{Z}_t) | \mathbf{Z}^{1:(t-1)}, \mathbf{Y}^{1:(t-1)}, \mathbf{X}^{1:t},$$

$$\mathbf{Z}_1 \perp Y_{i1}(\mathbf{Z}_1, \mathbf{Z}^{1:T} \setminus \mathbf{Z}_1) | \mathbf{X}^{1},$$

This assumption introduced by Imai et al. (2010) is a crucial and standard assumption for analysis of sequential treatments. This assumption means that given the observed pre-treatment variables, the treatment assignment is statistically independent of potential outcomes. It also implies that any mediators are also independent of outcomes given all observed variables. This means that the treatment assignment in period t, $\mathbf{Z}_t$, is result of perfect randomization conditional on past treatment assignments, past outcomes, and covariates that are not affected by $\mathbf{Z}_t$. Unlike the assumption of strict exogeneity in fixed effects models (Blackwell, 2013), sequential ignorability disallows any unobservable confounder. Therefore, if both the outcome and the assignment process are affected by some unobservable variables (e.g., unit fixed effects), the assumption will no longer hold.

Under Assumption 2 one can can use the information contained in the history to estimate the propensity scores at period t, $P(\mathbf{Z}_t = \mathbf{z}_t | \mathbf{Z}^{1:(t-1)}, \mathbf{Y}^{1:(t-1)}, \mathbf{X}^{1:t})$, which play a key role in the identification of the estimands (Imai et al., 2011, 2010).

Assumption 3 (Bernoulli design and Positivity)**.** *In any period* t*,* Z_{it} *is independent to each other for any* i*. Moreover* $0 < P(\mathbf{Z}_i^{1:t} = \mathbf{z}^{1:t}) < 1$ *for any* i *and* t*.*

We impose the common requirement of positivity (also known as overlapping) which is that each possible history for unit i should have a positive probability to occur on its support (Pearl, 2000). The roots of this assumption can be traced to seminal works of Basu (1971) and Horvitz and Thompson (1952). The bernoulli design assumption essentially is about randomization across units. It says Z_{it} can be dependent on the history, $\mathbf{Z}^{1:(t-1)}$, but not on the treatment status of other units in the same period, $\mathbf{Z}_t \setminus Z_{it}$. While this is implicitly assumed in essentially all causal literature (Austin, 2011; Rubin, 2010; Pearl, 2009) we make this explicit. For binary treatments, both positivity and bernoulli design, is can be ensured by ensuring that there is some randomization at each time step i.e. $0 < P(Z_{it} = z) < 1$.

Assumption 4 (No contagion)**.** *For any* i *and* t*, the probability* $P(Z_{it} = z)$ *is decided only by unit* i*'s own history.*

No contagion (Hudgens and Halloran, 2008; Ogburn and VanderWeele, 2014) means, treatment assigned to other units $\mathbf{Z}_j^{1:(t-1)}$ can not affect Z_{it}. In other words, the set of confounders excludes $\mathbf{Z}_j^{1:(t-1)}$, but may include $\mathbf{Z}_i^{1:(t-1)}$ or $\mathbf{Y}_i^{1:(t-1)}$. This is a version of the SUTVA assumption in literature (Rubin, 1990; Angrist et al., 1996; Rubin, 2005) via $\mathbf{Y}_i^{1:(t-1)}$. This allows us to focus only on the unit's history and not consider inter-unit interference.

[6] This is a structural assumption on the form of the outcome rather than the assignment process.

All of these assumption are common in causal inference literature (Robins et al., 2000; Blackwell, 2013; Pearl, 2000) and are in general necessary for any treatment effect estimation method to be function. Next we briefly state the specific assumptions needed by our approach.

A.1 DIET Assumptions

Assumption 5 (Outcome Model)**.**

$$Y_{it}(\mathbf{Z}_i^{1:t}) - Y_{it}(\vec{\mathbf{0}}) = (Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}})) * c_i(\mathbf{Z}_i^{1:t}) \tag{4}$$

Moreover if c_i are random functions, then c_i are conditonally independent given the covariates.

Assumption 6 (Limited Interference and Fragmentation)**.**

$$\begin{aligned}
&a) \exists H \qquad s.t. \qquad c_i(\mathbf{Z}_i^{1:t}) = c_i(\mathbf{Z}_i^{t-H:H}) \\
&b) \Pr(\bigcap_{j=1}^{H+1} \Delta_{i,t+j} = 0 | \Delta_{i,t} = 1) > \varepsilon > 0 \\
&c) \Delta_{i,t} \perp \mathbf{Z}_i^{1:t}, \mathbf{Y}_i^{1:T} | X_i
\end{aligned}$$

Parts b) and c) of the Assumption 6 assumption are implicit in the Assumptions 2 and 3. However, we make this explicit because deletion of cookies is an exogenous events and not directly observed, while other variables are observed.

Sequential ignorability implies there is no unmeasured confounding (Imai et al., 2011, 2010; Hernan and Robins, 2013). Since the event of cookie deletion leads to randomization in treatments (hence making these dependent), if cookie deletion is related to outcomes we effectively have a confounder. Part c) makes this explicit by making cookie deletion and outcomes independent. Similarly positivity implies that all possible treatment histories should be possible. If we include cookie deletion as exogenous treatments, treatment allocations where cookies are not deleted for atleast H time steps should have non zero probability. This is exactly what part b) specifies.

B Proofs

B.1 Proof for Theorem 1

We focus simply on the cumulative effect at time t as the direct effect at time t is the cumulative effect of treatment at time t.

$$\begin{aligned}
\tau_t(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t}) &= \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}} \left[\tau_{it}(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t}; \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}) \right] \\
&= \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}} \left[Y_{it}(\mathbf{z}^{s:t}, \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}) - Y_{it}(\tilde{\mathbf{z}}^{s:t}, \mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}) \right] \\
&\overset{(a)}{=} \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:T} \setminus \mathbf{Z}_i^{s:t}} \left[Y_{it}(\mathbf{z}^{s:t}, \mathbf{Z_i}^{1:T} \setminus \mathbf{Z}_i^{s:t}) - Y_{it}(\tilde{\mathbf{z}}^{s:t}, \mathbf{Z_i}^{1:T} \setminus \mathbf{Z}_i^{s:t}) \right] \\
&\overset{(b)}{=} \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:t} \setminus \mathbf{Z}_i^{s:t}} \left[Y_{it}(\mathbf{z}^{s:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t}) - Y_{it}(\tilde{\mathbf{z}}^{s:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t}) \right] \\
&\overset{(c)}{=} \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:t} \setminus \mathbf{Z}_i^{s:t}} \left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))(c_i(\mathbf{z}^{s:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t}) - c_i(\tilde{\mathbf{z}}^{s:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t})) \right] \\
&\overset{(d)}{=} \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:t} \setminus \mathbf{Z}_i^{s:t}} \left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))(c_i(\mathbf{z}^{t-H:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t}) - c_i(\tilde{\mathbf{z}}^{t-H:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t})) \right] \\
&\overset{(e)}{=} \mathbb{E}_i \mathbb{E}_{\mathbf{Z}^{1:t} \setminus \mathbf{Z}_i^{s:t}} \left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}})) \right] \mathbb{E}_{\mathbf{Z}^{1:t} \setminus \mathbf{Z}_i^{s:t}} \left[(c_i(\mathbf{z}^{t-H:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t}) - c_i(\tilde{\mathbf{z}}^{t-H:t}, \mathbf{Z_i}^{1:t} \setminus \mathbf{Z}_i^{s:t})) \right] \\
&\overset{(f)}{=} \mathbb{E}_i \mathbb{E} \left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}})) \right] \mathbb{E}_{P(\mathbf{Z})} C(\mathbf{z}^{t-H:t}, \tilde{\mathbf{z}}^{t-H:t}) \overset{(g)}{=} \mathbb{E}_i \mathbb{E} \left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}})) \right] \mathbb{E}_{\pi} C(\mathbf{z}^{t-H:t}, \tilde{\mathbf{z}}^{t-H:t}, P(\mathbf{Z}))
\end{aligned} \tag{5}$$

In the above derivation at (a) we used Assumption 4, to restrict the dependence only on $\mathbf{Z}_i$ instead of all treatments. Next in (b) we use Assumption 1 to drop any random variable of the future i.e. with time step more than t. Next in (c) we

use Assumption 5 and simple algebra to factor out $Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}})$. Next in (e) we use Assumption 2 which provides conditional independence to seperate the two expectations. In (f), we write the expectation over the c_i as a functional over the treatments $\mathbf{z}, \tilde{\mathbf{z}}$, and over the distribution induced by the treatment allocations. Finally in (g) we rewrite the expectation over the treatment allocation policy as an expectation over some behaviour policy π analogous to propensity weighting (Horvitz and Thompson, 1952; Hajek, 1971). This transformation requires knowing $P(Z)$ as well as full support over all possible allocations. Assumption 3 ensures positivity as well as gives us a way to compute the probability of a treatment allocation. For more discussion on the admissibility of such change in expectation/sampling we refer the readers to works of Basu (1971, 1980); Pearl (2000); Precup (2000); Shiba and Kawahara (2021).

Similarly, for the direct effect, we have

$$\tau_t = \mathbb{E}_i \mathbb{E}\left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))\right] \mathbb{E}_\pi C(1, 0, P(Z)) \tag{6}$$

We now consider the conditional effects $\tau|X$ instead of population effect τ. These are similar to the earlier expression, however instead of taking expectation over all units $\mathbb{E}_i$ we average over units with matching covariates. This then gives us:

$$\begin{aligned}\tau_t(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t})|X &= \mathbb{E}\left[\mathbb{E}\left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))\right] \mathbb{E}_\pi C(\mathbf{z}^{t-H:t}, \tilde{\mathbf{z}}^{t-H:t}, P(\mathbf{Z}))|X\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))\right]|X\right] C_X(\mathbf{z}^{t-H:t}, \tilde{\mathbf{z}}^{t-H:t}, P(\mathbf{Z}))\end{aligned} \tag{7}$$

$$\tau_t|X = \mathbb{E}\left[\mathbb{E}\left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))\right] \mathbb{E}_\pi C(1, 0, P(Z))|X\right] = \mathbb{E}\left[\mathbb{E}\left[(Y_{it}(\vec{\mathbf{1}}) - Y_{it}(\vec{\mathbf{0}}))\right]|X\right] C_X(1, 0, P(Z)) \tag{8}$$

Note since the treatment allocation policy as well as the risk coefficients c_i are functions of X; they are constant for the conditional expectations in Equations 7,8. Computing their ratio gives the result that:

$$\Lambda(X) := \frac{\tau_t(\mathbf{z}^{s:t}, \tilde{\mathbf{z}}^{s:t})|X}{\tau_t|X} = \frac{C_X(\mathbf{z}^{t-H:t}, \tilde{\mathbf{z}}^{t-H:t}, P_X(\mathbf{Z}))}{C_X(1, 0, P_X(Z))} = \text{constant} \tag{9}$$

B.2 Proof for Theorem 2

First we note that from Assumption 6 a, interference is limited to only a time period of H. Hence for cumulative effect estimation at time t we can only look at session of length which begining at $t - H$.

Next we note that Assumption 6 b, implies that if for a unit i the cookie is lost at time step t i.e. $\Delta_{i,t} = 1$, then the probability of it not being lost again within the next H steps is $> \varepsilon$. This implies for that unit there is atleast an ε chance of getting a trajectory of length H or higher after cookie loss.

Putting both these together as $N \xrightarrow{\infty}$ we will have atleast $N\varepsilon$ sessions of length $> H$. Let us assume that this number is M. Note that with high probability (w.h.p) $M > N\varepsilon$, as anytime a user loses a cookie, there is a chance to get another history; and if a user doesn't lose a cookie we have an even longer session.

Next, at step 3 of the Algorithm 1, we estimate the conditional direct effects $\tau_t(X)$. By standard results (Horvitz and Thompson, 1952; Pearl, 2000), IPW estimation provides an unbiased consistent estimator $\hat{\tau}_t^D(X)$ for the direct effect. Note that this is on the subset of data with history (so sample size is M), and the corresponding error in estimate is $O\Big(\frac{1}{\sqrt{M}}\Big)$.

By similar arguments, step 4 produces unbiased consistent estimator $\hat{\tau}_t^C(X)$ for $\tau_t(\vec{1}, \vec{0}, X)$ with a sample size of M.

By Theorem 1, we know that $\frac{\tau_t(\vec{1},\vec{0},X)}{\tau_t(X)}$ is a function of X (and perhaps t). Now in step 5, we compute $\hat{\Lambda}_t(X) = \hat{\tau}_t^C(X)/\hat{\tau}_t^D(X)$. Since as $\lim_{M\to\infty} \hat{\tau}_t^C(X) \to \tau_t(\vec{1}, \vec{0}, X)$ and $\lim_{M\to\infty} \hat{\tau}_t^D(X) \to \tau_t(X)$, by the Mann-Wald Continuous Mapping theorem, we have

$$\lim_{M\to\infty} \hat{\Lambda}_t(X) \to \Lambda_t(X)$$

.

From Equations 7,8, the function $C_X(.)$ depend only on the individual risks coefficients c_i and the distribution of treatments Z, and by Assumption 5 these are independent. More over since $Y_{i,t} \perp\!\!\!\perp Y_{i,t-1} | X_i$ (Neyman, 1923) and the treatment allocation Z_{it} are also conditionally independent by Assumption 4, so are $\hat{\Lambda}_t$ (as they are functions of independent random variables, in this case of observed outcomes).

In Step 6, we learn a model by fitting $\hat{\Lambda}_t$ to a function. Assuming the function class for the estimator is powerful enough, the independence between $\hat{\Lambda}_t$ implies that we get a consistent estimator for Λ – or, in the case of constant and linear models, an unbiased estimate of Λ.

In Step 8, we use consistent estimate of Λ with $\hat{\tau}_t^g(X)$ (which by earlier arguments is also a consistent estimator of $\tau_t(X)$, which once again via continuous mapping demonstrates the consistency of the final estimator of the conditional effect $\tau_t(\vec{1}, \vec{0}, X)$. By the definition of ATE, it is a weighted combination of the conditional ATEs (CATEs). Since we have consistent CATEs, continuous mapping also gives consistency of the final ATE estimate.

Remark 1. *Since in Step 8, the multiplication is with $\hat{\tau}$ which is from an independent subgroup (and is unbiased), the final estimates will unbiased estimates if Λ_t^f were unbiased as well. While for specific forms like constant/linear models, unbiased estimation can be proven, we need the target values $\hat{\Lambda}_t$ to be unbiased. However our method only gives consistent (but not unbiased) $\hat{\Lambda}_t$. Since the numerator of $\hat{\Lambda}_t$ is unbiased, if we can get high accuracy direct effects one can potentially get unbiasedness as well. We leave such estimation for future research.*

Variance Analysis The ratio Λ is obtained via a ratio of two distributions and is hence difficult to analyze in the general case. However if the number of users is large enough, both the estimates in the numerator and the denominator are asymtotically normal. If the denominator is very likely to be positive (i.e. its mean is much higher than the variance), then Katz et al. (1978) has shown that the above random variable is approximately log-Normal with log variance

$$\left[\frac{\mathrm{Var}[Y_t(1) - Y_t(0)]}{\mathbb{E}[Y_t(1) - Y_t(0)]^2} + \frac{\mathrm{Var}[Y_t(\vec{1}) - Y_t(\vec{0})]}{\mathbb{E}[Y_t(\vec{1}) - Y_t(\vec{0})]^2} \right]$$

where Var refers to the variance of the random variable. Furthermore if the individual variances are small then the Delta method approximation to the variance is:

$$\mathrm{Var}[\hat{\Lambda}] = \Lambda^2 \left[\frac{\mathrm{Var}[Y_t(1) - Y_t(0)]}{\mathbb{E}[Y_t(1) - Y_t(0)]^2} + \frac{\mathrm{Var}[Y_t(\vec{1}) - Y_t(\vec{0})]}{\mathbb{E}[Y_t(\vec{1}) - Y_t(\vec{0})]^2} \right]$$

B.3 Proof for bias under partial linkage

Proposition 1. *Fragmentation bias is non monotonic with respect to fraction of users with fragmented identities*

Let us consider a simple linear model between exposure and outcome.

$$y \sim \alpha + z'\beta \tag{10}$$

Here, y is the outcome of interest (such as the dollar value of purchases); z represents a vector of variables including the treatment i.e. advertising exposure and other relevant covariates, and ε is the error component. For simplicity we assume a two device fragmentation scenario. The user accesses the website through 2 unlinked devices, and has corresponding exposure and purchase recorded for each of the device associated identities. Let y_j, z_j denote the corresponding variables on device $j \in \{1, 2\}$. By construction, $y = y_1 + y_2$ and $z = z_1 + z_2$, representing the aggregate spend and treatment exposure.

The un-fragmented data consists of N identically and independently distributed observations corresponding to N unique consumers. Let $Y = [y_{(1)}, ..., y_{(N)}]'$, $Z_j = [z'_{(1)j}, ..., z'_{(N)j}]$, and define $Z = [\eta \quad Z_1 + Z_2]$ where η is a length-N vector of ones. If consumer-level identities were observed, β can be obtained by regressing Y on Z. When the data is *fragmented*, however, the advertiser instead observes

$$\tilde{Y} \equiv \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \quad \tilde{Z} \equiv \begin{bmatrix} \eta & Z_1 \\ \eta & Z_2 \end{bmatrix}.$$

We use the variable $s_{(i)}$ to capture the purchase preference of the users. In particular, $s_{(i)} = 1$ implies purchase corresponding to user i was on device 1; else it was made on device 2. We can stack these $s_{(i)}$ variables into a diagonal matrix where the i_{th} diagonal is $s_{(i)}$. Then we put this matrix and its complement into a larger matrix of the form $S = \begin{bmatrix} \mathbf{s} \\ I - \mathbf{s} \end{bmatrix}$.

S can then express the relationship between the fragmented and un-fragmented purchases as

$$\tilde{Y} = SY. \tag{11}$$

Similarly we have a relation between the unfragmented covariates (Z) and its fragmented version $\left(\widetilde{Z}\right)$

$$Z = W\widetilde{Z}\Omega. \tag{12}$$

where W denote the block identity matrix $[I_{N\times N} \quad I_{N\times N}]$, and Ω is a $K+1$ diagonal matrix diag$(1/2, 1_K)$.

The OLS regression estimator using fragmented data is given by:

$$\widehat{\theta} = (\widetilde{Z}'\widetilde{Z})^{-1}(\widetilde{Z}'\widetilde{Y}) \tag{13}$$

$$= (\widetilde{Z}'\widetilde{Z})^{-1}(\widetilde{Z}'SY) \tag{14}$$

$$= (\widetilde{Z}'\widetilde{Z})^{-1}[\widetilde{Z}'S(W\widetilde{Z}'\Omega\theta + \varepsilon)] \tag{15}$$

where θ denotes the combined vector $[\alpha\beta]$.

Now let us suppose the pooled data contain a proportion r of fragmented users. In their seminal works Durbin (1953); Theil and Goldberger (1992) show how, for linear models, the parameter obtained from mixed statistical estimation is the weighted average of pure estimators. Specifically in our case it can be shows that

$$\widehat{\theta}^m = \omega\widehat{\theta}^f + (I-\omega)\widehat{\theta}^l, \quad \text{where} \quad \omega = (r\widetilde{Z}'\widetilde{Z} + (1-r)Z'Z)^{-1}r\widetilde{Z}'\widetilde{Z},$$

where $\widehat{\theta}^f$ is the estimator using only the fragmented data, while $\widehat{\theta}^l$ is the estimator obtained from the unfragmented data alone. It follows that $E[\widehat{\theta}^m|Z] - \theta = \omega(E[\widehat{\theta}^f|Z] - \theta)$. Since ω is a matrix and its dependence on r is mediated by two factors, ω does not have a monotonic dependence on r. Secondly, the observed per parameter bias is a combination of different components the vector $E[\widehat{\theta}^f|Z] - \theta$ (mediated by ω). Depending on individual terms in ω, the observed bias can show both inverse sign or amplified magnitude. This means that unlike the attenuation bias of (Coey and Bailey, 2016), where the parameters get shrunk closer to 0, there is no easily quantifiable bias direction or magnitude in the case of partial linkage.

B.4 Proof for Attenuation Bias

We start with the linear model used in Lin and Misra (2021).

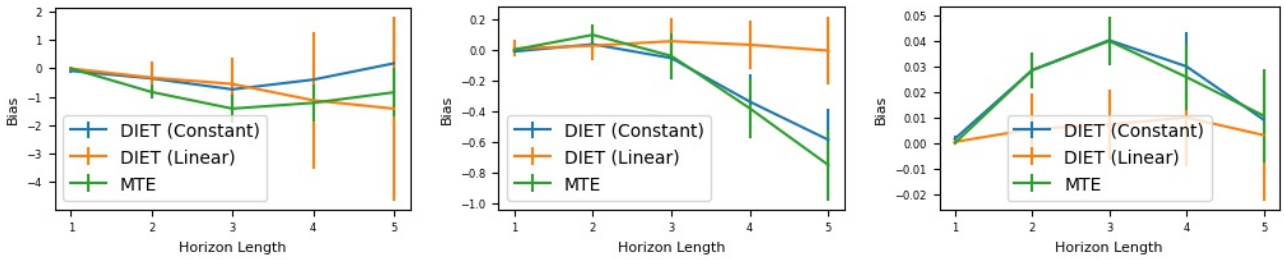
$$Y \sim \alpha + z\eta$$

with a cap on maximum exposure B. Then the expected highest outcome is $[\eta B + \alpha]$ while the lowest outcome is $[\alpha]$. Identifying these as $Y(\vec{1})$ and $Y(\vec{0})$ we can see that the outcome $Y(z)$ at any other exposure z is given by:

$$Y(z) = z\underbrace{\frac{Y(\vec{1}) - Y(\vec{0})}{B}}_{\eta} + \underbrace{Y(\vec{0})}_{\alpha}$$

It is clear that this expression satisfies our outcome model assumption. Even more specifically for a sequence $\vec{z}$, the risk factor $c_i(\vec{z})$ is given by $\sum z_t/H$ where we have divided by H for sake of normalization.

Next Lin and Misra (2021) prove that estimating treatment from the split data, (which would in our terminology be direct treatment effect), provides an attenuated value of the cumulative effect. From our Theorem 1 we know that the direct and



(a) Glucose (b) White Blood Cell (c) Haemoglobin

Figure 6: Bias (with standard error across trials) of different estimators against the horizon of temporal interference on sepsis patients from MIMIC data for different outcome measures.

cumulative effects are related by a constant. Thus the basic result of Lin and Misra (2021) is a direct corollary of Theorem 1.

The only thing left now is to show that the constant is $1/K$, showing attenuation by the number of identities K . This requires translating their assumption about number of identities to temporal fragmentation. Since we have a H horizon period of interference, we assume that this is split sequentially and equally among the K identities. Plugging the corresponding histories into $\mathbf{z}, \mathbf{z}'$ in the definition of C (Equation 5) we see that

$$c_i(\mathbf{z}, Z) - c_i(\mathbf{z}', Z)) = \sum \frac{\mathbf{z}_t + \sum Z}{H} - \frac{\mathbf{z}'_t + \sum Z}{H} = K/H$$

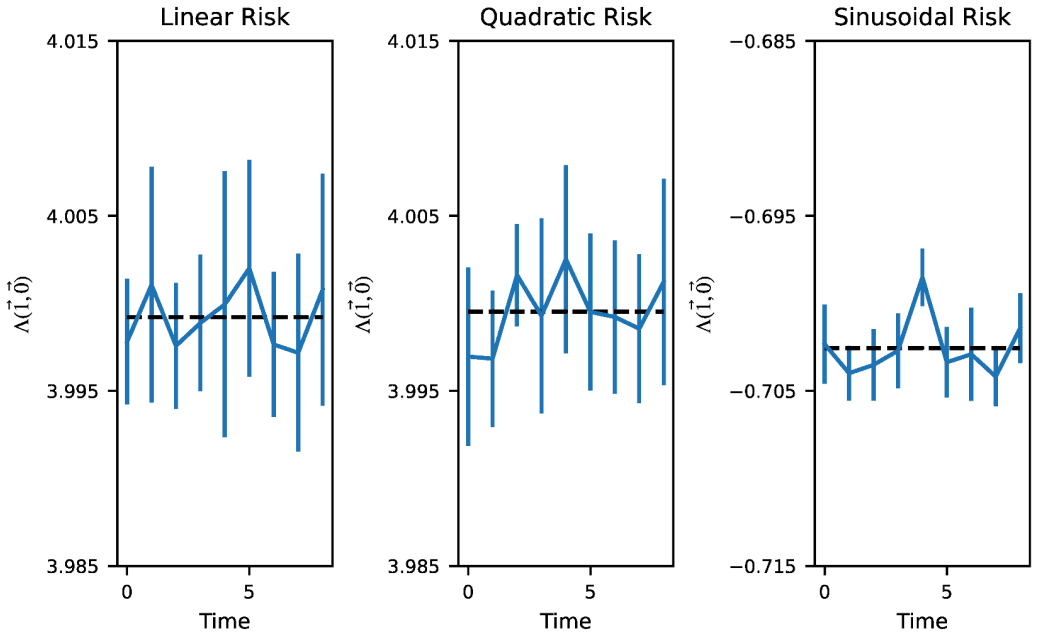
Similarly $c_i(1, Z) - c_i(0, Z)) = 1/H$.

Pluggin this result into Equation 9 gives $\Lambda = \frac{K/H}{1/H} = K$, proving that the cumulative effect is K (number of identities) times the direct effect.

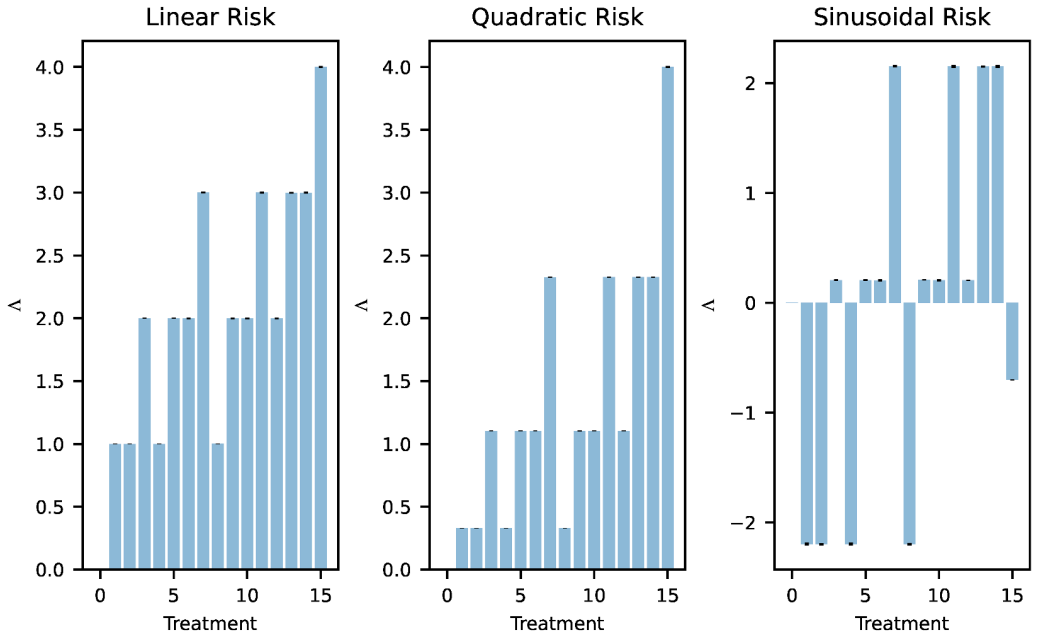
C Further Experiments

We repeat the experiments we performed earlier but focus on the MTE. We first demonstrate temporal interference/history dependence on the outcomes. Next we assess the error of the estimator as the percentage of users with available history varies. In Figure6 we observe a statistically significant bias of using MTE for two of the outcomes (WBC counts and Haemoglobin). We also see greater variance and difference between the two different DIET estimators. We also see some minor improvement in MSE by using DIET and greater variability. This can be partially attributed to the smaller amount of data availability.

D Supplementary Figures



(a) Value of $\lambda(\vec{1})$ for different time steps. Note that this is the population level average of ratio of the marginal and global treatment effect i.e. $\mathbb{E}[Y_t(\vec{1}) - Y_t(\vec{0})]/\mathbb{E}[Y_t(1) - Y_t(0)]$.



(b) Value of $\Lambda(H)$ for different treatment histories.