On Minimal Observation Time for Recovering Latent Hawkes Networks

Jonas Linkerhägner
University of Basel
jonas.linkerhaegner@unibas.ch

Michele Bortolasi
University of Basel
michele.bortolasi@unibas.ch

Lorenzo Baldassari
University of Basel
lorenzo.baldassari@unibas.ch

Maarten V. de Hoop
Rice University
mvd2@rice.edu

Ivan Dokmanić
University of Basel
ivan.dokmanic@unibas.ch

Abstract

Dynamics of interacting systems in engineering, society, and nature often evolve over latent networks that govern which entities can interact. We study the problem of inferring these networks from event-based observations, which arise naturally in finance, seismology, and neuroscience. While there is substantial algorithmic work addressing this important problem, theoretical results are scarce. In this paper we ask the following fundamental question: what is the minimum time that one must observe the dynamics in order to exactly recover the underlying network, as a function of the number d of interacting entities? For a class of stationary Hawkes processes with sparse, weak interactions, we prove that an observation time of order $\log d$ is necessary and sufficient. For the upper bound we construct a two-stage estimator that uses clipped and binned event data for screening, followed by a least-squares refinement, and apply concentration bounds derived from the Poisson cluster representation. For the lower bound we combine Fano's inequality with Jacod's Girsanov formula for point processes on a suitable subclass of networks.

1 Introduction

Reconstructing the latent interaction network is a central problem in the statistical analysis of point processes. In this paper, we ask: *How must the observation time scale with the number of nodes d in order to recover the directed interaction network exactly?*

We give an answer for Hawkes processes [1, 3, 22], when excitations of events occur over a sparse, weakly coupled interaction network. Such processes model a variety of phenomena, including aftershock propagation after earthquakes [47, 67, 41], high-frequency trading [10, 36, 4], neuronal spike trains [56, 51], social activity and information cascades [65, 27], and epidemic spreading [17].

Suppose, for example, that we record earthquake occurrences in a seismic region. The nodes of the underlying network could be subsurface elements such as fault segments or spatial cells; the edges may describe direct triggering influence, mediated, for instance, by stress transfer: that is, how a seismic event in one element modulates the probability that events will occur in other elements. Recovering this directed network helps understand how seismic activity is coordinated and how triggering effects propagate [45, 32].

Such interacting systems are high-dimensional but sparse: a seismic region may contain many fault segments, but each segment directly interacts only with a small number of nearby segments as dictated by the fault geometry. This motivates studying how the observation time required for network reconstruction scales with the number of nodes d, when each node is directly influenced by fixed

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

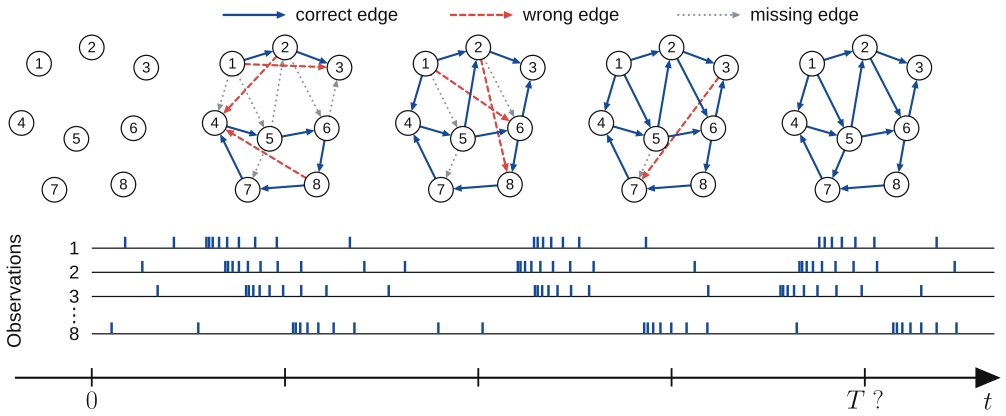


Figure 1: We observe event streams for each node over a window $[0, T]$, and the goal is to recover the hidden directed interaction network. Independent of the choice of a specific estimator, for short observation times, different networks remain statistically indistinguishable, leading to wrong and missing edges. With sufficiently long time, the true network becomes identifiable. Our work quantifies how the observation-time requirement scales with the network size d.

$k \ll d$ others, and these direct influences are individually moderate. Figure 1 provides an illustration. If the observation time T is too small, the observations do not yet include enough information to recover the network correctly. When T is of the correct order in d, exact recovery becomes possible.

Here we provide a theoretical analysis of the network recovery problem. We prove that for a fixed Hawkes model, with fixed sparsity k and sufficiently weak coupling α, the observation time required for exact network recovery scales as $\log d$. The upper bound in Section 3 is constructive. It uses a two-stage estimator that utilizes clipped and binned event data for screening, followed by local least-squares regression and thresholding. Combined with concentration bounds derived from the Poisson cluster representation [26], it shows that exact support recovery is achievable with high probability when

$$T \gtrsim \log d.$$

In Section 4, we prove a matching information-theoretic lower bound by applying Fano's inequality [23] to an explicit subclass of interaction networks. This reduces exact recovery to controlling the mutual information between the latent network and the observations, which we bound via the corresponding Kullback–Leibler divergences using Jacod's Girsanov formula for point processes [29]. As a consequence, for all d large enough, any estimator fails to achieve exact support recovery with non-vanishing probability when the observation time is $T \lesssim \log d$.

Related Work

Hawkes processes are self- and mutually-exciting point processes [25], where events can increase the probability of future events. The cluster representation characterizes events as branching cascades [26]: one event may trigger further events, which may again trigger subsequent generations of events. Classical work on stability and stationarity conditions [13, 12] and likelihood-based estimation methods [46, 49] provide the foundations for our analysis. Since then, Hawkes processes have been widely used to model events such as aftershocks [45], neuronal spike trains [35], social activities [20, 54, 15], epidemics [43, 57], crime [44, 18, 66] and trades and volatility in finance [5, 2, 9].

Existing approaches to recover latent Hawkes networks use penalized [24, 3, 14] and nonparametric estimation [6, 21], causal graph recovery [1], and graph screening [16]. More recent approaches include Bayesian structure learning [38, 52, 40] and Granger-causal recovery [62, 60]. Neural relational inference uses neural networks to infer relations in dynamical systems [34, 50, 61], including point processes [64]. Other works focus on learning full intensity functions of point processes [48, 39], often using Neural ODEs [30, 63]. We complement this line of work with a theoretical result on how the observation time for exact graph recovery scales with network size.

The main challenge compared to other high-dimensional graphical models [42] is that for Hawkes processes, the network to be recovered enters the model endogenously: the interactions of the latent

intensities are governed by the network we want to recover. This makes the problem challenging, since indirect excitation can mimic direct edges that are not present, and even the absence of events is informative. It is qualitatively distinct from observation-time scaling results in Gaussian graphical models [58, 59], Markov random fields [53] and Ornstein–Uhlenbeck diffusions [7, 8], where network dependence is linear and noise is additive.

2 Model, assumptions, and main result

In this section, we introduce the Hawkes model under study, formulate the exact support-recovery problem, and specify the sparse, weak interaction setting in which our analysis takes place.

2.1 Exponential Hawkes process

We consider a d-dimensional exponential Hawkes process observed over a stationary time window $[0, T]$. Writing $N := (N_1, \dots, N_d)$ for the counting process, and using the exponential kernel with decay rate $\beta > 0$, let

$$X_j(t) := \int_{(-\infty,t]} e^{-\beta(t-s)}\, dN_j(s),$$

denote the exponential shot-noise process associated with node j. Each coordinate evolves according to the SDE

$$dX_j(t) = -\beta X_j(t)\, dt + dN_j(t).$$

Thus, with the exponential kernel, the Hawkes process admits a Markov representation. The state variable used throughout the proofs is

$$X(t) = (X_1(t), \dots, X_d(t)).$$

The conditional intensity of node i is

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{d} \theta_{ij} X_j(t-), \qquad i \in [d].$$

Here, $\mu_i > 0$ is the background intensity of node i and $[d] = \{1, \dots, d\}$. $X_j(t-)$ denotes the left limit of the state just before time t, ensuring that the intensity is predictable. The matrix

$$\Theta := (\theta_{ij})_{i,j\in[d]}$$

is the interaction matrix. We consider the excitatory case, so for all $i, j \in [d]$ we have $\theta_{ij} \geq 0$. This model is a standard choice for multivariate event data with self- and cross-excitation, where $\theta_{ij} > 0$ means that the activity at node j directly increases the event rate at node i through the current state $X_j(t-)$. $\theta_{ij} = 0$ means that there is no edge from j to i. We do not consider inhibitory Hawkes processes in this work, since this would require additional constraints or a nonlinear intensity function to ensure that the intensities remain nonnegative.

To focus narrowly on the task of interaction recovery, we treat the decay rate $\beta > 0$ as known throughout the paper. Also, we take the observed data to be

$$Y_T = (X(0), N_{[0,T]}),$$

where $N_{[0,T]}$ is the counting process restricted to the window $[0, T]$. For exponential Hawkes with known β, $X(0)$ is the Markov state summarizing the past. The fact that $X(0)$ is treated as observable is an initialization assumption rather than additional information. In practice, $X(0)$ is available in stationary simulations after burn-in, and can be approximated from a warm-up window of pre-0 event data.

2.2 Exact support-recovery

Our goal is exact network recovery from the observed data Y_T. In practice, this corresponds, for example, to inferring regions with a higher aftershock risk from observing earthquakes, identifying assets that transmit volatility shocks to other assets, or identifying epidemic spreading routes from

observing infections. Since $\theta_{ij} > 0$ corresponds to a directed edge from node j to node i, network recovery corresponds to support recovery of the interaction matrix Θ. For each row i, let

$$S_i := \{j \in [d] : \theta_{ij} > 0\}$$

denote the parent set of node i, that is, the set of source nodes that directly excite node i. We ask how the required observation time T must scale with the ambient dimension d to recover the full support pattern $(S_i)_{i=1}^d$ exactly with high probability.

2.3 Sparse weak-interaction model class

We now formalize the sparse, weak-interaction setting in which exact support recovery will be studied. We fix constants

$$\beta > 0, \qquad 0 < \mu_- \le \mu_+ < \infty, \qquad 0 < w_- \le w_+ < \infty,$$

an integer $k \ge 1$, and a coupling level $\alpha > 0$.

Sparse network structure. Let $\mathcal{B}_{d,k}$ denote the class of nonnegative matrices $B = (b_{ij}) \in \mathbb{R}_+^{d\times d}$ such that

- each row of B has at most k nonzero entries:

$$\|b_{i\cdot}\|_0 \le k, \qquad i \in [d];$$

- every nonzero entry of B lies in $[w_-, w_+]$:

$$b_{ij} \ne 0 \quad \Longrightarrow \quad b_{ij} \in [w_-, w_+].$$

$\mathcal{B}_{d,k}$ specifies the structural part of the model class: each node has at most k direct parents, and every nonzero edge weight is bounded above and below.

Hawkes parameters. We then consider Hawkes models whose interaction matrix is of the form

$$\Theta = \alpha B, \qquad B \in \mathcal{B}_{d,k},$$

and whose background intensities satisfy

$$\mu_i \in [\mu_-, \mu_+], \qquad i \in [d].$$

In addition, we define

$$\theta_- := \alpha w_-, \qquad \theta_+ := \alpha w_+, \qquad \gamma := \frac{k\theta_+}{\beta},$$

where γ is a uniform upper bound on the total excitation that one row can receive, measured relative to the exponential decay rate β. Indeed, since each row has at most k nonzero entries and each interaction weight is at most θ_+,

$$\frac{1}{\beta}\rho(\Theta) \le \frac{1}{\beta}\|\Theta\|_\infty \le \frac{k\theta_+}{\beta} = \gamma.$$

We impose the subcriticality condition $\gamma < 1$, a sufficient form of the standard Hawkes stability condition [25]. It guarantees existence of a unique stationary law and excludes explosive excitation cascades. Together, this yields the class of Hawkes parameters studied in the paper:

$$\mathcal{G}_{d,k} := \Big\{(\mu, \Theta) : \Theta = \alpha B,\ B \in \mathcal{B}_{d,k},\ \mu_i \in [\mu_-, \mu_+]\ \forall i,\ \gamma < 1\Big\}.$$

The weak-interaction setting captures large sparse systems where individual interactions are weak, such as neuronal networks, social activity cascades, and epidemic propagation. Since excitation is sufficiently moderate in these cases, the contribution of a true parent remains detectable, while higher-order cascade effects remain controlled. This is precisely the regime in which one can hope to prove sample complexity bounds by combining support separation with concentration uniformly in d.

Finally, the smallest possible nonzero interaction strength in this class has to fulfill $\theta_- > 0$. This is the minimum-signal condition needed when d grows: without a positive lower bound on nonzero edges, true interactions could become arbitrarily weak, and exact support recovery becomes impossible.

2.4 Main result

In the regime set by $\mathcal{G}_{d,k}$, our main result identifies the correct scaling of the observation time with dimension d for exact network recovery: for fixed k and fixed weak coupling level α,[1] the necessary and sufficient observation-time scaling for exact support recovery is

$$T \asymp \log d.$$

To prove this, we combine an achievability result in Theorem 3.1 with an impossibility result in Theorem 4.1, provided in Sections 3 and 4, respectively.

3 Upper bound on the observation time

In this section, we prove the upper bound by constructing an explicit estimator. Since the estimator serves as a proof tool rather than as a practical recovery algorithm, we do not discuss its computational aspects here. Figure 2 summarizes the procedure. The estimator is built node by node because exact support recovery is equivalent to identifying, for each node i, the set of nodes with a direct influence on it. The estimator proceeds in two stages.

The screening stage asks which nodes are active shortly before node i has an event. This is measured by the empirical covariance score between the current activity of a candidate node j and the occurrence of a future event of node i. The activity is reconstructed from the observed event counts, sampled at time resolution h, and clipped at level R. Screening does not require knowing the exact maximum number of parents per node k, but only an upper bound $m \geq k$. The m nodes with the largest scores form a candidate list of possible parents. In the weak-interaction regime, these scores are shown to separate true parents from non-parents, ensuring that the true support is contained in the candidate set (sure screening).

In the second stage, the estimator reevaluates this reduced set to separate genuine direct influences from multi-hop correlations. It fits a local ordinary least-squares regression on the selected candidates and thresholds the estimated coefficients. Row sparsity keeps this low-dimensional, while clipping and binning allow the relevant empirical quantities to concentrate.

Overall, this strategy turns the high-dimensional recovery problem into a collection of smaller, local subproblems. We now make the construction precise.

Definition of the estimator. It is defined row by row. Fix a target node $i \in [d]$. We choose a bin size $h > 0$, a clipping level $R > 0$, and a candidate set size $m \geq k$. Let $n = \lfloor T/h \rfloor$, so that we use the discrete time points $0, h, \ldots, (n-1)h$. For each bin $r = 0, \ldots, n-1$, we record two quantities. First, we sample the state on a grid and define

$$U_{j,r} := X_j(rh), \qquad U_r = (U_{1,r}, \ldots, U_{d,r}),$$

and the clipped and sampled state of node j as

$$\psi_R(x) := \min\{x, R\}, \qquad Z_{j,r}^{(R)} := \psi_R(U_{j,r}).$$

Second, we record the binned event indicator for each node i as

$$Y_{i,r}^{(h)} := \mathbf{1}\{N_i((rh, (r+1)h]) \geq 1\}.$$

$Z_{j,r}^{(R)}$ is the clipped and sampled candidate state, measuring the past activity of node j, while $Y_{i,r}^{(h)}$ measures whether node i has an event shortly afterwards.

The first stage screens candidate parents using the empirical covariance between these two quantities:

$$\widehat{F}_{ij}^{(h,R)} := \frac{1}{n}\sum_{r=0}^{n-1} Z_{j,r}^{(R)} Y_{i,r}^{(h)} - \Big(\frac{1}{n}\sum_{r=0}^{n-1} Z_{j,r}^{(R)}\Big)\Big(\frac{1}{n}\sum_{r=0}^{n-1} Y_{i,r}^{(h)}\Big). \tag{1}$$

For the target node i, let $\widehat{C}_i$ be the set of indices corresponding to the m largest values of $\{\widehat{F}_{ij}^{(h,R)} : j \in [d]\}$. The set $\widehat{C}_i$ is the screening set: it contains the nodes whose clipped and sampled activity is

[1] See Appendix A.6 for a preliminary discussion of the dependence on the remaining parameters, whose optimal scaling remains open.

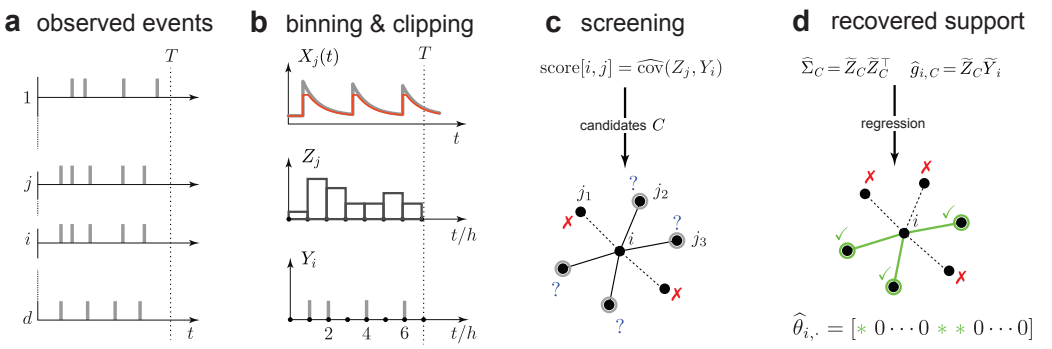


Figure 2: **a** We observe event streams from a Hawkes process with an underlying latent network. To recover the network, our two-stage estimator works node by node. **b** For each node i, it converts the observed event times into a binned activity trace Y_i. Also, for all candidates $j \in [d]$, it computes the state $X_j(t)$ and its sampled and clipped variant Z_j. **c** The first stage screens for nodes that are active shortly before an event at node i, retaining the nodes with the highest scores as a candidate set C of possible parents (grey). **d** The second stage centers Y_i and Z_j, performs a local regression on this reduced candidate set, and thresholds the fitted coefficients to recover the correct support (green).

most correlated with future events of node i. The second stage refines this candidate set by a local least-squares regression. We first center the clipped and sampled state and the binned event indicator,

$$\widetilde{Z}_{j,r} := Z_{j,r}^{(R)} - \frac{1}{n}\sum_{r=0}^{n-1} Z_{j,r}^{(R)}, \qquad \widetilde{Y}_{i,r} := Y_{i,r}^{(h)} - \frac{1}{n}\sum_{r=0}^{n-1} Y_{i,r}^{(h)}.$$

For any candidate set $C \subset [d]$, we write

$$\widetilde{Z}_{C,r} := (\widetilde{Z}_{j,r})_{j \in C}.$$

We then define the empirical covariance matrix and the empirical cross-covariance:

$$\widehat{\Sigma}_C^{(R)} := \frac{1}{n}\sum_{r=0}^{n-1} \widetilde{Z}_{C,r}\widetilde{Z}_{C,r}^\top, \qquad \widehat{g}_{i,C}^{(h,R)} := \frac{1}{n}\sum_{r=0}^{n-1} \widetilde{Z}_{C,r}\widetilde{Y}_{i,r}. \tag{2}$$

On the screened set $\widehat{C}_i$, if $\widehat{\Sigma}_{\widehat{C}_i}^{(R)}$ is invertible, we define the local least-squares coefficient vector by

$$\widehat{y}_{i,\widehat{C}_i}^{(h,R)} := \big(\widehat{\Sigma}_{\widehat{C}_i}^{(R)}\big)^{-1} \widehat{g}_{i,\widehat{C}_i}^{(h,R)}. \tag{3}$$

The estimated parent set of node i is obtained by thresholding the local least-squares coefficients. We choose the threshold $\tau_{\alpha,h} := \alpha w_- h/2$ and define

$$\widehat{S}_i := \{j \in \widehat{C}_i : \widehat{y}_{ij}^{(h,R)} \geq \tau_{\alpha,h}\}. \tag{4}$$

We can now state the exact support-recovery result for the bounded-block estimator.

Theorem 3.1. *Fix $m \geq k$. There exists $\alpha_0 > 0$ such that for every fixed $0 < \alpha \leq \alpha_0$ there exist constants $C_*, c_* > 0$ only depending on $(\beta, \mu_\pm, w_\pm, k, m, \alpha)$, and $d_0 \geq m$ such that for every $d > d_0$, every $B \in \mathcal{B}_{d,k}$, and every observation time satisfying*

$$T \geq C_* \log d,$$

for every row i, the empirical covariance matrix in (2) *is invertible, and the bounded-block estimator defined in* (1), (3) *and* (4) *satisfies*

$$\mathbb{P}\big(\widehat{S}_i = S_i, \ \forall i\big) \geq 1 - d^{-c_*}.$$

Sketch of proof. The proof has three ingredients; the full details are in Appendix A. First, for fixed (h, R), the theoretical screening scores

$$F_{ij}^{(h,R)} := \mathrm{Cov}\big(Z_{j,0}^{(R)}, Y_{i,0}^{(h)}\big)$$

admit the first-order expansion

$$F_{ij}^{(h,R)} = hH_{ij}^{(R)} + O(Rh^2), \qquad H_{ij}^{(R)} \approx \frac{\mu_j \theta_{ij}}{2\beta},$$

uniformly over the weak-interaction class. More precisely, for suitable fixed choices $R = R(\alpha) \asymp \alpha^{-1}$ and $h = h(\alpha) \asymp \alpha^2$,

$$\min_{j \in S_i} F_{ij}^{(h,R)} - \max_{j \notin S_i} F_{ij}^{(h,R)} \geq c_{\mathrm{scr}}\, \alpha w_- h,$$

uniformly over $\mathcal{G}_{d,k}$; see Proposition A.5. Thus, every true parent has a strictly larger theoretical screening score than every non-parent.

Second, the bounded block statistics concentrate uniformly. Since $Z_{j,r}^{(R)} \leq R$ and $Y_{i,r}^{(h)} \in \{0,1\}$, the Poisson cluster representation gives a Bernstein-type inequality for the dependent block averages. Consequently, with probability at least $1 - d^{-c}$, simultaneously for all rows, all coordinates, and all candidate sets C with $|C| \leq m$,

$$\max_{i,j} |\widehat{F}_{ij}^{(h,R)} - F_{ij}^{(h,R)}| + \max_{i,C} \|\widehat{g}_{i,C}^{(h,R)} - g_{i,C}^{(h,R)}\|_\infty + \max_{C} \|\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)}\|_{\mathrm{op}} \leq C\delta_T,$$

where

$$\delta_T := \sqrt{\frac{(h^{-1} + R^2 h + R^2/h)\log d}{n}}, \qquad n = \lfloor T/h \rfloor;$$

see Lemma A.12. If $C\delta_T \leq c_{\mathrm{scr}} \alpha w_- h/2$, the empirical screening scores preserve the screening gap, and hence $S_i \subseteq \widehat{C}_i$ for every row i; see Proposition A.9.

Third, on this same high-probability event, the local least-squares step is stable; see Lemma A.4. The theoretical Gram matrices on all sets $C \supseteq S_i$, $|C| \leq m$, have eigenvalues bounded away from zero, and their theoretical regression coefficients satisfy the separation

$$\min_{j \in S_i} y_{ij,C}^{(h,R)} \geq \frac{3}{4}\alpha w_- h, \qquad \max_{j \in C \setminus S_i} |y_{ij,C}^{(h,R)}| \leq \frac{1}{4}\alpha w_- h;$$

see Proposition A.6. The preceding concentration bounds and a standard matrix-inversion perturbation argument imply

$$\|\widehat{y}_{i,\widehat{C}_i}^{(h,R)} - y_{i,\widehat{C}_i}^{(h,R)}\|_\infty \leq C'\delta_T.$$

Taking $C'\delta_T \leq \alpha w_- h/8$, thresholding at $\tau_{\alpha,h} = \alpha w_- h/2$ includes every parent and excludes every non-parent; see Proposition A.10. Since $h(\alpha)$ and $R(\alpha)$ are fixed once α is fixed, the required small-deviation conditions are implied by $T \geq C \log d$. This proves the theorem. □

4 Information-theoretic lower bound

Theorem 3.1 shows that there exists a procedure that achieves exact network recovery once the observation time T is of order $\log d$. We now check whether this is the intrinsic scaling of the problem, or whether it comes from how we constructed the procedure. In this section, we prove that the logarithmic scaling of the observation time in d is optimal for the problem by establishing an information-theoretic lower bound. It holds uniformly over all estimators based on the observed process, regardless of their construction or computational cost. Rather than treating the full model class of Section 2 in complete generality, we restrict our attention to a simple subclass. This is sufficient for proving optimality: since Theorem 3.1 already gives an upper bound of order $\log d$ over the full class, it remains only to find a subclass with matching scaling in d. Any such lower bound then applies *a fortiori* to the full class.

The subclass we consider consists of interaction matrices with all rows equal to zero except for one target row $i_\star$. Fix $i_\star \in [d]$ and background intensities $\bar{\mu}, \bar{\mu}_\star \in [\mu_-, \mu_+]$. The only unknown part of the interaction matrix is the support of size k of the $i_\star$-th row. We denote the set of admissible supports by

$$\mathcal{S}_{d,k}(i_\star) := \{S \subset [d] \setminus \{i_\star\} : |S| = k\}.$$

Thus, each $S \in \mathcal{S}_{d,k}(i_\star)$ specifies which k nodes can influence node $i_\star$. For a given support S, we define the corresponding interaction matrix $\Theta^{(S)} = (\theta_{ij}^{(S)})$ by

$$\theta_{i_\star j}^{(S)} = \begin{cases} \theta_-, & j \in S, \\ 0, & j \notin S, \end{cases} \qquad \theta_{ij}^{(S)} = 0 \quad \text{for } i \neq i_\star.$$

Thus, in this subclass, exact support recovery amounts to identifying the unknown set S from the observation $Y_T = (X(0), N_{[0,T]})$. This construction deliberately removes most interactions and indirect cascades rather than relying on them for the analysis.

We now place a uniform prior on S. The initial uncertainty about the correct support is then $\log |\mathcal{S}_{d,k}(i_\star)| = \log \binom{d-1}{k}$ nats (using natural logarithmic convention). Therefore, an estimator can succeed only if the observation Y_T carries a comparable amount of information about S. This information is measured by the mutual information $I(S; Y_T)$. Fano's inequality turns this intuition into a lower bound on the error probability: if $I(S; Y_T)$ is much smaller than $\log |\mathcal{S}_{d,k}(i_\star)|$, then many supports remain statistically indistinguishable, and every estimator must fail with non-negligible probability. For more details, see Appendix B.

Since $\bar{\mu}, \bar{\mu}_\star \in [\mu_-, \mu_+]$ and every active edge has the minimum allowed signal $\theta_- = \alpha w_-$, every $\Theta^{(S)}$, with $S \in \mathcal{S}_{d,k}(i_\star)$ belongs to $\mathcal{G}_{d,k}$. Thus, any impossibility result proved for this subclass applies *a fortiori* to the full class $\mathcal{G}_{d,k}$. This is made precise by the following theorem, whose complete proof is given in Appendix C.

Theorem 4.1. *For fixed $k \geq 1$, there exist constants $c > 0$ and $\varepsilon \in (0, 1)$, depending only on $(\beta, \bar{\mu}, \bar{\mu}_\star, \theta_-, k)$, such that for all sufficiently large d, if $T \leq c \log d$, then*

$$\inf_{\widehat{S}} \sup_{S \in \mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S(\widehat{S} \neq S) \geq \varepsilon.$$

Consequently,

$$\inf_{\widehat{G}} \sup_{(\mu,\Theta) \in \mathcal{G}_{d,k}} \mathbb{P}_{\mu,\Theta}\big(\widehat{G}(Y_T) \neq \operatorname{supp}(\Theta)\big) \geq \varepsilon.$$

In particular, along any sequence $T_d = o(\log d)$, the worst-case exact-recovery error over the full class is bounded away from zero. Combined with Theorem 3.1, this identifies the optimal dimension dependence as logarithmic for fixed k and fixed coupling level.

Sketch of proof. Let $S^\star$ be uniformly distributed on $\mathcal{S}_{d,k}(i_\star)$, and let Y_T be the stationary observation generated under the interaction matrix $\Theta^{(S^\star)}$ and the fixed background intensities $\bar{\mu}, \bar{\mu}_\star$. Fano's inequality gives, for any estimator $\widehat{S} = \widehat{S}(Y_T)$ of $S^\star$,

$$\sup_{S \in \mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S(\widehat{S} \neq S) \geq 1 - \frac{I(S^\star; Y_T) + \log 2}{\log |\mathcal{S}_{d,k}(i_\star)|}.$$

Since $\log |\mathcal{S}_{d,k}(i_\star)| = \log \binom{d-1}{k}$, it remains to bound the mutual information $I(S^\star; Y_T)$. Let P_S^T be the law of the observation Y_T when the parent set is S. Let P_0^T be the reference law in which all interaction coefficients are set to zero, while the background intensities $\bar{\mu}, \bar{\mu}_\star$ are kept unchanged. The mutual information $I(S^\star; Y_T)$ measures how much information the observation contains about the unknown support $S^\star$. We do not bound this mutual information directly. Instead, we compare every possible law P_S^T with the same law P_0^T. Since P_0^T does not depend on S, Lemma C.2 gives

$$I(S^\star; Y_T) \leq \frac{1}{|\mathcal{S}_{d,k}(i_\star)|} \sum_{S \in \mathcal{S}_{d,k}(i_\star)} \mathrm{KL}(P_S^T \| P_0^T). \tag{5}$$

It is therefore enough to obtain a bound on $\mathrm{KL}(P_S^T \| P_0^T)$ that holds uniformly over $S \in \mathcal{S}_{d,k}(i_\star)$. For each S, the KL divergence splits into two terms: one coming from the stationary law of $X(0)$, and one coming from the conditional law of the counting process $N_{[0,T]}$ given $X(0)$; see Lemma C.1. The stationary-law contribution is bounded by a constant C_{init}, independent of both d and T. The reason is that, after a permutation of the nodes, the stationary laws differ only on the fixed $(k+1)$-dimensional block formed by node $i_\star$ and its k parents; see Proposition C.1.

For the conditional-law contribution, the laws P_S^T and P_0^T differ only through the intensity of the counting process $N_{i_\star}$:

$$\lambda_{i_\star}^{(S)}(t) - \lambda_{i_\star}^{(0)}(t) = \theta_- \sum_{j \in S} X_j(t-).$$

Under P_0^T, all interaction coefficients are zero, so the intensity of $N_{i_\star}$ is equal to its background intensity: $\lambda_{i_\star}^{(0)}(t) \equiv \bar{\mu}_\star$ for all t. Jacod's point-process Girsanov formula [29] therefore yields

$$\mathrm{KL}(P_S^T \| P_0^T) \le C_{\mathrm{init}} + \frac{\theta_-^2}{\bar{\mu}_\star} \int_0^T \mathbb{E}_S\Big[\Big(\sum_{j \in S} X_j(t-)\Big)^2\Big]\, dt;$$

see Proposition C.2. Now the advantage of the subclass becomes clear: for $j \neq i_\star$, the processes X_j are independent stationary Poisson shot-noise processes with background intensity $\bar{\mu}$. Hence,

$$\mathbb{E}_S\Big[\Big(\sum_{j \in S} X_j(t-)\Big)^2\Big] = \frac{k^2 \bar{\mu}^2}{\beta^2} + \frac{k\bar{\mu}}{2\beta} =: C_{\mathrm{path}},$$

uniformly in S, d, and t; see Lemma C.5. Therefore,

$$\mathrm{KL}(P_S^T \| P_0^T) \le C_{\mathrm{init}} + \frac{\theta_-^2}{\bar{\mu}_\star} C_{\mathrm{path}}\, T.$$

Substituting this bound into Fano's inequality gives

$$\inf_{\widehat{S}} \sup_{S \in \mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S(\widehat{S} \neq S) \ge 1 - \frac{C_{\mathrm{init}} + \frac{\theta_-^2}{\bar{\mu}_\star} C_{\mathrm{path}} T + \log 2}{\log \binom{d-1}{k}}; \tag{6}$$

see Proposition C.3. Since k is fixed, $\log \binom{d-1}{k} \asymp \log d$, so the right-hand side is bounded away from zero whenever $T \le c \log d$ for $c > 0$ small enough.

Finally, any estimator of the full support matrix over $\mathcal{G}_{d,k}$ induces, on this subclass, an estimator of the unknown support S by restricting the estimated support to the $i_\star$-th row. Hence the lower bound obtained on the subclass also applies to exact support recovery over the full class $\mathcal{G}_{d,k}$. □

Based on Theorem 4.1, we can make a few remarks:

- The term C_{init} is the bound on the information contained in stationary initial state $X(0)$. It does *not* depend on d and does not grow with T. Therefore, in the dimension-asymptotic regime with all other parameters fixed, it contributes only an additive $O(1)$ term. The part of the KL bound that grows with the observation time is

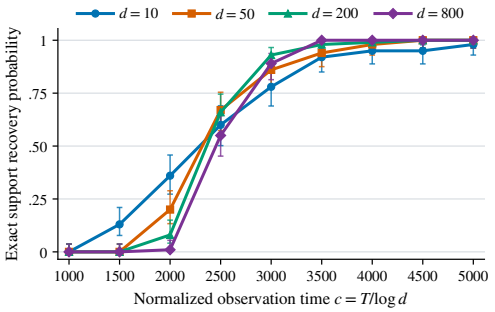
  $$\frac{\theta_-^2}{\bar{\mu}_\star}\, C_{\mathrm{path}}\, T, \qquad C_{\mathrm{path}} = \frac{k^2 \bar{\mu}^2}{\beta^2} + \frac{k\bar{\mu}}{2\beta}.$$

  Thus the lower-bound argument identifies the necessary information scale

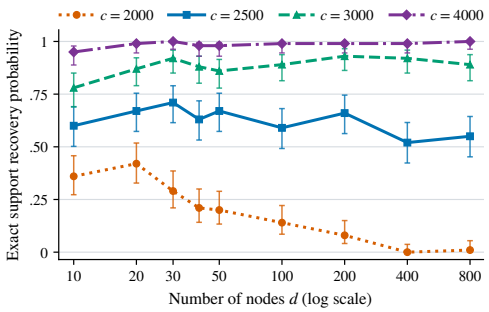
  $$T \gtrsim \frac{\bar{\mu}_\star}{\theta_-^2\, C_{\mathrm{path}}} \log \binom{d-1}{k},$$

  up to dimension-independent additive constants depending on $(\beta, \bar{\mu}, \bar{\mu}_\star, \theta_-, k)$.

- The logarithmic dependence on d is the combinatorial part of the required observation time. The constant prefactor reflects the amount of information generated per unit time. For d sufficiently large, it can be read informally as follows: a larger minimum edge strength θ_- makes recovery easier, since distinguishability scales quadratically in θ_-; a larger background intensity at node $i_\star$ makes recovery harder, because it increases the background activity against which parental influence must be detected; a larger background activity $\bar{\mu}$ of the parent process makes recovery easier, as it increases the aggregated signal power C_{path} of the parents; and a smaller decay rate β makes recovery easier, since longer memory strengthens the cumulative effect of parent events.



(a) Empirical recovery probability versus normalized observation time $c = T/\log d$ for representative d.



(b) Recovery probability versus number of nodes d at representative c, using a logarithmic horizontal axis.

Figure 3: Empirical illustrations of the observation time scaling for finite dimensions. Error bars indicate 95% Wilson confidence intervals based on 100 independent repetitions per configuration.

5 An illustration

Figure 3 numerically illustrates that the proven $\log d$ scaling is already visible at finite dimensions. We simulate stationary exponential Hawkes networks with $\beta = 1$, $\mu_i = 1$, at most $k = 5$ parents per node, and homogeneous interaction weights $\theta_{ij} = \alpha = 0.1$. In each repetition, we sample a new random network support. The estimator uses $m = 10$, $h = 0.02$, $R = 80$, and $\tau = 10^{-3}$. We consider different d at observation times $T = c \log d$, discard a burn-in period of length 50, and run 100 independent repetitions. Each point reports the fraction of repetitions in which the entire support is recovered exactly, together with its 95% Wilson confidence interval. The smaller dimensions are included to illustrate finite-size effects. The code is publicly available on GitHub.

Figure 3a shows that the recovery curves approximately align when the observation time is normalized by $\log d$. Figure 3b provides a complementary perspective by plotting recovery probability against network dimension at fixed values of c. Even for large network dimensions d, comparable recovery performance is achieved at approximately the same normalized observation times.

6 Conclusion and future work

We proved that the observation time required for exact network recovery in Hawkes processes scales logarithmically with the network size d. This puts the success of empirical work on a theoretical basis: a large network size does not make exact recovery prohibitive, because d only enters through the combinatorial cost of identifying a sparse parent set. To the best of our knowledge, this is the first result of its kind. We also identify the key factors that make our proof possible. The known exponential kernel provides a Markov state representation, the intensity depends linearly on this state, and stationarity removes burn-in effects while providing uniform distributional control over the observation window. Weak interactions play an equally important role. Indirect excitation cascades may create correlations between non-adjacent nodes and thereby mimic direct causal influence. The weak-interaction condition ensures that such effects remain smaller than the contribution of a true parent. This separation is the central statistical mechanism behind both the screening gap and the stability of the local regression step.

A natural direction for future work is to sharpen the dependence on other parameters, especially the coupling strength. This is particularly important for applications where the available observation window is usually constrained. Other extensions include signed or inhibitory interactions, unknown or coordinate-dependent decay rates, marked Hawkes processes, and spatially structured networks. Signed interactions are challenging because cancellations may destroy separation arguments and the intensity may approach zero. Marked and spatial models would bring the theory closer to applications.

One speculative but promising application area is earthquake dynamics. Earthquake catalogs are naturally event data, and classical aftershock models already use Hawkes processes to describe triggering. It would be interesting to develop a multiscale (marked and spatial) extension of the theory and connect it to emerging network-based forecasting approaches.

Acknowledgments

This work was funded by the Swiss National Science Foundation (SNSF) under Grant No. 10007161.

References

[1] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. *Journal of Machine Learning Research*, 18(192):1–28, 2018.

[2] Yacine Aït-Sahalia, Julio Cacho-Diaz, and Roger JA Laeven. Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606, 2015.

[3] Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Jean-Francois Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020.

[4] Emmanuel Bacry, Khalil Dayri, and Jean-François Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85(5):157, 2012.

[5] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes Processes in Finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015.

[6] Emmanuel Bacry and Jean-François Muzy. First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

[7] José Bento, Morteza Ibrahimi, and Andrea Montanari. Learning Networks of Stochastic Differential Equations. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[8] José Bento, Morteza Ibrahimi, and Andrea Montanari. Information theoretic limits on learning stochastic differential equations. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 855–859. IEEE, 2011.

[9] Alessandro Bondi, Sergio Pulido, and Simone Scotti. The rough Hawkes Heston stochastic volatility model. *Mathematical Finance*, 34(4):1197–1241, 2024.

[10] Clive G Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[11] Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer Series in Statistics. Springer-Verlag, New York, Heidelberg, Berlin, 1981.

[12] P. Brémaud, G. Nappo, and G. L. Torrisi. Rate of Convergence to Equilibrium of Marked Hawkes Processes. *Journal of Applied Probability*, 39(1):123–136, 2002.

[13] Pierre Brémaud and Laurent Massoulié. Stability of Nonlinear Hawkes Processes. *The Annals of Probability*, 24(3):1563–1588, 1996.

[14] Biao Cai, Jingfei Zhang, and Yongtao Guan. Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association*, 119(545):95–108, 2024.

[15] Feng Chen and Wai Hong Tan. Marked self-exciting point process modelling of information diffusion on Twitter. *The Annals of Applied Statistics*, 12(4):2175–2196, 2018.

[16] Shizhe Chen, Daniela Witten, and Ali Shojaie. Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electronic journal of statistics*, 11(1):1207, 2017.

[17] Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International journal of forecasting*, 38(2):505–520, 2022.

[18] Nicholas J. Clark and Philip M. Dixon. Modeling and Estimation for Self-Exciting Spatio-Temporal Models of Terrorist Activity. *The Annals of Applied Statistics*, 12(1):633–653, 2018.

[19] Rama Cont and Peter Tankov. *Financial modelling with jump processes*. Chapman and Hall/CRC, 2003.

[20] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[21] Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics*, 48(5):2698 – 2727, 2020.

[22] Paul Embrechts and Matthias Kirchner. Hawkes graphs. *Theory of Probability & Its Applications*, 62(1):132–156, 2018.

[23] Sébastien Gerchinovitz, Pierre Ménard, and Gilles Stoltz. Fano's Inequality for Random Variables. *Statistical Science*, 35(2):178 – 201, 2020.

[24] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

[25] Alan G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1):83–90, 1971.

[26] Alan G. Hawkes and David Oakes. A Cluster Process Representation of a Self-Exciting Process. *Journal of Applied Probability*, 11(3):493–503, 1974.

[27] Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, and Yan Liu. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *International Conference on Machine Learning*, pages 871–880. PMLR, 2015.

[28] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24. Elsevier, 2014.

[29] Jean Jacod. Multivariate point processes: predictable projection, Radon-Nikodym derivatives, representation of martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 31(3):235–253, 1975.

[30] Junteng Jia and Austin R Benson. Neural Jump Stochastic Differential Equations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] Yu M Kabanov. The capacity of a channel of the Poisson type. *Theory of Probability & Its Applications*, 23(1):143–147, 1978.

[32] Kiymet Kaya, Murat Sahin, Furkan Mangir, Alp Baris Beydemir, Cenk Yaltirak, and Sule Gunduz Oguducu. Deep quake dynamics: A multimodal fault-aware approach to earthquake magnitude and occurrence time forecasting. *Geoscience Data Journal*, 13(2):e70066, 2026. e70066 GDJ-2025-02-0014.

[33] Alexander Kechris. *Classical descriptive set theory*. Springer Science & Business Media, 2012.

[34] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural Relational Inference for Interacting Systems. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2688–2697. PMLR, 10–15 Jul 2018.

[35] Ryota Kobayashi, Shuhei Kurita, Anno Kurth, Katsunori Kitano, Kenji Mizuseki, Markus Diesmann, Barry J Richmond, and Shigeru Shinomoto. Reconstructing neuronal circuitry from parallel spike trains. *Nature communications*, 10(1):4468, 2019.

[36] Jeremy Large. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1):1–25, 2007.

[37] Günter Last and Mathew Penrose. *Lectures on the Poisson process*, volume 7. Cambridge University Press, 2018.

[38] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421. PMLR, 2014.

[39] David Lüdke, Marin Biloš, Oleksandr Shchur, Marten Lienen, and Stephan Günnemann. Add and thin: Diffusion for temporal point processes. *Advances in Neural Information Processing Systems*, 36:56784–56801, 2023.

[40] Noa Malem-Shinitski, César Ojeda, and Manfred Opper. Variational Bayesian inference for nonlinear Hawkes process with Gaussian process self-effects. *Entropy*, 24(3):356, 2022.

[41] David Marsan and Olivier Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

[42] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436 – 1462, 2006.

[43] Sebastian Meyer, Johannes Elias, and Michael Höhle. A space–time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616, 2012.

[44] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[45] Shyam Nandan, Guy Ouillon, and Didier Sornette. Are large earthquakes preferentially triggered by other large events? *Journal of Geophysical Research: Solid Earth*, 127(8):e2022JB024380, 2022. e2022JB024380 2022JB024380.

[46] Yosihiko Ogata. Estimators for stationary point processes. *Ann. Inst. Statist. Math*, 30(Part A):243–261, 1978.

[47] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[48] Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. *Advances in neural information processing systems*, 32, 2019.

[49] Tohru Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

[50] Liming Pan, Cheng Shi, and Ivan Dokmanic. A graph dynamics prior for relational inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14508–14516, 2024.

[51] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[52] Jakob Gulddahl Rasmussen. Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

[53] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287 – 1319, 2010.

[54] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be hip: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pages 735–744, 2017.

[55] Ken-Iti Sato. *Lévy processes and infinitely divisible distributions*, volume 68. Cambridge university press, 1999.

[56] Wilson Truccolo, Uri T Eden, Matthew R Fellows, John P Donoghue, and Emery N Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, 93(2):1074–1089, 2005.

[57] H Juliette T Unwin, Isobel Routledge, Seth Flaxman, Marian-Andrei Rizoiu, Shengjie Lai, Justin Cohen, Daniel J Weiss, Swapnil Mishra, and Samir Bhatt. Using Hawkes Processes to model imported and local malaria cases in near-elimination settings. *PLoS Computational Biology*, 17(4):e1008830, 2021.

[58] Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741, 2009.

[59] Martin J Wainwright. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[60] Dongxia Wu, Tsuyoshi Ide, Georgios Kollias, Jiri Navratil, Aurelie Lozano, Naoki Abe, Yian Ma, and Rose Yu. Learning Granger Causality from Instance-wise Self-attentive Hawkes Processes. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 415–423. PMLR, 02–04 May 2024.

[61] Ruizi Wu, Liming Pan, and Linyuan Lü. RIVA: Efficient relational inference with variate attention. *Neural Networks*, 191:107748, 2025.

[62] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning Granger Causality for Hawkes Processes. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1717–1726, New York, New York, USA, 20–22 Jun 2016. PMLR.

[63] Shuai Zhang, Chuan Zhou, Yang Aron Liu, Peng Zhang, Xixun Lin, and Zhi-Ming Ma. Neural jump-diffusion temporal point processes. In *Forty-first International Conference on Machine Learning*, 2024.

[64] Yunhao Zhang and Junchi Yan. Neural Relation Inference for Multi-dimensional Temporal Point Processes via Message Passing Graph. In *IJCAI*, pages 3406–3412, 2021.

[65] Ke Zhou, Hongyuan Zha, and Le Song. Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 641–649, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.

[66] Jiancang Zhuang and Jorge Mateu. A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(3):919–942, 2019.

[67] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.

A Upper bound proof details

Appendix A contains the proof of the upper bound. We organize it into five sections. We begin with an intuitive description and a short overview before turning to the technical details.

As presented in Section 3, the whole proof for the upper bound is based on the construction of an estimator composed by a first screening stage and a second Ordinary Least Squares regression stage. The choice for this two-steps procedure is not unique: a method that optimizes over the remaining

parameters, or a more "compressed" one, likely exists. Nevertheless, we believe these choices simplify the study of the scaling in d, which is the main goal of this work. In fact, it is a natural approach to first reduce the search space by choosing a candidate superset and, after guaranteeing that it contains the information about the true interactions, refine the search via regression. Note also that screening first is necessary[2] because, at $n = O(logd) \ll d$, the unrestricted empirical Gram matrix for the regression step is singular.

Instead of using the raw and unbounded process, we clip the covariates (i.e., the states or specific combinations of them) at a finite level R, and then discretize time into small bins of size h. This yields a discrete formulation of the problem, which simplifies the analysis.

The immediate next step is to show that this preprocessing preserves the signal: the covariance used to distinguish true parents of events (trigger nodes) from non-parents must survive clipping and binning, and the resulting regression problem must remain well posed. Once this foundation is established, the next challenge is statistical: the empirical averages computed from the data must be shown to concentrate around their true values. This is nontrivial because the binned observations are not independent, as events in a Hawkes process can trigger later events. The Poisson cluster representation, however, isolates the independent cluster roots, making it possible to control dependence across bins and derive explicit exponential tail bounds. Finally, we combine the theoretical separation between parents and non-parents with these concentration bounds to show that the two-stage estimator succeeds with high probability.

The contents of the sections are outlined in the following list:

- **Section A1** contains some lemmas resulting from the choice of the framework and the assumptions made; these results are not specific to the estimator but rather reflect fundamental properties of stable multivariate Hawkes processes, and they serve as the basic toolkit invoked repeatedly throughout the subsequent sections
- **Section A2** introduces clipping and binning and analyzes the effect on covariances and, consequently, on screening gaps;
- **Section A3** shows that the least-squares regression on clipped covariates, run on any candidate set containing the true parents, produces slope estimates that are meaningful and that the associated Gram matrix remains well-conditioned uniformly;
- **Section A4** derives concentration inequalities for the empirical averages of the bounded observables used by the estimator, by decomposing the Hawkes process into clusters with independent roots, bounding the impact of each cluster on the statistics of interest, and combining these bounds into explicit tail probabilities that decay exponentially in the sample size;
- **Section A5** combines the theoretical screening gap and regression separation with the empirical concentration to show that, with probability tending to one, the two-stage estimator first includes all true parents in the screened superset and then recovers them exactly by thresholding the bounded least-squares coefficients.

A.1 Preliminary Lemmas

First of all we prove the uniform bounds for any fixed order moments of $X_j(0)$:

Proposition A.1. *For every $P \in \mathbb{N}$ there exists $\alpha_P > 0$, depending only on $(P, \beta, \mu_+, k, w_+)$, such that for all $0 < \alpha \leq \alpha_P$,*

$$\sup_{d\geq 1} \sup_{B\in\mathcal{B}_{d,k}} \sup_{1\leq j\leq d} \mathbb{E}[X_j(0)^p] < \infty, \qquad p = 1, \dots, P.$$

Proof. Fix $p \geq 1$ and j. For $M \geq 1$, let $\phi_M(x) = \min\{x, M\}$ and set $F_{p,M}(x) := \phi_M(x_j)^p$. Define the approximation $F_{p,M,\epsilon}$ using a smooth mollifier $\phi_{M,\epsilon}$ and apply Dynkin's formula in stationarity with the coherent infinitesimal generator of the process. Then letting the smoothing parameter tend to zero leads to

$$\beta p\, \mathbb{E}\Big[X_j \min\{X_j, M\}^{p-1} \mathbf{1}_{\{X_j<M\}}\Big] = \mathbb{E}\Big[\lambda_j\big(X(0)\big)\big(\min\{X_j+1, M\}^p - \min\{X_j, M\}^p\big)\Big].$$

[2]Screening alone could potentially be sufficient. However, since it would not change the result obtained for the scaling in d, investigating this possibility further is beyond the scope of this work.

Both sides are nonnegative and increase to their untruncated counterparts as $M \uparrow \infty$, so monotone convergence gives

$$\beta p\, \mathbb{E}[X_j^p] = \mathbb{E}\Big[\lambda_j\big(X(0)\big)\big((X_j+1)^p - X_j^p\big)\Big].$$

Now, since $(x+1)^p - x^p \le c_p(1+x^{p-1})$ for $x \ge 0$ and $\lambda_j\big(X(0)\big) \le \mu_+ + \theta_+ \sum_{\ell \in S_j,\, |S_j| \le k} X_\ell(0)$, we obtain from the previous equation

$$\beta p\, \mathbb{E}[X_j^p] \le c_p\mu_+\big(1 + \mathbb{E}[X_j^{p-1}]\big) + c_p\theta_+ \sum_{\ell \in S_j,\, |S_j| \le k} \mathbb{E}\big[X_\ell(1 + X_j^{p-1})\big]. \tag{7}$$

Using Young's inequality for the mixed term gives

$$\mathbb{E}\big[X_\ell(1 + X_j^{p-1})\big] \le \mathbb{E}[X_\ell] + \mathbb{E}[X_\ell^p] + \mathbb{E}[X_j^p] \le M_1 + M_p + M_p,$$

where, with the convention $M_0 := 1$, $M_p := \sup_{d,B,j} \mathbb{E}[X_j(0)^p]$. Therefore taking the supremum over (d, B, j) in (7), after enlarging the constant if necessary, gives:

$$\beta p\, M_p \le C_p\mu_+\big(1 + M_{p-1}\big) + C_p\beta\gamma\big(M_1 + M_p\big),$$

Equivalently, the recursion equation

$$\beta(p - C_p\gamma)\, M_p \le C_p\big(\mu_+ + \beta\gamma M_1 + \mu_+ M_{p-1}\big).$$

For $p = 1$ this yields $\beta(1 - 2C_p\gamma)M_1 \le 2C_p\mu_+$ and for α small enough this implies $M_1 < \infty$. Choosing then α small enough in order to have $\beta(p - C_p\gamma) > 0$ for $p = 1, \dots, P$ proves the claim by induction. □

Before establishing bounds on the covariance of $X_j(0)$, we prove the following lemma regarding the stationary mean.

Lemma A.1. *Under the stability condition, let $m := \mathbb{E}[X(0)]$ and $\bar\lambda := \mathbb{E}[\lambda(0)]$. Then*

$$m = (\beta I - \Theta)^{-1}\mu, \qquad \bar\lambda = \beta m.$$

Moreover, uniformly in d and $B \in \mathcal{B}_{d,k}$

$$\frac{\mu_-}{\beta} \le \min_j m_j \le \max_j m_j \le \frac{\mu_+}{\beta(1-\gamma)}, \tag{8}$$

$$\max_j \left| m_j - \frac{\mu_j}{\beta} \right| \le \frac{\gamma\mu_+}{\beta(1-\gamma)}. \tag{9}$$

Proof. Stationarity of the process leads to $0 = -\beta m + \bar\lambda$ and $\bar\lambda = \mu + \Theta m$, therefore $(\beta I - \Theta)m = \mu > 0$ and $\bar\lambda = \beta m$. The lower bound in (8) follows from $\beta m_j \ge \mu_j \ge \mu_-$, while for the upper bound:

$$\beta m_j = \mu_j + \sum_\ell \theta_{j\ell} m_\ell \le \mu_+ + \gamma\beta \max_\ell m_\ell,$$

especially $\beta(1-\gamma)\max_j m_j \le \mu_+$. Finally, exploiting the stability condition in the inequality,

$$\left| m_j - \frac{\mu_j}{\beta} \right| = \frac{1}{\beta} \sum_\ell \theta_{j\ell} m_\ell \le \gamma \max_\ell m_\ell,$$

whence (9) by the upper bound in (8). □

We are now in the position of proving bounds on the covariance of $X_j(0)$.

Proposition A.2. *There exists $\alpha_{\mathrm{cov}} > 0$, depending only on (β, μ_+, w_+, k), such that whenever $0 < \alpha \le \alpha_{\mathrm{cov}}$,*

$$\sup_{j \ne \ell} |\operatorname{Cov}(X_j(0), X_\ell(0))| \le C_{\mathrm{off}}\alpha, \qquad \sup_j \left| \operatorname{Var}(X_j(0)) - \frac{\mu_j}{2\beta} \right| \le C_{\mathrm{diag}}\alpha, \tag{10}$$

with constants depending only on $(\beta, \mu_-, \mu_+, w_+, k)$.

Proof. Set $m = \mathbb{E}[X(0)]$, $\bar{\lambda} = \mathbb{E}[\lambda(0)]$ and $Y_t = X_t - m$. Then

$$dY_t = (\Theta - \beta I)Y_t\, dt + dM_t, \qquad M_t := N_t - \int_0^t \lambda_s\, ds.$$

Denote $\Sigma := \mathrm{Cov}(X(0), X(0))$, $D := \sup_j \Sigma_{jj}$ and $S := \sup_{j\neq\ell} |\Sigma_{j\ell}|$: applying the semimartingale product rule to $Y_t Y_t^\top$ and taking expectations in stationarity yield the (entrywise-)Lyapunov equation

$$2\beta\Sigma_{j\ell} = \sum_r \theta_{jr}\Sigma_{r\ell} + \sum_r \theta_{\ell r}\Sigma_{jr} + \bar{\lambda}_j \mathbf{1}_{\{j=\ell\}}. \tag{11}$$

From the Lyapunov equation for $j = \ell$ and using $|\Sigma_{jr}| \le \sqrt{\Sigma_{jj}\Sigma_{rr}} \le D$ and $\sum_r \theta_{jr} \le \gamma\beta$, we get

$$2\beta\Sigma_{jj} \le 2\gamma\beta D + \bar{\lambda}_j.$$

Taking the supremum over j and using the previous Lemma

$$2\beta(1-\gamma)D \le \max_j \bar{\lambda}_j \le \frac{\mu_+}{1-\gamma} \qquad \Longrightarrow \qquad D \le \frac{\mu_+}{2\beta(1-\gamma)^2} =: C_D. \tag{12}$$

In a similar manner, if $j \neq \ell$, (11) yields $2\beta|\Sigma_{j\ell}| \le 2\gamma\beta D$, whence $S \le \gamma D \le (C_D k w_+/\beta)\,\alpha =: C_{\mathrm{off}}\alpha$ (which is the stated off-diagonal bound).
From the diagonal equation, subtracting $\mu_j/2\beta$ in both sides and taking the absolute value lead to

$$\left|\Sigma_{jj} - \frac{\mu_j}{2\beta}\right| \le \frac{1}{\beta}\sum_r \theta_{jr}|\Sigma_{jr}| + \frac{|\bar{\lambda}_j - \mu_j|}{2\beta}.$$

By (12), the first term in right side is at most $\gamma D = O(\alpha)$, while the second term is $O(\alpha)$ by the previous Lemma because $\bar{\lambda}_j = \beta m_j$. This proves the diagonal bound in (10). □

We conclude this section by showing that the screening scores effectively distinguish between the nodes $j \in S_i$ and $j \notin S_i$.

Proposition A.3. *Define*

$$G_{ij} := \mathrm{Cov}(X_j(0), \lambda_i(0)).$$

Assume $0 < \alpha \le \alpha_{\mathrm{cov}}$. *Then there exists* $C_G > 0$, *depending only on* $(\beta, \mu_-, \mu_+, w_+, k)$, *such that:*

$$\left|G_{ij} - \frac{\mu_j}{2\beta}\theta_{ij}\right| \le C_G\alpha^2 \quad \textit{if } j \in S_i \qquad \textit{and} \qquad |G_{ij}| \le C_G\alpha^2 \quad \textit{if } j \notin S_i$$

Consequently there exists $\alpha_{scr} \in (0, \alpha_{\mathrm{cov}}]$ *such that for every* $0 < \alpha \le \alpha_{scr}$,

$$\min_{j\in S_i} G_{ij} - \max_{j\notin S_i} G_{ij} \ge \Delta_{scr}(\alpha) := \frac{\mu_- w_-}{4\beta}\alpha \tag{13}$$

for every row i *with* $S_i \neq \varnothing$, *while* $G_{ij} = 0$ *for all* j *whenever* $S_i = \varnothing$.

Proof. From the definition of $\lambda_i(0)$, if $j \in S_i$, then

$$G_{ij} = \theta_{ij}\Sigma_{jj} + \sum_{\ell\in S_i\setminus\{j\}} \theta_{i\ell}\Sigma_{j\ell}.$$

Hence using Proposition A.2 and $\theta_{ij} \le \theta_+$:

$$\left|G_{ij} - \frac{\mu_j}{2\beta}\theta_{ij}\right| \le \theta_{ij}\left|\Sigma_{jj} - \frac{\mu_j}{2\beta}\right| + \sum_{\ell\in S_i\setminus\{j\}} \theta_{i\ell}|\Sigma_{j\ell}| \le C_G\alpha^2. \tag{14}$$

If $j \notin S_i$, then

$$|G_{ij}| = \left|\sum_{\ell\in S_i} \theta_{i\ell}\Sigma_{j\ell}\right| \le k\theta_+ \sup_{j\neq\ell} |\Sigma_{j\ell}| \le C_G\alpha^2. \tag{15}$$

Finally, using (14) and (15), for α small enough the gap (13) follows. □

A.2 Clipping and binning

Before we show that the screening signal survives clipping, we need to quantify how much the covariances change when we replace the unbounded covariates $X_j(0)$ by their clipped versions $Z_j^{(R)} = \min\{X_j(0), R\}$. Intuitively, clipping only affects the tails, i.e. the rare events where $X_j(0)$ is very large, but since these events become rarer as R grows, the "distortion" introduced should vanish as R$\to \infty$. The following lemma makes this precise, showing that the covariances change by at most C_{clip}/R.

We fix $h > 0$ and $R > 0$ and denote $Z_j^{(R)} := \psi_R(X_j(0))$ and $F_{ij}^{(h,R)} := \text{Cov}(Z_j^{(R)}, Y_{i,0}^{(h)})$.

Lemma A.2. *There exist a constant $C_{\text{clip}} > 0$, depending only on $(\beta, \mu_-, \mu_+, w_+, k)$, such that uniformly in d, B, i, j, ℓ,*

$$|\text{Cov}(Z_j^{(R)}, \lambda_i(0)) - \text{Cov}(X_j(0), \lambda_i(0))| \le \frac{C_{\text{clip}}}{R}, \tag{16}$$

$$|\text{Cov}(Z_j^{(R)}, Z_\ell^{(R)}) - \text{Cov}(X_j(0), X_\ell(0))| \le \frac{C_{\text{clip}}}{R}. \tag{17}$$

Proof. Writing $\delta_j^{(R)} := X_j(0) - Z_j^{(R)} = (X_j(0) - R)_+$,

$$|\text{Cov}(Z_j^{(R)}, \lambda_i) - \text{Cov}(X_j, \lambda_i)| = |\text{Cov}(\delta_j^{(R)}, \lambda_i)| \le \mathbb{E}[\delta_j^{(R)}\lambda_i] + \mathbb{E}[\delta_j^{(R)}]\mathbb{E}[\lambda_i].$$

Because $\delta_j^{(R)} \le X_j \mathbf{1}_{\{X_j > R\}}$, Cauchy–Schwarz and Markov give

$$\mathbb{E}[\delta_j^{(R)}] \le \sqrt{\mathbb{E}[X_j^2]\mathbb{P}(X_j > R)} \le \frac{\sqrt{\mathbb{E}[X_j^2]\mathbb{E}[X_j^2]}}{R} \le \frac{C_{\text{clip}}}{R}.$$

Moreover, for every $x \ge 0$, $(x-R)_+^2 \le x^2 \mathbf{1}\{x > R\} \le \frac{x^4}{R^2}$. Proposition A.1 with $P = 4$ yields

$$\mathbb{E}[(\delta_j^{(R)})^2] \le \frac{\mathbb{E}[X_j^4]}{R^2} \le \frac{C_{\text{clip}}}{R^2}. \tag{18}$$

Since $\lambda_i \le \mu_+ + \alpha w_+ \sum_{\ell \in S_i} X_\ell$ and $|S_i| \le k$, Proposition A.1 with $P = 4$ also implies $\sup_{i,d,B} \mathbb{E}[\lambda_i^2] < \infty$. Therefore

$$\mathbb{E}[\delta_j^{(R)}\lambda_i] \le \sqrt{\mathbb{E}[(\delta_j^{(R)})^2]\mathbb{E}[\lambda_i^2]} \le \frac{C_{\text{clip}}}{R}.$$

This proves (16). The proof of (17) is identical after replacing λ_i by X_ℓ. □

Proposition A.4. *Define*

$$H_{ij}^{(R)} := \text{Cov}\big(Z_j^{(R)}, \lambda_i(0)\big).$$

There exist $\alpha_1 > 0$, $R_1 : (0, \alpha_1] \to [1, \infty)$, and constants $c_H, C_H > 0$, depending only on $(\beta, \mu_-, \mu_+, w_-, w_+, k)$, such that for every fixed $0 < \alpha \le \alpha_1$ and every $R \ge R_1(\alpha)$,

$$\min_{j \in S_i} H_{ij}^{(R)} \ge c_H \alpha, \tag{19}$$

$$\max_{j \notin S_i} H_{ij}^{(R)} \le C_H \alpha^2 + \frac{C_H}{R}, \tag{20}$$

for every row i, and therefore

$$\min_{j \in S_i} H_{ij}^{(R)} - \max_{j \notin S_i} H_{ij}^{(R)} \ge \frac{\mu_- w_-}{8\beta}\alpha. \tag{21}$$

And $H_{ij}^{(R)} = 0$ for all j whenever $S_i = \varnothing$.

Proof. By Lemma A.2 and Proposition A.3, $H_{ij}^{(R)} = G_{ij} + O(R^{-1})$ uniformly in (i, j, d, B). Using again the bounds in Proposition A.3, choose α_1 such that $C_G \alpha \le \mu_- w_-/(16\beta)$ for $0 < \alpha \le \alpha_1$, and then choose $R_1(\alpha)$ such that $C_{\text{clip}}/R \le \mu_- w_- \alpha/(16\beta)$ for all $R \ge R_1(\alpha)$. Using these bounds proves (19)–(21). □

Since we are using discrete bin indicators $Y_{i,0}^{(h)}$, which record whether at least one event of type-i occurred in the first bin, the next Lemmas show that the covariance between $Z_j^{(R)}$ and $Y_{i,0}^{(h)}$ is well approximated by h times the covariance between $Z_j^{(R)}$ and $\lambda_i(0)$, up to an error of order Rh^2. Intuitively this is true because, over a short bin of width h, the probability of observing at least one event is approximately $h\lambda_i(0)$, so the indicator $Y_{i,0}^{(h)}$ behaves roughly like $h\lambda_i(0)$. We first present the following auxiliary Lemma about the approximation on one bin.

Lemma A.3. *Then there exists a constant $C_{\mathrm{bin}} > 0$, depending only on (β, μ_+, w_+, k), such that uniformly in d, B, i, j,*

$$|F_{ij}^{(h,R)} - hH_{ij}^{(R)}| \le C_{\mathrm{bin}} R h^2 \qquad (0 < h \le 1,\ R \ge 1). \tag{22}$$

Proof. Fix (i, j) and write $Z := \psi_R(X_j(0))$, $N_i^{(h)} := N_i((0, h])$ and $\Lambda_i^{(h)} := \int_0^h \lambda_i(t)\, dt$ (note that Z is $\mathcal{F}_0$-measurable). By the compensator identity for N_i and using $\mathbf{1}\{N_i((0,t)) \ge 1\} \le N_i((0,t])$ for the inequality, we have

$$\mathbb{E}\big[N_i^{(h)} - Y_{i,0}^{(h)}\big] = \mathbb{E}\int_0^h \mathbf{1}\{N_i((0,t)) \ge 1\}\lambda_i(t)\, dt \le \int_0^h \mathbb{E}\big[N_i((0,t])\lambda_i(t)\big] dt. \tag{23}$$

Now, for $t \le 1$, using $X_\ell(t) \le X_\ell(0) + N_\ell((0,t])$

$$\lambda_i(t) \le \mu_+ + \theta_+ \sum_{\ell \in S_i} X_\ell(t) \le C_1 \left(1 + \sum_{\ell \in S_i} X_\ell(0) + \sum_{\ell \in S_i} N_\ell((0,t])\right),$$

Hence the expectation in the previous integral satisfies

$$\begin{aligned}\mathbb{E}\big[N_i((0,t])\lambda_i(t)\big] \le\ & C_1 \mathbb{E}[N_i((0,t])] + C_1 \sum_{\ell \in S_i} \mathbb{E}\big[N_i((0,t])X_\ell(0)\big] \\ & + C_1 \sum_{\ell \in S_i} \mathbb{E}\big[N_i((0,t])N_\ell((0,t])\big].\end{aligned} \tag{24}$$

The first term satisfies $\mathbb{E}[N_i((0,t])] \le C_2 t$ by compensator identity and Proposition A.1. For the mixed term, using the compensator identity multiple times, $\lambda_i(s) \le C_3(1 + \sum_{m \in S_i} X_m(s))$, Cauchy Schwarz, stationarity, and Proposition A.1 with $P = 2$, it holds $\mathbb{E}[N_i((0,t])X_\ell(0)] \le C_3 t$. For the last term, we use the martingale decomposition for $N_q((0,t])$ (for any q), Itô isometry, and stationarity to state that M_q is a square-integrable martingale, and then for $t \le 1$ it holds

$$\mathbb{E}\big[N_q((0,t])^2\big] \le 2\mathbb{E}[M_q(t)^2] + 2\mathbb{E}\left[\left(\int_0^t \lambda_q(s)\, ds\right)^2\right] \le C_4 t \qquad \text{uniformly in } (q, d, B),$$

where again the elementary inequality $(x+y)^2 \le C(x^2 + y^2)$, the Itô isometry, Cauchy Schwarz for integrals, and Proposition A.1 with $P = 2$ have been used. Then it is easy to bound the term $\mathbb{E}\big[N_i((0,t])N_\ell((0,t])\big]$ by Cauchy Schwarz. Hence, using $|S_i| \le k$, combining all the bounds obtained above for the right-hand terms in (24) and defining $C_{bin,1}$ as a convenient combination of the constants in (24) lead to $\mathbb{E}\big[N_i((0,t])\lambda_i(t)\big] \le C_{bin,1} t$, and substituting this bound in (23) gives:

$$\mathbb{E}\big[N_i^{(h)} - Y_{i,0}^{(h)}\big] \le \int_0^h \mathbb{E}\big[N_i((0,t])\lambda_i(t)\big] dt \le \frac{C_{bin,1}}{2} h^2$$

Exploiting the facts that $0 \le Z \le R$ and that $\mathrm{Cov}(Z, N_i^{(h)}) = \mathrm{Cov}(Z, \Lambda_i^{(h)})$ (this follows from substituting the martingale decomposition in the expectations, because Z is $\mathcal{F}_0$-measurable and $N_i(t) - \int_0^t \lambda_i(s)ds$ is an $(\mathcal{F}_t)$-martingale), then it is possible to show the following:

$$|\mathrm{Cov}(Z, Y_{i,0}^{(h)}) - \mathrm{Cov}(Z, N_i^{(h)})| \le 2\|Z\|_\infty \mathbb{E}\big[N_i^{(h)} - Y_{i,0}^{(h)}\big] \le C_{bin,1} R h^2. \tag{25}$$

$$\mathrm{Cov}(Z, N_i^{(h)}) - h\,\mathrm{Cov}(Z, \lambda_i(0)) = \mathrm{Cov}(Z, \Lambda_i^{(h)}) - h\,\mathrm{Cov}(Z, \lambda_i(0)) = \int_0^h \mathrm{Cov}\big(Z, \lambda_i(t) - \lambda_i(0)\big) dt.$$

As before, it is easy to show that $|\mathrm{Cov}\big(Z, \lambda_i(t) - \lambda_i(0)\big)| \le 2R\,\mathbb{E}|\lambda_i(t) - \lambda_i(0)|$. Now:

$$\lambda_i(t) - \lambda_i(0) = \sum_{\ell \in S_i} \theta_{i\ell}(X_\ell(t) - X_\ell(0)),$$

Writing down the integral form of SDE for $X_\ell(t)$, taking the absolute value, and using stationarity and Lemma A.1, we get

$$\mathbb{E}|X_\ell(t) - X_\ell(0)| \le \beta \int_0^t \mathbb{E}[X_\ell(s)]ds + \mathbb{E}[N_\ell(t) - N_\ell(0)] \le C_5 t.$$

Therefore, we can define conveniently $C_{bin,2}$ such that $\mathbb{E}|\lambda_i(t) - \lambda_i(0)| \le C_{bin,2}t$ uniformly in (i, d, B), and thus

$$|\mathrm{Cov}(Z, \Lambda_i^{(h)}) - hH_{ij}^{(R)}| \le 2R \int_0^h C_{bin,2}t\,dt \le C_{bin,2}Rh^2.$$

Combining this with (25) and defining $C_{bin} := C_{bin,1} + C_{bin,2}$ proves (22). □

Lemma A.3, together with the screening gap already established for the clipped covariances, yields Proposition A.5 below: the separation between true parents and non-parents is preserved by the observable quantities used by the estimator, provided R is sufficiently large and h is sufficiently small relative to α.

Proposition A.5. *There exist $\alpha_2 > 0$, a function $R_2 : (0, \alpha_2] \to [1, \infty)$, and a constant $c_{\mathrm{scr}} > 0$ depending only on $(\beta, \mu_\pm, w_\pm, k)$, such that for every fixed $0 < \alpha \le \alpha_2$, every $R \ge R_2(\alpha)$, and every*

$$0 < h \le \min\Big\{\frac{\mu_- w_-}{32\beta C_{\mathrm{bin}}}\frac{\alpha}{R}, 1\Big\},$$

one has, for every row i with $S_i \neq \varnothing$,

$$\min_{j \in S_i} F_{ij}^{(h,R)} - \max_{j \notin S_i} F_{ij}^{(h,R)} \ge c_{\mathrm{scr}}\alpha h, \tag{26}$$

while $F_{ij}^{(h,R)} = 0$ for all j whenever $S_i = \varnothing$.

Proof. If $S_i = \varnothing$, then $\lambda_i(t) \equiv \mu_i$ and $Y_{i,0}^{(h)}$ is independent of $X(0)$, so $F_{ij}^{(h,R)} = 0$ for all j. Assume $S_i \neq \varnothing$: by Lemma A.3, $F_{ij}^{(h,R)} = hH_{ij}^{(R)} + O(Rh^2)$ and by Proposition A.4, uniformly in i,

$$\min_{j \in S_i} F_{ij}^{(h,R)} - \max_{j \notin S_i} F_{ij}^{(h,R)} \ge h \cdot \frac{\mu_- w_-}{8\beta}\alpha - 2C_{\mathrm{bin}}Rh^2.$$

Choose $\alpha_2 \le \alpha_1$ from Proposition A.4, and then choose $R_2(\alpha) \ge R_1(\alpha)$. If $R \ge R_2(\alpha)$ and

$$0 < h \le \min\Big\{\frac{\mu_- w_-}{32\beta C_{\mathrm{bin}}}\frac{\alpha}{R}, 1\Big\},$$

then

$$C_{\mathrm{bin}}Rh \le \frac{\mu_- w_-}{32\beta}\alpha,$$

so combining this with the previous inequality proves (26). □

A.3 Bounded local regression

Fix $R \ge 1$ and a candidate set $C \subset \{1, \dots, d\}$ with $|C| \le m$. Define:

$$\Sigma_C^{(R)} := \mathrm{Cov}\big(Z_C^{(R)}(0), Z_C^{(R)}(0)\big), \qquad Z_C^{(R)}(0) := (Z_j^{(R)})_{j \in C}, \qquad g_{i,C}^{(h,R)} := \mathrm{Cov}\big(Z_C^{(R)}(0), Y_{i,0}^{(h)}\big).$$

Before stating the uniform nondegeneracy of the clipped Gram matrix $\Sigma_C^{(R)}$, we define, whenever $\Sigma_C^{(R)}$ is invertible, the centered theoretical least-squares slope as

$$y_{i,C}^{(h,R)} := \big(\Sigma_C^{(R)}\big)^{-1} g_{i,C}^{(h,R)} \tag{27}$$

Lemma A.4. *There exist $\alpha_3 > 0$, $R_3 : (0, \alpha_3] \to [1, \infty)$, and $\kappa > 0$, depending only on $(\beta, \mu_-, \mu_+, w_+, k, m)$, such that for every fixed $0 < \alpha \le \alpha_3$, every $R \ge R_3(\alpha)$, every d, every $B \in \mathcal{B}_{d,k}$, and every $C \subset \{1, \dots, d\}$ with $|C| \le m$, we have $\lambda_{\min}(\Sigma_C^{(R)}) \ge \kappa$.*

Proof. Under $\alpha = 0$ the coordinates are independent shot-noise processes. Hence, for any fixed C:

$$\Sigma_{C,0}^{(R)} = \operatorname{diag}\Big(\operatorname{Var}_0\big(\psi_R(X_j(0))\big)\Big)_{j \in C}.$$

Now $\operatorname{Var}_0(X_j(0)) = \mu_j/(2\beta) \ge \mu_-/(2\beta)$, then Lemma A.2 with $i = j$ and $\ell = j$ and choosing R large enough gives $\operatorname{Var}_0\big(\psi_R(X_j(0))\big) \ge \mu_-/4\beta$ for all j . Thus $\lambda_{\min}(\Sigma_{C,0}^{(R)}) \ge \mu_-/4\beta$ for all $|C| \le m$. Next compare $\Sigma_C^{(R)}$ to $\Sigma_{C,0}^{(R)}$. Fix $j, \ell \in C$. By adding and subtracting the unclipped covariances at couplings α and 0,

$$\begin{aligned}
&\operatorname{Cov}_\alpha\big(Z_j^{(R)}, Z_\ell^{(R)}\big) - \operatorname{Cov}_0\big(Z_j^{(R)}, Z_\ell^{(R)}\big) \\
&\quad = \Big[\operatorname{Cov}_\alpha\big(Z_j^{(R)}, Z_\ell^{(R)}\big) - \operatorname{Cov}_\alpha(X_j, X_\ell)\Big] + \Big[\operatorname{Cov}_\alpha(X_j, X_\ell) - \operatorname{Cov}_0(X_j, X_\ell)\Big] \\
&\qquad + \Big[\operatorname{Cov}_0(X_j, X_\ell) - \operatorname{Cov}_0\big(Z_j^{(R)}, Z_\ell^{(R)}\big)\Big].
\end{aligned}$$

The first and third brackets are $O(R^{-1})$ uniformly by Lemma A.2. For the middle bracket, if $j = \ell$ then Proposition A.2 gives

$$|\operatorname{Var}_\alpha(X_j) - \operatorname{Var}_0(X_j)| \le C\alpha,$$

while if $j \ne \ell$ then independence at $\alpha = 0$ gives $\operatorname{Cov}_0(X_j, X_\ell) = 0$ and Proposition A.2 yields

$$|\operatorname{Cov}_\alpha(X_j, X_\ell) - \operatorname{Cov}_0(X_j, X_\ell)| = |\operatorname{Cov}_\alpha(X_j, X_\ell)| \le C\alpha.$$

Hence every entry of $\Sigma_C^{(R)} - \Sigma_{C,0}^{(R)}$ is $O(\alpha) + O(R^{-1})$, uniformly over all (d, B, C). Recalling the bound for operator norm given by the Frobenius norm, since the matrix dimension is at most m, it holds that

$$\sup_{|C| \le m} \|\Sigma_C^{(R)} - \Sigma_{C,0}^{(R)}\|_{\mathrm{op}} \le C_m\big(\alpha + R^{-1}\big).$$

We take $\alpha_3 > 0$ small and then $R_3(\alpha)$ large enough so that the right-hand side is at most $\mu_-/(8\beta)$. Weyl's inequality and bound $\lambda_{\min}(\Sigma_{C,0}^{(R)}) \ge \mu_-/4\beta$ yield the desired claim with $\kappa := \mu_-/8\beta$ □

The next proposition shows how the theoretical OLS slope computed on any candidate set $C \supset S_i$ of bounded size is close to the rescaled true coefficient vector $\theta_{i,C}$ up to small errors from clipping and binning and that this approximation is tight enough to correctly separate true parents from non-parents:

Proposition A.6. *There exist $\alpha_4 > 0$, a function $R_4 : (0, \alpha_4] \to [1, \infty)$, and a constant $C_y > 0$ depending only on $(\beta, \mu_\pm, w_\pm, k, m)$ such that for every fixed $0 < \alpha \le \alpha_4$, every $R \ge R_4(\alpha)$, and every*

$$0 < h \le \min\left\{\frac{w_-}{8C_y}\frac{\alpha}{R}, 1\right\},$$

one has, for every row i and every candidate set $C \supset S_i$ with $|C| \le m$,

$$\|y_{i,C}^{(h,R)} - h\theta_{i,C}\|_\infty \le C_y\left(\frac{\alpha h}{R} + Rh^2\right). \tag{28}$$

In particular,

$$\min_{j \in S_i} y_{ij,C}^{(h,R)} \ge \frac{3}{4}\theta_- h, \qquad \max_{j \in C \setminus S_i} |y_{ij,C}^{(h,R)}| \le \frac{1}{4}\theta_- h. \tag{29}$$

Proof. Fix (i, C). Recalling the definition of $g_{i,C}^{(h,R)}$, Lemma A.3 applied coordinatewise gives

$$g_{i,C}^{(h,R)} = h\operatorname{Cov}\big(Z_C^{(R)}(0), \lambda_i(0)\big) + r_{i,C}^{(h,R)}, \qquad \|r_{i,C}^{(h,R)}\|_\infty \le CRh^2. \tag{30}$$

Now, because $C \supset S_i$, using the definition for $\lambda_i(0)$:

$$\mathrm{Cov}\big(Z_C^{(R)}, \lambda_i\big) = \sum_{\ell \in C} \theta_{i\ell}\, \mathrm{Cov}\big(Z_C^{(R)}, X_\ell\big).$$

Writing $X_\ell = Z_\ell^{(R)} + \delta_\ell^{(R)}$ with $\delta_\ell^{(R)} = (X_\ell - R)_+$ gives $\mathrm{Cov}\big(Z_C^{(R)}, \lambda_i\big) = \Sigma_C^{(R)} \theta_{i,C} + q_{i,C}^{(R)}$, where $q_{i,C}^{(R)} := \sum_{\ell \in C} \theta_{i\ell}\, \mathrm{Cov}\big(Z_C^{(R)}, \delta_\ell^{(R)}\big)$, then

$$g_{i,C}^{(h,R)} = h\Sigma_C^{(R)} \theta_{i,C} + h q_{i,C}^{(R)} + r_{i,C}^{(h,R)}. \tag{31}$$

To bound $q_{i,C}^{(R)}$, fix $j, \ell \in C$. Since $0 \le Z_j^{(R)} \le X_j$ and $\delta_\ell^{(R)} = (X_\ell - R)_+$,

$$|\mathrm{Cov}\big(Z_j^{(R)}, \delta_\ell^{(R)}\big)| \le \mathbb{E}\big[Z_j^{(R)} \delta_\ell^{(R)}\big] + \mathbb{E}[Z_j^{(R)}]\mathbb{E}[\delta_\ell^{(R)}] \le \mathbb{E}\big[X_j \delta_\ell^{(R)}\big] + \mathbb{E}[X_j]\mathbb{E}[\delta_\ell^{(R)}].$$

By Cauchy–Schwarz, Proposition A.1 with $P = 4$ and (18),

$$\mathbb{E}\big[X_j \delta_\ell^{(R)}\big] \le \sqrt{\mathbb{E}[X_j^2]\mathbb{E}[(\delta_\ell^{(R)})^2]} \le \frac{C}{R}, \qquad \mathbb{E}[X_j]\mathbb{E}[\delta_\ell^{(R)}] \le \frac{C}{R}.$$

Hence each coordinate of $\mathrm{Cov}(Z_C^{(R)}, \delta_\ell^{(R)})$ is $O(R^{-1})$, uniformly in (d, B, i, C, ℓ). Since $|C| \le m$ and $\theta_{i\ell} \le \alpha w_+$,

$$\|q_{i,C}^{(R)}\|_\infty \le C\alpha / R.$$

Now, rewriting (27) using (31), gives

$$y_{i,C}^{(h,R)} - h\theta_{i,C} = \big(\Sigma_C^{(R)}\big)^{-1}\Big(h q_{i,C}^{(R)} + r_{i,C}^{(h,R)}\Big).$$

Therefore, using Lemma A.4 implication $\|(\Sigma_C^{(R)})^{-1}\|_{\mathrm{op}} \le \kappa^{-1}$ and the obtained bound for $\|q_{i,C}^{(R)}\|_\infty$ gives

$$\|y_{i,C}^{(h,R)} - h\theta_{i,C}\|_\infty \le C_y \left(\frac{\alpha h}{R} + Rh^2\right)$$

for a constant C_y depending only on the model parameters. Choose $\alpha_4 \le \alpha_3$ and then choose $R_4(\alpha) \ge \max\{R_3(\alpha), 8C_y/w_-\}$ and $0 < h \le \min\{1, \theta_-/(8C_y R)\}$. If $R \ge R_4(\alpha)$ then

$$C_y \frac{\alpha h}{R} \le \frac{1}{8}\theta_- h, \qquad C_y R h^2 \le \frac{1}{8}\theta_- h,$$

so the right-hand side of (28) is at most $h\theta_-/4$. Since $\theta_{ij} \ge \theta_-$ on S_i and $\theta_{ij} = 0$ on $C \setminus S_i$, (29) follows. □

A.4 Bounded-block concentration via Poisson clusters

First of all, set $0 < h \le h_0 := 1$ and recall $\gamma < 1$. The observables we are focusing on are the following four:

$$Z_{j,r}^{(R)}, \qquad Y_{i,r}^{(h)}, \qquad Z_{j,r}^{(R)} Y_{i,r}^{(h)}, \qquad Z_{a,r}^{(R)} Z_{b,r}^{(R)} \tag{32}$$

Now, we introduce the terminology used in the context of the cluster representation of Hawkes processes:

- **Root**: these are events that arrive according to a independent Poisson process with background intensity μ_v. They are "exogenous" because they are not caused by previous events;
- **Type-j**: since the process X has d components indexed by $j = 1, \dots, d$, each event in the system carries a label j indicating which component it belongs to. This label is called its "type";
- **Offspring**: each root can trigger direct offspring, which in turn can trigger their own offspring (called multiple generations). Thus, each cluster $\mathcal{C}$ has an independent root and each event of type-j produces type-i children according to an independent Poisson point process on $(0, \infty)$ with intensity $\theta_{ij} \exp\{-\beta t\}\, dt$. Therefore the mean offspring matrix is $K := \Theta/\beta$ and consequently $\|K\|_\infty \le \gamma < 1$;
- **Progeny**: while offspring refers only to the events that are immediately triggered by one specific root, the progeny of one root includes every single event in the cluster, across all generations;

- **Age**: in the cluster representation, each cluster has a root event that is "born" at some time t, namely the time at which the root arrives. The age of any event in the cluster is the time that passed since the root was born. Thus if an event occurs at time s, its age is $s - t$. Hence, "age" is not absolute time, but rather time measured relative to the birth time of the root of the cluster to which the event belongs.

Now, set $a > 1$ s.t. $a\gamma < 1$, fix a finite coordinate set $J \subset [d]$ with $|J| \le 2$ and denote the type-j events at time t in a cluster $\mathcal{C}$ as $(j,t) \in \mathcal{C}$. Define the type weights $q_v^{(a)}$ as

$$q_v^{(a)} := \sum_{\ell=0}^{\infty} a^\ell \sum_{j \in J} (K^\ell)_{jv}, \qquad v = 1, \dots, d. \tag{33}$$

Note that $q_v^{(a)} = \mathbf{1}\{v \in J\} + a \sum_{i=1}^d K_{iv} q_i^{(a)}$, thus $q_v^{(a)} \ge 1$ for every $v \in J$. Define also:

$$a_h(t) := \sum_{r=0}^{\infty} e^{-\beta(rh-t)} \mathbf{1}\{t \le rh\} \le C_h \times \begin{cases} e^{\beta t}, & t < 0, \\ 1, & t \ge 0, \end{cases} \qquad C_h := \frac{e^{\beta h}}{1 - e^{-\beta h}}, \tag{34}$$

where the inequality follows from a simple calculation. Lastly, the weighted local cluster size is:

$$H_J(\mathcal{C}) := \sum_{(j,t) \in \mathcal{C},\, j \in J} a_h(t) + \sum_{(j,t) \in \mathcal{C},\, j \in J} \mathbf{1}\{t \ge 0\}. \tag{35}$$

Because $q_j^{(a)} \ge 1$ on J and $a_h(t) \le C_h$, defining $W_J^{(a)}(\mathcal{C}) := \sum_{(j,t) \in \mathcal{C}} q_j^{(a)}$, it holds

$$H_J(\mathcal{C}) \le (C_h + 1) \sum_{(j,t) \in \mathcal{C},\, j \in J} 1 \le (C_h + 1) W_J^{(a)}(\mathcal{C}). \tag{36}$$

The following lemma shows that the exponential moment of the weighted cluster size is uniformly bounded, regardless of the dimension d and the network configuration.

Lemma A.5. *There exist constants $\eta_A > 0$ and $C_A > 0$, depending only on (h, R, β, γ), such that for every d, every $B \in \mathcal{B}_{d,k}$, every $J \subset [d]$ with $|J| \le 2$, and every root type v,*

$$\mathbb{E}_v\big[e^{\eta_A H_J(\mathcal{C})}\big] \le C_A. \tag{37}$$

Proof. Choose $a \in (1, \gamma^{-1})$ and abbreviate $q_v := q_v^{(a)}$. By (36), it suffices to prove the Lemma for $W_J^{(a)}(\mathcal{C})$. First of all note that:

$$0 \le q_v^{(a)} \le \sum_{\ell=0}^{\infty} a^\ell \sum_{j \in J} \|K\|_\infty^\ell \le |J| \sum_{\ell=0}^{\infty} (a\gamma)^\ell = \frac{|J|}{1 - a\gamma} \le \frac{2}{1 - a\gamma} =: Q. \tag{38}$$

Let $W_{J,v}^{(a)}$ denote the weighted total progeny when the root has type v, and let $W_{J,v}^{(a,m)}$ be the same quantity truncated at generation depth at most $m \ge 0$. Similarly, if root v has M_i type-i offspring (note that M_i is Poisson with parameter K_{iv}), denote $W_{J,i,k}^{(a,m)}$ the truncated weighted total progeny generated by type-i offspring. Then defining $F_v^{(m)}(\eta) := \mathbb{E}_v\big[e^{\eta W_{J,v}^{(a,m)}}\big]$,

$$\begin{aligned} F_v^{(m+1)}(\eta) &= \mathbb{E}_v\left[e^{\eta\left(q_v + \sum_{i=1}^d \sum_{k=1}^{M_i} W_{J,i,k}^{(a,m)}\right)}\right] = e^{\eta q_v} \prod_{i=1}^d \mathbb{E}_v\left[\mathbb{E}\left[\prod_{k=1}^{M_i} e^{\eta W_{J,i,k}^{(a,m)}} \middle| M_i = m_i\right]\right] = \\ &= e^{\eta q_v} \prod_{i=1}^d \mathbb{E}_v\left[\left(F_i^{(m)}(\eta)\right)^{M_i}\right] = e^{\eta q_v} \prod_{i=1}^d e^{-K_{iv}} \sum_{m_i=0}^{\infty} \frac{\left(F_i^{(m)} K_{iv}\right)^{m_i}}{m_i!} = \\ &= \exp\left(\eta q_v + \sum_{i=1}^d K_{iv}\big(F_i^{(m)}(\eta) - 1\big)\right), \end{aligned}$$

where $F_v^{(0)}(\eta) = e^{\eta q_v}$. Set $A := 2/(1 - a^{-1}) > 1$, then $A(1 - a^{-1}) > 1$ and $\exists x_0 > 0$ depending only on a, such that

$$e^x - 1 \le A \frac{x}{1 + A/a} \qquad \text{for every } 0 \le x \le x_0. \tag{39}$$

Now choose $\eta_A > 0$ such that $\eta_A Q(1 + A/a) \le x_0$. This choice, for $m = 0$ and (38), implies $\eta_A q_v \le x_0$ and therefore $F_v^{(0)}(\eta_A) - 1 = e^{\eta_A q_v} - 1 \le A\eta_A q_v$. Assume that the same inequality holds for a fixed $m > 0$ and $\forall v$ as inductive step; using the obtained expression for $F_v^{(m+1)}(\eta)$, $\sum_{i=1}^{d} K_{iv} q_i \le q_v/a$ and the same reasoning used for the zero-step, it holds

$$F_v^{(m+1)}(\eta_A) \le \exp\left(\eta_A q_v + A\eta_A \sum_{i=1}^{d} K_{iv} q_i\right) \le \exp\left(\eta_A q_v \left(1 + \frac{A}{a}\right)\right) \le 1 + A\eta_A q_v,$$

concluding the induction. Since $W_{J,v}^{(a,m)} \uparrow W_{J,v}^{(a)}$ almost surely as $m \to \infty$, monotone convergence yields $\mathbb{E}_v\left[e^{\eta_A W_J^{(a)}(\mathcal{C})}\right] \le 1 + A\eta_A Q =: C'_A < \infty$, uniformly in (d, B, J, v). Combining with (36) and shrinking η_A by the factor $(C_h + 1)$ and renaming it proves (37). □

The proof of Lemma A.5 gives a little more than the uniform bound stated there: monotone convergence in previous proof's inductive step yields

$$\mathbb{E}_v\left[e^{\eta_A W_{J,v}^{(a)}} - 1\right] \le A\eta_A q_v^{(a)}. \tag{40}$$

Moreover, by the row-sum bound,

$$\sum_{v=1}^{d} \mu_v q_v^{(a)} \le \mu_+ \sum_{\ell=0}^{\infty} a^\ell \sum_{j\in J} \sum_{v=1}^{d} (K^\ell)_{jv} \le \mu_+ \sum_{\ell=0}^{\infty} a^\ell \sum_{j\in J} \|K\|_\infty^\ell \le 2\mu_+ \sum_{\ell=0}^{\infty} (a\gamma)^\ell \le C_q. \tag{41}$$

Consequently, for every integer $m \ge 1$ there exists $\eta_m > 0$ such that

$$\sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\left[(W_{J,v}^{(a)})^m e^{\eta_m W_{J,v}^{(a)}}\right] \le C_m, \tag{42}$$

where $C_m > 0$ depends only on $(m, \beta, \mu_+, k, w_+, \gamma)$. Indeed, choose $\eta_m > 0$ such that $2\eta_m \le \eta_A$ and use the pointwise bound $x^m e^{\eta_m x} \le C_{m,\eta_m}(e^{\eta_A x} - 1)$, valid for every $m \ge 1$, together with (40) and (41). For $m = 0$ one simply uses (37) directly. We now use this result to prove the following lemma, which controls how long a cluster keeps producing events of the types we are interested in (for example J). Specifically, it shows that the probability of a cluster still having type-J events at a late age u decays exponentially in u, uniformly over all dimensions and all network configurations.

Lemma A.6. *Fix $J \subset [d]$ with $|J| \le 2$. For one cluster started from a root of type v at time 0, let*

$$L_J := \sup\{u \ge 0 : \textit{ there exists a type-}j \in J \textit{ event in the full cluster at age } u\},$$

with the convention $\sup \varnothing = 0$. Thus the root itself is counted at age 0 when $v \in J$. Then there exist constants $a \in (1, \gamma^{-1})$, $c_L > 0$, and $C_L > 0$, depending only on (β, k, w_+, γ), such that for every $u \ge 0$,

$$\mathbb{P}_v(L_J \ge u) \le q_v^{(a)} e^{-c_L u}, \tag{43}$$

$$\sum_{v=1}^{d} \mu_v \mathbb{P}_v(L_J \ge u) \le C_L e^{-c_L u}. \tag{44}$$

Proof. Set $c_L := \beta(1 - a^{-1}) > 0$. Let $N_J([u, \infty))$ denote the number of type-J events in the full cluster whose ages lie in $[u, \infty)$, again counting the root when applicable. Then

$$\mathbb{P}_v(L_J \ge u) = \mathbb{E}_v[\mathbf{1}\{L_J \ge u\}] \le \mathbb{E}_v\left[N_J([u, \infty))\right].$$

For $u = 0$, the root contributes exactly when $v \in J$, so

$$\mathbb{E}_v\left[N_J([0, \infty))\right] = \mathbf{1}\{v \in J\} + \sum_{\ell \ge 1} \sum_{j\in J} (K^\ell)_{jv} \le q_v^{(a)},$$

because $a^\ell \ge 1$ and $q_v^{(a)}$ contains the $\ell = 0$ root term explicitly. For $u > 0$, the root does not contribute, and a type-$j \in J$ offspring in generation $\ell \ge 1$ has age distributed as $S_\ell := E_1 + \cdots + E_\ell$ with $E_m \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\beta)$ and expected multiplicity $(K^\ell)_{jv}$. Hence

$$\mathbb{E}_v\left[N_J([u, \infty))\right] = \sum_{\ell \ge 1} \sum_{j\in J} (K^\ell)_{jv} \mathbb{P}(S_\ell \ge u).$$

By Chernoff's bound,

$$\mathbb{P}(S_\ell \geq u) \leq e^{-c_L u}\Big(\frac{\beta}{\beta - c_L}\Big)^\ell = e^{-c_L u} a^\ell.$$

Therefore, for $u > 0$,

$$\mathbb{P}_v(L_J \geq u) \leq e^{-c_L u} \sum_{\ell \geq 1} a^\ell \sum_{j \in J} (K^\ell)_{jv} \leq q_v^{(a)} e^{-c_L u}.$$

Together with the $u = 0$ case this proves (43). Summing against μ_v and using (41) gives (44). □

We are now ready to prove the next lemma, which combines (i) the exponential decay in u of the probability that a cluster contains a type-J event at age larger than u (Lemma A.6) and (ii) the uniform bound on the exponential moments of the total cluster size (equation (42)). These two ingredients give finite moment bounds on the local cluster counts that do not depend on the dimension d. For real u, let $N_{J,u}^{\mathrm{tail}}$ be the number of type-J events in the full cluster with ages in $[u, \infty)$, and let $N_{J,u}^{\mathrm{win}}$ be the number of type-J events in the truncated unit window $[u, u+1) \cap [0, \infty)$.

Lemma A.7. *Fix $J \subset [d]$ with $|J| \leq 2$. Then there exist constants $\eta_C > 0$, $c_{cl} > 0$, and $C_{cl} > 0$, depending only on $(\beta, \mu_+, k, w_+, \gamma)$, such that for every $u \geq 0$ and every $s \geq -1$,*

$$\sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_{J,u}^{\mathrm{tail}})^2 e^{\eta_C N_{J,u}^{\mathrm{tail}}}\big] \leq C_{cl} e^{-c_{cl} u}, \quad \sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_{J,s}^{\mathrm{win}})^2 e^{\eta_C N_{J,s}^{\mathrm{win}}}\big] \leq C_{cl} e^{-c_{cl}(s)_+}. \quad (45)$$

Proof. Let N_J^{tot} denote the total number of type-J events in the whole cluster, root included when $v \in J$. Since $N_J^{\mathrm{tot}} = \sum_{(j,t) \in \mathcal{C}, j \in J} 1 \leq \sum_{(j,t) \in \mathcal{C}} q_j^{(a)} = W_{J,v}^{(a)}$, (42) with $m = 4$ yields:

$$\sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_J^{\mathrm{tot}})^4 e^{2\eta_C N_J^{\mathrm{tot}}}\big] \leq C \quad (46)$$

for some sufficiently small $\eta_C > 0$.

Now $N_{J,u}^{\mathrm{tail}} \leq N_J^{\mathrm{tot}}$ and the event $\{N_{J,u}^{\mathrm{tail}} > 0\} \subset \{L_J \geq u\}$. Hence, by Cauchy–Schwarz,

$$\begin{aligned}
\sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_{J,u}^{\mathrm{tail}})^2 e^{\eta_C N_{J,u}^{\mathrm{tail}}}\big] &\leq \sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_J^{\mathrm{tot}})^2 e^{\eta_C N_J^{\mathrm{tot}}} \mathbf{1}\{L_J \geq u\}\big] \\
&\leq \left(\sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_J^{\mathrm{tot}})^4 e^{2\eta_C N_J^{\mathrm{tot}}}\big]\right)^{1/2} \left(\sum_{v=1}^{d} \mu_v \mathbb{P}_v(L_J \geq u)\right)^{1/2} \\
&\leq C e^{-c_{cl} u},
\end{aligned}$$

using (44) and after decreasing $c_{cl} > 0$ if necessary. For the window counts, first take $s \geq 0$. Then $N_{J,s}^{\mathrm{win}} \leq N_J^{\mathrm{tot}}$ and $\{N_{J,s}^{\mathrm{win}} > 0\} \subset \{L_J \geq s\}$, so the same argument gives

$$\sum_{v=1}^{d} \mu_v \, \mathbb{E}_v\big[(N_{J,s}^{\mathrm{win}})^2 e^{\eta_C N_{J,s}^{\mathrm{win}}}\big] \leq C e^{-c_{cl} s}.$$

If $-1 \leq s < 0$, then $N_{J,s}^{\mathrm{win}} \leq N_{J,0}^{\mathrm{win}}$ because $[s, s+1) \cap [0, \infty) = [0, s+1) \subset [0, 1)$, so the same bound follows from the already proved case $s = 0$. □

From now on, we consider independent root time-blocks: for every integer $q \in \mathbb{Z}$, let $I_q := [qh, (q+1)h)$ and $U_q :=$ all clusters rooted in I_q. The sequence $(U_q)_{q \in \mathbb{Z}}$ is independent (note that the observed bin variables $(\xi_0, \ldots, \xi_{n-1})$, where ξ_r is any in (32), are *not* independent) and consequently define the filtration $\mathcal{F}_q := \sigma(U_p : \ p \leq q)$. Thus, for any statistic from (32), the random variable

$$S_n := \sum_{r=0}^{n-1} (\xi_r - \mathbb{E}\xi_0) \quad (47)$$

is measurable with respect to $(U_q)_{q \leq n-1}$. For one root born at $x = (v, t)$, let $\mathcal{C}_x$ denote its full cluster, including the root itself. For $r = 0, \ldots, n-1$, the family-specific one-cluster envelopes are:

- for a type-j, $B_j^x(r) := \sum_{e\in\mathcal{C}_x,\mathrm{type}(e)=j,\, s_e\le rh} e^{-\beta(rh-s_e)}$,
- for a type-i, $A_i^x(r) := \mathbf{1}\{\mathcal{C}_x \text{ contributes at least one type-}i \text{ event in } (rh,(r+1)h]\}$.

Set also:

$$\Gamma_{Z_j}(x) \coloneqq \sum_{r=0}^{n-1} B_j^x(r), \qquad \Gamma_{Y_i}(x) \coloneqq \sum_{r=0}^{n-1} A_i^x(r) \tag{48}$$

$$\Gamma_{Z_jY_i}(x) \coloneqq 2\Gamma_{Z_j}(x) + R\Gamma_{Y_i}(x), \qquad \Gamma_{Z_aZ_b}(x) \coloneqq 2R\Gamma_{Z_a}(x) + 2R\Gamma_{Z_b}(x). \tag{49}$$

Lemma A.8. *Fix one sequence (ξ_r) from (32). Let $x=(v,t)$ and let $\mathcal{C}_x$ be one cluster rooted at x. For every fixed configuration of all other blocks, adding $\mathcal{C}_x$ changes the centered statistic by at most*

$$\big|S_n(\mathcal{C}_x) - S_n(0)\big| \le \Gamma_\xi(x,\mathcal{C}_x),$$

where $S_n(0)$ denotes the value obtained when that one cluster is absent.

Proof. First of all, note that the centering is deterministic and cancels when taking the difference, so we will not consider it. We can prove the claim for each of the element from (32):

- $\xi = Z_{j,r}^{(R)}$. ψ_R is 1-Lipschitz, so adding one type-j event at time $s_e \le rh$ changes $X_j(rh)$ by exactly $e^{-\beta(rh-s_e)}$, and after clipping the change is at most this. Thus
$$\left|\Delta_{\mathcal{C}_x}\sum_{r=0}^{n-1} Z_{j,r}^{(R)}\right| \le \sum_{e\in\mathcal{C}_x,\ \mathrm{type}(e)=j}\ \sum_{r=0}^{n-1} e^{-\beta(rh-s_e)}\cdot\mathbf{1}\{s_e\le rh\} = \sum_{r=0}^{n-1} B_j^x(r) = \Gamma_{Z_j}(x,\mathcal{C}_x).$$
- $\xi = Y_{i,r}^{(h)}$. Note that $A_i^x(r)=0$ iff $\mathcal{C}_x$ has no type-i offspring in $(rh,(r+1)h]$. Therefore removing $\mathcal{C}_x$ gives $|\Delta_{\mathcal{C}_x}Y_{i,r}^{(h)}| = 0 \le A_i^x(r)$. Now, note that $A_i^x(r)=1$ in two cases: the first one, if $\mathcal{C}_x$ is the only cluster having type-i offspring in $(rh,(r+1)h]$. Therefore removing $\mathcal{C}_x$ gives $|\Delta_{\mathcal{C}_x}Y_{i,r}^{(h)}| = 1 \le A_i^x(r)$. The second one, if $\mathcal{C}_x$ has type-i offspring in $(rh,(r+1)h]$, but there exists $\mathcal{C}_y$ for some $y=(v',t')$ with type-i offspring in $(rh,(r+1)h]$. Then, removing $\mathcal{C}_x$ has no impact and therefore $|\Delta_{\mathcal{C}_x}Y_{i,r}^{(h)}| = 0 \le A_i^x(r)$. Then the result follows immediately.
- $\xi = Z_{j,r}^{(R)}Y_{i,r}^{(h)}$. Use the inequality $|(z+\delta)(y+a)-zy| \le 2|\delta| + R|a|$ for $y\ge 0$, $y+a\le 1$, $z\ge 0$ and $z+\delta\le R$. Then $z+\delta$ plays the role of $Z_{j,r}^{(R)}$ when considering $\mathcal{C}_x$, while z plays that role when it is not considered. Similarly, $y+a$ plays the role of $Y_{i,r}^{(h)}$ when considering $\mathcal{C}_x$, while y plays that role when it is not considered. Therefore the result follows taking δ as $\Delta_{\mathcal{C}_x}Z_{j,r}^{(R)}$ and a as $\Delta_{\mathcal{C}_x}Y_{i,r}^{(h)}$, and using the previous results.
- $\xi = Z_{a,r}^{(R)}Z_{b,r}^{(R)}$. Same as the previous step, but using $|(z_a+\delta_a)(z_b+\delta_b)-z_az_b| \le 2R|\delta_a|+2R|\delta_b|$.

□

Now we show an auxiliary lemma useful for proving the next one about conditional mgf of one block increment:

Lemma A.9. *Let $(E,\mathcal{E})$ be a standard Borel space, let η be a Poisson point process on E with intensity measure ν, and let $F=F(\eta)$ be a measurable functional with $\mathbb{E}[F]=0$. Assume there exists a deterministic measurable function $g: E\to[0,\infty)$ such that for every configuration η and every point $z\in E$,*

$$|D_zF(\eta)| := |F(\eta+\delta_z) - F(\eta)| \le g(z).$$

Assume moreover that for some $\eta_*>0$,

$$\int_E g(z)^2 e^{\eta_* g(z)}\,\nu(dz) < \infty.$$

Then there exist a constant $\lambda_{\mathrm{P}}>0$, depending only on η_*, such that

$$\mathbb{E}\big[e^{\lambda F}\big] \le \exp\left(\lambda^2\int_E g(z)^2 e^{\eta_* g(z)}\,\nu(dz)\right) \qquad (|\lambda|\le\lambda_{\mathrm{P}}). \tag{50}$$

Proof. For $K \geq 1$, let $F^{(K)} := (-K) \vee (F \wedge K) - \mathbb{E}[(-K) \vee (F \wedge K)]$. The clipping map is 1-Lipschitz, so $|D_z F^{(K)}| \leq g(z)$. It therefore suffices to prove the claim for bounded centered functionals and then let $K \uparrow \infty$.
Assume for the moment that F is bounded. Take $B \subseteq [0,1]$ with $card(B) = card(E)$ and s.t. $B \in \mathcal{B}([0,1])$. Since $([0,1], \mathcal{B}([0,1]))$ is standard Borel, then Corollary (13.4) in [33] holds and $(B, \mathcal{B}([0,1])|_B)$ is standard Borel. By assumption $(E, \mathcal{E})$ is standard Borel, then by Theorem (15.6) in [33] there exists an isomorphism $\phi : E \to B \subseteq [0,1]$ such that ϕ and ϕ^{-1} are Borel measurable. Being η a Poisson point process with intensity ν and being ϕ Borel measurable, by Theorem (5.1) in [37], $\phi_{\#}\eta$ is again Poisson with intensity $\phi_{\#}\nu$ and distributional properties are preserved. The induced functional $\widetilde{F}(\phi_{\#}\eta) := F(\eta)$ is well defined because ϕ is a bijection and $\mathbb{E}[\tilde{F}(\phi_{\#}\eta)] = \mathbb{E}[F(\eta)] = 0$. Moreover, for any λ it holds $\mathbb{E}[e^{\tilde{F}(\phi_{\#}\eta)}] = \mathbb{E}[e^{\lambda F(\eta)}]$, thus any mgf bound for $\tilde{F}(\phi_{\#}\eta)$ is exactly equivalent for $F(\eta)$. From the definition of pushforward:

$$|D_w \tilde{F}| = |\tilde{F}(\phi_{\#}\eta + \delta_w) - \tilde{F}(\phi_{\#}\eta)| = |F(\eta + \delta_z) - F(\eta)| \leq g(z) = (g \circ \phi^{-1})(w) = \tilde{g}(w)$$

so the add-one bound is preserved by redefining g. Lastly note that:

$$\int_B \tilde{g}(w)^2 e^{\eta_* \tilde{g}(w)} (\phi_{\#}\nu)(dw) = \int_E \tilde{g}(\phi(z))^2 e^{\eta_* \tilde{g}(\phi(z))} \nu(dz) = \int_E g(z)^2 e^{\eta_* g(z)} \nu(dz) < \infty.$$

So it is possible to prove Lemma A.9 over the transported system $(B, \mathcal{B}([0,1])|_B, \phi_{\#}\eta, \tilde{F}, \tilde{g}, \phi_{\#}\nu)$, renaming it as the original one $(E, \mathcal{E}, \eta, F, g, \nu)$.
Now, realize η as the time-1 section of a marked Poisson process $\bar{\eta}$ on $E \times [0,1]$ with intensity $\nu(dz)\, ds$. Write

$$\eta_s := \bar{\eta}(\cdot \times [0,s]), \qquad \eta^s := \bar{\eta}(\cdot \times (s,1]), \qquad \mathcal{G}_s := \sigma\big(\bar{\eta}|_{E \times [0,s]}\big),$$

and define the bounded Doob martingale $M_s := \mathbb{E}[F(\eta_s + \eta^s) \mid \mathcal{G}_s] = \mathbb{E}[F(\eta) \mid \mathcal{G}_s]$, then $M_0 = \mathbb{E}[F] = 0$ and $M_1 = F$. If a marked point (z,s) appears at time s, the jump of M is

$$J_s(z) := \mathbb{E}\Big[F(\eta_{s-} + \delta_z + \eta^s) - F(\eta_{s-} + \eta^s) \,\Big|\, \mathcal{G}_{s-}, z\Big],$$

hence $|J_s(z)| \leq g(z)$ and, because F is bounded, the martingale M is square-integrable. For the completed natural filtration $(\mathcal{G}_s)$ of the Poisson random measure $N(ds,dz) := \bar{\eta}(ds,dz)$, the Theorem (6.7) II in [28] yields a predictable integrand $H \in L^2(\Omega \times (0,1] \times E, d\mathbb{P}\, ds\, \nu(dz))$ such that

$$M_t = \int_{(0,t] \times E} H_s(z)\, \widetilde{N}(ds,dz), \qquad \widetilde{N}(ds,dz) := N(ds,dz) - \nu(dz)\, ds.$$

At a jump time s of N with mark z, from Doob martingale definition

$$\Delta M_s = \mathbb{E}\Big[F(\eta_{s-} + \delta_z + \eta^s) - (\eta_{s-} + \eta^s) \,\Big|\, \mathcal{G}_{s-}, z\Big] = J_s(z).$$

At same time, the stochastic integral representation has jump size $H_s(z)$. By uniqueness of the predictable representation, $H_s(z) = J_s(z)$ for $d\mathbb{P}\, ds\, \nu(dz)$-a.e. (ω, s, z), and therefore

$$M_t = \int_{(0,t] \times E} J_s(z)\, \widetilde{N}(ds,dz).$$

Apply Itô formula to $e^{\lambda M_t}$ (Proposition 8.19 [19]), take expectations and use $|J_s(z)| \leq g(z)$:

$$\mathbb{E}[e^{\lambda M_t}] \leq 1 + \int_0^t \mathbb{E}[e^{\lambda M_{s-}}] \left(\int_E \big(e^{|\lambda| g(z)} - 1 - |\lambda| g(z)\big)\, \nu(dz) \right) ds.$$

Note that for $|\lambda| \leq \eta_*/2$,

$$0 \leq e^{|\lambda| g(z)} - 1 - |\lambda| g(z) \leq \frac{\lambda^2}{2} g(z)^2 e^{|\lambda| g(z)} \leq \frac{\lambda^2}{2} g(z)^2 e^{\eta_* g(z)},$$

so the compensator term is finite by the integrability assumption $\int_E g(z)^2 e^{\eta_* g(z)}\, \nu(dz) < \infty$. Since F is bounded, also $|M_t| \leq \|F\|_\infty$ almost surely for every t, hence $0 \leq \mathbb{E}[e^{\lambda M_t}] \leq e^{|\lambda| \|F\|_\infty} < \infty$. Thus the function $f(t) := \mathbb{E}[e^{\lambda M_t}]$ is finite on $[0,1]$, and Gronwall's inequality applies to give

$$\mathbb{E}[e^{\lambda F}] \leq \exp\left(\int_E \big(e^{|\lambda| g(z)} - 1 - |\lambda| g(z)\big)\, \nu(dz) \right) \qquad (\lambda \in \mathbb{R}).$$

Now choose $\lambda_{\mathrm{P}} := \eta_*/2$. For $y \geq 0$, again $e^y - 1 - y \leq y^2 e^y$, so for $|\lambda| \leq \lambda_{\mathrm{P}}$,

$$e^{|\lambda| g(z)} - 1 - |\lambda| g(z) \leq \lambda^2 g(z)^2 e^{|\lambda| g(z)} \leq \lambda^2 g(z)^2 e^{\eta_* g(z)}.$$

Substituting this into the previous display gives (50) for bounded centered F. Finally apply the bounded case to $F^{(K)}$. Because $F^{(K)} \to F$ almost surely and in L^1, Fatou's lemma for $|\lambda| \leq \lambda_{\mathrm{P}}$ yields

$$\mathbb{E}[e^{\lambda F}] \leq \liminf_{K \to \infty} \mathbb{E}[e^{\lambda F^{(K)}}] \leq \exp\left(\lambda^2 \int_E g(z)^2 e^{\eta_* g(z)} \, \nu(dz) \right).$$

□

Lemma A.10. *Fix one family (ξ_r) from (32). For a root birth at $x = (v, t)$, define*

$$b_\xi(v, t) := \mathbb{E}\Big[\Gamma_\xi(x)^2 e^{\eta_1 \Gamma_\xi(x)} \Big],$$

where $\eta_1 = \eta_1(h, R) > 0$ is chosen so that all these expectations are finite. Let

$$V_q(\xi) := \int_{I_q} \sum_{v=1}^{d} \mu_v \, b_\xi(v, t) \, dt, \qquad I_q = [qh, (q+1)h), \ q \in \mathbb{Z}.$$

Then there exist a constant $c_\lambda > 0$ depending only on $(\beta, \mu_+, k, w_+, \gamma, h_0)$, such that for every finite $q \leq n - 1$, the Doob martingale difference $D_q := \mathbb{E}[S_n \mid \mathcal{F}_q] - \mathbb{E}[S_n \mid \mathcal{F}_{q-1}]$ satisfies

$$\mathbb{E}[e^{\lambda D_q} \mid \mathcal{F}_{q-1}] \leq \exp\big(\lambda^2 V_q(\xi)\big) \qquad (|\lambda| \leq c_\lambda h/R). \tag{51}$$

Proof. Fix one finite block $q \leq n - 1$. Conditional on $\mathcal{F}_{q-1}$, the random block U_q is an independent marked Poisson point process, its state space is the space of marked triples $(v, t, \mathcal{C})$, where $v \in [d]$, $t \in I_q$, and $\mathcal{C}$ is a full cluster rooted at (v, t). By Marking Theorem (see [37]), the intensity measure is $\nu_q(dv\, dt\, d\mathcal{C}) = \mu_v \, dt \, \mathbb{P}_{v,t}(d\mathcal{C})$. Define the centered block functional

$$F_q(U_q) := \mathbb{E}[S_n \mid \mathcal{F}_{q-1}, U_q] - \mathbb{E}[S_n \mid \mathcal{F}_{q-1}].$$

Since $\mathcal{F}_q = \sigma(\mathcal{F}_{q-1}, U_q)$, we have $F_q(U_q) = D_q$. Now fix one marked cluster $z = (x, \mathcal{C})$ with $x = (v, t) \in [d] \times I_q$. Let η^+ denote the future root blocks $(U_{q+1}, \ldots, U_{n-1})$. For every fixed realization of $\mathcal{F}_{q-1}$, every block configuration $U = (\ldots, U_{q-1}, U_q, \eta_+)$, and every realization of η^+, Lemma A.8 gives

$$\big| S_n(U + \delta_z) - S_n(U) \big| \leq \Gamma_\xi(x, \mathcal{C}).$$

Taking conditional expectation over η^+ and using Jensen therefore yields

$$\big| F_q(U + \delta_z) - F_q(U) \big| = \Big| \mathbb{E}\big[S_n(U + \delta_z) - S_n(U) \mid \mathcal{F}_{q-1}, U_q \big] \Big| \leq \Gamma_\xi(x, \mathcal{C}).$$

Hence Lemma A.9, conditional to $\mathcal{F}_{q-1}$, applied with $g(z) = \Gamma_\xi(z)$ and $\eta_* = \eta_1(h, R)$, gives:

$$\mathbb{E}[e^{\lambda D_q} \mid \mathcal{F}_{q-1}] \leq \exp\left(\lambda^2 \int g(z)^2 e^{\eta_1 g(z)} \, \nu_q(dz) \right) \qquad (|\lambda| \leq \lambda_{\mathrm{P}}).$$

By the definition of ν_q and of $b_\xi(v, t)$,

$$\int g(z)^2 e^{\eta_1 g(z)} \, \nu_q(dz) = \int_{I_q} \sum_{v=1}^{d} \mu_v \, \mathbb{E}\Big[\Gamma_\xi(v, t)^2 e^{\eta_1 \Gamma_\xi(v,t)} \Big] dt = V_q(\xi).$$

Since Lemma A.9 gives $\lambda_{\mathrm{P}} = \eta_1(h, R)/2 = c_\eta h/(2R)$, (51) holds with $c_\lambda := c_\eta/2$, i.e. the admissible range is $|\lambda| \leq c_\lambda h/R$. □

We now bound the aggregated second-order exponential moments of the one-cluster envelopes Γ_ξ:

Lemma A.11. *Assume $0 < h \leq h_0$, and $R \geq 1$. Then there exist constants $c_\eta > 0$, $c_\Gamma > 0$, and $C_\Gamma > 0$, depending only on $(\beta, \mu_+, k, w_+, \gamma, h_0)$, such that with $\eta_1(h, R) := c_\eta h/R$, one has, uniformly over all $d \geq 1$, all $B \in \mathcal{B}_{d,k}$, all indices, and all $t \in \mathbb{R}$,*

$$\sum_{v=1}^{d} \mu_v b_{Z_j}(v, t) \leq C_\Gamma h^{-2} e^{-c_\Gamma (-t)_+}, \qquad \sum_{v=1}^{d} \mu_v b_{Y_i}(v, t) \leq C_\Gamma e^{-c_\Gamma (-t)_+}, \tag{52}$$

$$\sum_{v=1}^{d} \mu_v b_{Z_j Y_i}(v, t) \leq C_\Gamma \big(h^{-2} + R^2 \big) e^{-c_\Gamma (-t)_+}, \quad \sum_{v=1}^{d} \mu_v b_{Z_a Z_b}(v, t) \leq C_\Gamma R^2 h^{-2} e^{-c_\Gamma (-t)_+}. \tag{53}$$

Proof. For Y_i, if the root is born at time t, then every contributing type-i event must occur at age at least $u := (-t)_+$. Because $N^{\mathrm{tail}}_{\{i\},u}$ counts events in the full cluster, including the root when $v = i$ and $u = 0$, this bound also covers the case $t \geq 0$ in which the only contribution may come from the root itself. Thus $\Gamma_{Y_i}(x) \leq N^{\mathrm{tail}}_{\{i\},u}$. Applying (45) gives the first in (52).

For Z_j, first suppose $t \geq 0$. Then, recall the first definition in (48), then every type-j event contributes at most $C_h \leq C/h$ to the sampled sum (see definition (35)), so considering every type-j event $\Gamma_{Z_j}(x) \leq C_h N^{\mathrm{tail}}_{\{j\},0}$. Here again $N^{\mathrm{tail}}_{\{j\},0}$ counts the root itself when $v = j$, so the bound remains valid even if the cluster has no offspring at all. Then (45) yields the bound $\sum_v \mu_v b_{Z_j}(v,t) \leq Ch^{-2}$. Now suppose $t < 0$ and write $u := -t > 0$. Split the type-j events of the cluster according to their ages: events with age at least u satisfy $t + s \geq 0$, so we only use $a_h(t+s) \leq C_h$. For the remaining events, ages in the strip

$$[u - m - 1, u - m) \cap [0, \infty), \qquad m \geq 0,$$

satisfy $t + s \in [-m - 1, -m)$ and therefore $a_h(t+s) \leq C_h e^{-\beta m}$. This includes the first partial window $[0, u)$ when $0 < u < 1$, namely the case $m = 0$. Hence

$$\Gamma_{Z_j}(x) \leq C_h N^{\mathrm{tail}}_{\{j\},u} + C_h \sum_{m\geq 0} e^{-\beta m} N^{\mathrm{win}}_{\{j\},u-m-1} =: \sum_{r\geq 0} w_r A_r. \tag{54}$$

where in the last definition we set

$$A_0 := N^{\mathrm{tail}}_{\{j\},u}, \qquad A_{m+1} := N^{\mathrm{win}}_{\{j\},u-m-1}, \quad w_0 := C_h, \qquad w_{m+1} := C_h e^{-\beta m} \quad (m \geq 0).$$

Then let

$$W_h := \sum_{r\geq 0} w_r = C_h\Big(1 + \sum_{m\geq 0} e^{-\beta m}\Big) \leq C_\beta h^{-1}.$$

For $\phi_{\eta,W}(y) := y^2 e^{\eta W y}$, convexity of $\phi_{\eta,W}$ on $[0,\infty)$, and Jensen's inequality for the probability weights $p_r := w_r/W_h$ give

$$\Big(\sum_{r\geq 0} w_r A_r\Big)^2 e^{\eta \sum_{r\geq 0} w_r A_r} = W_h^2\, \phi_{\eta,W_h}\Big(\sum_{r\geq 0} p_r A_r\Big) \leq W_h^2 \sum_{r\geq 0} p_r \phi_{\eta,W_h}(A_r) = W_h \sum_{r\geq 0} w_r A_r^2 e^{\eta W_h A_r}.$$

Applying this with $\eta = \eta_1$ and using $W_h \leq C_\beta h^{-1}$ yields

$$\Gamma_{Z_j}(x)^2 e^{\eta_1 \Gamma_{Z_j}(x)} \leq Ch^{-2}\left((N^{\mathrm{tail}}_{\{j\},u})^2 e^{c_h \eta_1 N^{\mathrm{tail}}_{\{j\},u}} + \sum_{m\geq 0} e^{-\beta m} (N^{\mathrm{win}}_{\{j\},u-m-1})^2 e^{c_h \eta_1 N^{\mathrm{win}}_{\{j\},u-m-1}}\right),$$

for some $c_h \asymp W_h \lesssim h^{-1}$. Let $\eta_{\mathrm{C}} > 0$ be the exponential-moment parameter from Lemma A.7. In order to apply Lemma A.7, we need $c_h \eta_1$ sufficiently small. With $\eta_1(h,R) := c_\eta h/R$ and $W_h \leq C_\beta h^{-1}$, one has $c_h \eta_1(h,R) \leq C_\beta \eta_1(h,R)/h = C_\beta c_\eta/R \leq C_\beta c_\eta$ uniformly in $R \geq 1$ and $h \in (0, h_0]$. It therefore suffices to choose $c_\eta \leq \min\Big\{\frac{\eta_{\mathrm{C}}}{4C_\beta}, \frac{\eta_{\mathrm{C}}}{2h_0}\Big\}$, which depends only on $(\beta, \mu_+, k, w_+, \gamma, h_0)$. With this choice, summing over v, using Lemma A.7 and decreasing $c_\Gamma > 0$ if necessary lead to

$$\sum_{v=1}^{d} \mu_v b_{Z_j}(v,t) \leq Ch^{-2}\left(e^{-c_\Gamma u} + \sum_{m\geq 0} e^{-\beta m} e^{-c_\Gamma (u-m-1)_+}\right) \leq Ch^{-2} e^{-c_\Gamma u}$$

For the mixed $Z_j Y_i$, write $a := 2\Gamma_{Z_j}$ and $b := R\Gamma_{Y_i}$ and apply $(a+b)^2 e^{\eta_1(a+b)} \leq C(a^2 e^{2\eta_1 a} + b^2 e^{2\eta_1 b})$. The first term $4\Gamma^2_{Z_j} e^{4\eta_1 \Gamma_{Z_j}}$ is handled exactly as in the Z_j case, since the effective parameter $4c_h\eta_1 = 4C_\beta c_\eta/R \leq \eta_{\mathrm{C}}$. The second term $R^2 \Gamma^2_{Y_i} e^{2R\eta_1 \Gamma_{Y_i}}$ has exponent $2R\eta_1 = 2R \cdot c_\eta h/R = 2c_\eta h \leq \eta_{\mathrm{C}}$; since $\Gamma_{Y_i} \leq N^{\mathrm{tail}}_{\{i\},u}$, Lemma A.7 applies and R^2 gives the bound $C_\Gamma(h^{-2} + R^2) e^{-c_\Gamma(-t)_+}$ in (53). For $Z_a Z_b$, set $a := 2R\Gamma_{Z_a}$ and $b := 2R\Gamma_{Z_b}$; the effective parameter for each Γ_{Z_a} term is $4Rc_h\eta_1 = 4R \cdot (C_\beta/h) \cdot (c_\eta h/R) = 4C_\beta c_\eta \leq \eta_{\mathrm{C}}$, so Lemma A.7 applies and yields the bound $C_\Gamma R^2 h^{-2} e^{-c_\Gamma(-t)_+}$ in (53). □

We now sum the profile bounds of Lemma A.11 over all root blocks to obtain the total variance proxy $\sum_q V_q(\xi)$ needed by the next bounded-block concentration lemma.

Proposition A.7. *Assume $0 < h \le h_0$, $R \ge 1$, and $T \ge 1$. Then there exists a constant $C > 0$, depending only on $(h, R, \beta, \mu_+, k, w_+, \gamma, h_0)$, such that uniformly over all $d \ge 1$, all $B \in \mathcal{B}_{d,k}$, and all indices,*

$$\sum_{q=-\infty}^{n-1} V_q(Z_j) \le C\,\frac{n}{h}, \qquad \sum_{q=-\infty}^{n-1} V_q(Y_i) \le C\,nh, \tag{55}$$

$$\sum_{q=-\infty}^{n-1} V_q(Z_jY_i) \le C\,n\Big(\frac{1}{h} + R^2h\Big), \qquad \sum_{q=-\infty}^{n-1} V_q(Z_aZ_b) \le C\,n\,\frac{R^2}{h} \tag{56}$$

Proof. Recall the definition for $V_q(\xi)$ given in Lemma A.10. For $q = 0, \dots, n-1$, by Lemma A.11:

$$V_q(Z_j) \le Ch^{-1}, \qquad V_q(Y_i) \le Ch, \qquad V_q(Z_jY_i) \le C(h^{-1}+R^2h), \qquad V_q(Z_aZ_b) \le CR^2h^{-1}.$$

For $q < 0$, write $t \in [qh, (q+1)h)$ so that $(-t)_+ \asymp |q|h$ and therefore

$$V_q(Z_j) \le Ch^{-1}e^{-c|q|h}, \qquad V_q(Y_i) \le Che^{-c|q|h},$$
$$V_q(Z_jY_i) \le C(h^{-1}+R^2h)e^{-c|q|h}, \qquad V_q(Z_aZ_b) \le CR^2h^{-1}e^{-c|q|h}.$$

Summing the geometric tails and using $T \ge 1$, hence $n = \lfloor T/h \rfloor \gtrsim h^{-1}$, gives

$$\sum_{q=-\infty}^{n-1} V_q(Z_j) \le C\left(\frac{n}{h} + \sum_{m\ge1} h^{-1}e^{-cmh}\right) \le C\left(\frac{n}{h} + \frac{1}{h^2}\right) \le C\frac{n}{h},$$

$$\sum_{q=-\infty}^{n-1} V_q(Y_i) \le C\left(nh + \sum_{m\ge1} he^{-cmh}\right) \le C(nh+1) \le Cnh,$$

and the two product bounds follow in the same way. □

Now, we are ready to state the central result for this section: the lemmas above show that each individual cluster has a limited impact on the empirical averages (Lemma A.8 and Lemma A.10), and that this impact decays as the cluster ages (Lemma A.6 and Lemma A.11). Combining these bounds across all clusters and all time blocks, the following proposition shows that every empirical average used by the estimator concentrates around its true value, with a tail probability that decays exponentially in n. This is the key statistical guarantee that will allow the estimator to succeed with high probability once n is large enough.

Proposition A.8. *Assume $0 < h \le h_0$, $R \ge 1$, and $T \ge 1$. Then there exist constants $c, u_0 > 0$, depending on $(h, R, \beta, \mu_+, k, w_+, \gamma, h_0)$, such that uniformly over all $d \ge 1$, all $B \in \mathcal{B}_{d,k}$, and all choices of (ξ_r) from the family* (32)*, the following holds for every u with $0 < u \le u_0$:*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{r=0}^{n-1}(Z_{j,r}^{(R)} - \mathbb{E}Z_{j,0}^{(R)})\right| > u\right) \le 2\exp\left(-cnhu^2\right), \tag{57}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{r=0}^{n-1}(Y_{i,r}^{(h)} - \mathbb{E}Y_{i,0}^{(h)})\right| > u\right) \le 2\exp\left(-cn\frac{u^2}{h}\right), \tag{58}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{r=0}^{n-1}(Z_{j,r}^{(R)}Y_{i,r}^{(h)} - \mathbb{E}[Z_{j,0}^{(R)}Y_{i,0}^{(h)}])\right| > u\right) \le 2\exp\left(-cn\frac{u^2}{h^{-1}+R^2h}\right), \tag{59}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{r=0}^{n-1}(Z_{a,r}^{(R)}Z_{b,r}^{(R)} - \mathbb{E}[Z_{a,0}^{(R)}Z_{b,0}^{(R)}])\right| > u\right) \le 2\exp\left(-cn\frac{u^2}{R^2h^{-1}}\right). \tag{60}$$

Proof. Let S_n as in (47) and D_q as in Lemma A.10 with $q \le n-1$. For $M \ge 1$, set

$$S_n^{(M)} := \sum_{q=-M}^{n-1} D_q = \mathbb{E}[S_n \mid \mathcal{F}_{n-1}] - \mathbb{E}[S_n \mid \mathcal{F}_{-M-1}].$$

By Lemma A.10, $\mathbb{E}[e^{\lambda D_q} \mid \mathcal{F}_{q-1}] \le \exp\big(\lambda^2 V_q(\xi)\big)$ with $|\lambda| \le c_\lambda h/R$. Iterating for $q = -M, \dots, n-1$ gives

$$\mathbb{E}[e^{\lambda S_n^{(M)}}] \le \exp\Big(\lambda^2 \sum_{q=-M}^{n-1} V_q(\xi)\Big).$$

By Proposition A.7, $\sum_{q=-\infty}^{n-1} V_q(\xi) < \infty$. Moreover, differentiating (51) at $\lambda = 0$ shows $\mathbb{E}[D_q^2] \le V_q(\xi)$, so $\sum_{q=-\infty}^{n-1} \mathbb{E}[D_q^2] < \infty$. In particular $(S_n^{(M)})_{M\ge1}$ is Cauchy in L^2. To identify its limit, note that S_n is $\mathcal{F}_{n-1}$-measurable and $\mathbb{E}[S_n] = 0$, hence

$$S_n^{(M)} = \mathbb{E}[S_n \mid \mathcal{F}_{n-1}] - \mathbb{E}[S_n \mid \mathcal{F}_{-M-1}] = S_n - \mathbb{E}[S_n \mid \mathcal{F}_{-M-1}].$$

Note that $(\mathcal{F}_{-M-1})_{M\ge1}$ decrease to the tail sigma-field of the independent root blocks $(U_q)_{q\in\mathbb{Z}}$

$$\mathcal{F}_{-\infty} := \bigcap_{M\ge1} \mathcal{F}_{-M-1},$$

By the reverse martingale convergence theorem, $\mathbb{E}[S_n \mid \mathcal{F}_{-M-1}] \to \mathbb{E}[S_n \mid \mathcal{F}_{-\infty}]$ in L^2 and by Kolmogorov's 0-1 law the tail sigma-field $\mathcal{F}_{-\infty}$ is trivial, so $\mathbb{E}[S_n \mid \mathcal{F}_{-\infty}] = \mathbb{E}[S_n] = 0$. Taking $|2\lambda| \le c_\lambda h/R$, the same mgf bound applied with the sign-correct parameter 2λ gives

$$\sup_M \mathbb{E}[e^{2\lambda S_n^{(M)}}] \le \exp\Big(4\lambda^2 \sum_{q=-\infty}^{n-1} V_q(\xi)\Big) < \infty.$$

where, for simplicity on exposition, we choose to introduce a suitable positive constant C instead of a numerical factor. Thus, the family $\{e^{\lambda S_n^{(M)}}\}_M$ is bounded in L^2 and therefore uniformly integrable. Therefore

$$\mathbb{E}[e^{\lambda S_n}] = \lim_{M\to\infty} \mathbb{E}[e^{\lambda S_n^{(M)}}] \le \exp\Big(\lambda^2 \sum_{q=-\infty}^{n-1} V_q(\xi)\Big).$$

In order to have $\mathbb{E}[e^{\lambda S_n}] \le e^{K_\xi \lambda^2}$ with $|\lambda| \le c_\lambda h/(2R)$, write

$$K_\xi := \sum_{q=-\infty}^{n-1} V_q(\xi).$$

For any $\lambda \in [0, c_\lambda h/(2R)]$, Chernoff's bound gives $\mathbb{P}(S_n/n > u) \le \exp\big(-\lambda n u + K_\xi \lambda^2\big)$. Choose $\lambda = nu/2K_\xi$ (we can do this whenever $u \le u_0$, as we will see in a moment), then

$$\mathbb{P}\left(\frac{S_n}{n} > u\right) \le \exp\left(-\frac{n^2u^2}{4K_\xi}\right),$$

and the same bound holds for the lower tail. Proposition A.7 gives

$$K_{Z_j} \lesssim \frac{n}{h}, \qquad K_{Y_i} \lesssim nh, \qquad K_{Z_jY_i} \lesssim n(h^{-1} + R^2h), \qquad K_{Z_aZ_b} \lesssim \frac{nR^2}{h},$$

with constants depending only on $(h, R, \beta, \mu_+, k, w_+, \gamma, h_0)$. Thus the four displayed tail bounds (57)–(60) follow once

$$u \le u_0 := \frac{c_\lambda h}{2R} \min\left\{\frac{1}{h},\ h,\ h^{-1} + R^2h,\ \frac{R^2}{h}\right\},$$

after absorbing fixed factors into the constant c. □

Next lemma translates Proposition A.8 into simultaneous high-probability control over all empirical screening statistics $\widehat{F}_{ij}^{(h,R)}$, Gram matrices $\widehat{\Sigma}_C^{(R)}$, and score vectors $\widehat{g}_{i,C}^{(h,R)}$ across all d^2 index pairs and all candidate sets of size at most m, via a union bound costing a $\log d$ factor.

Lemma A.12. *Fix $0 < h \le 1$, $R \ge 1$, and $m \in \mathbb{N}$. There exist constants $C_{\mathrm{prob}}, c_{\mathrm{prob}}, C_0 > 0$, depending only on $(m, h, R, \beta, \mu_+, k, w_+, \gamma, h_0)$, such that for every d, every $B \in \mathcal{B}_{d,k}$, and every $T \ge 1$ satisfying*

$$n \ge C_0 \max\left\{\frac{h^{-1} + R^2h}{u_0(h,R)^2},\ \frac{R^2}{h\,u_0(h,R)^2}\right\} \log d, \tag{61}$$

with probability at least $1 - d^{-c_{\mathrm{prob}}}$ the following hold simultaneously:

(i) for all rows i and all coordinates j,

$$|\widehat{F}_{ij}^{(h,R)} - F_{ij}^{(h,R)}| \le C_{\mathrm{prob}}\sqrt{\frac{(h^{-1}+R^2h)\log d}{n}}; \tag{62}$$

(ii) for all candidate sets C with $|C| \le m$,

$$\|\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)}\|_{\mathrm{op}} \le C_{\mathrm{prob}} R\sqrt{\frac{\log d}{nh}}; \tag{63}$$

(iii) for all rows i and all candidate sets C with $|C| \le m$,

$$\|\widehat{g}_{i,C}^{(h,R)} - g_{i,C}^{(h,R)}\|_\infty \le C_{\mathrm{prob}}\sqrt{\frac{(h^{-1}+R^2h)\log d}{n}}. \tag{64}$$

Proof. (i) Under (61), the screening/score deviation level $\sqrt{(h^{-1}+R^2h)(\log d)/n}$ and the Gram deviation level $R\sqrt{(\log d)/(nh)}$ are both at most $u_0(h,R)$ after enlarging C_0 if necessary. Hence Proposition A.8 is applicable in each instance. For screening, write

$$\hat{m}_{ZY} := \frac{1}{n}\sum_{r=0}^{n-1} Z_{j,r}^{(R)} Y_{i,r}^{(h)}, \qquad \hat{m}_Z := \frac{1}{n}\sum_{r=0}^{n-1} Z_{j,r}^{(R)}, \qquad \hat{m}_Y := \frac{1}{n}\sum_{r=0}^{n-1} Y_{i,r}^{(h)},$$

and similarly m_{ZY}, m_Z, m_Y for the expectations. Then by the definition at the beginning of Appendix A.2, $\widehat{F}_{ij}^{(h,R)} - F_{ij}^{(h,R)} = (\hat{m}_{ZY} - m_{ZY}) - (\hat{m}_Z\hat{m}_Y - m_Z m_Y)$. By Proposition A.8,

$$|\hat{m}_{ZY}-m_{ZY}| \le C\sqrt{\frac{(h^{-1}+R^2h)\log d}{n}}, \qquad |\hat{m}_Z-m_Z| \le C\sqrt{\frac{\log d}{nh}}, \qquad |\hat{m}_Y-m_Y| \le C\sqrt{\frac{h\log d}{n}}.$$

Also $m_Y = \mathbb{E}[Y_{i,0}^{(h)}] \le Ch$ and $|\hat{m}_Z| \le R$. Hence summing and subtracting $\hat{m}_Z m_Y$ in $|\hat{m}_Z\hat{m}_Y - m_Z m_Y|$, using the bounds just obtained and the fact that $h \le 1$ prove (62).

(ii) Now for the Gram matrix, write

$$\hat{M}_{ab} := \frac{1}{n}\sum_{r=0}^{n-1} Z_{a,r}^{(R)} Z_{b,r}^{(R)}, \qquad M_{ab} := \mathbb{E}[Z_{a,0}^{(R)} Z_{b,0}^{(R)}],$$

and let $\hat{m}_a, m_a$ be the empirical and theoretical means of $Z_a^{(R)}$. Then as before, $(\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)})_{ab} = (\hat{M}_{ab} - M_{ab}) - (\hat{m}_a\hat{m}_b - m_a m_b)$. Now Proposition A.8 yields

$$|\hat{M}_{ab} - M_{ab}| \le CR\sqrt{\frac{\log d}{nh}}, \qquad |\hat{m}_a - m_a| \le C\sqrt{\frac{\log d}{nh}}.$$

Therefore

$$|(\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)})_{ab}| \le CR\sqrt{\frac{\log d}{nh}} + 2R\cdot C\sqrt{\frac{\log d}{nh}} \le CR\sqrt{\frac{\log d}{nh}}.$$

Since $|C| \le m$ is fixed, (63) follows.

(iii) For the score vector, write

$$\hat{s}_a := \frac{1}{n}\sum_{r=0}^{n-1} Z_{a,r}^{(R)} Y_{i,r}^{(h)}, \qquad s_a := \mathbb{E}[Z_{a,0}^{(R)} Y_{i,0}^{(h)}].$$

Then

$$(\widehat{g}_{i,C}^{(h,R)} - g_{i,C}^{(h,R)})_a = (\hat{s}_a - s_a) - (\hat{m}_a\hat{m}_Y - m_a m_Y),$$

so the same bounds as in the screening case give (64). Finally, there are at most d^2 screening pairs (i,j), at most $\sum_{r=0}^m \binom{d}{r} \le C_m d^m$ candidate sets C, and at most $d\sum_{r=0}^m \binom{d}{r} \le C_m d^{m+1}$ row-model pairs (i,C). A union bound completes the proof. □

A.5 Sure screening and exact recovery

Proposition A.9. *Under the assumptions of Proposition A.5 and Lemma A.12, there exists a constant $C_{\mathrm{scr}} > 0$ such that if*

$$n \geq C_{\mathrm{scr}} \frac{h^{-1} + R^2 h}{\alpha^2 h^2} \log d, \tag{65}$$

then with probability at least $1 - d^{-c}$ one has $S_i \subset \widehat{C}_i$ for every $i = 1, \ldots, d$.

Proof. On the event (62), adding and subtracting $\min_{j \in S_i} F_{ij}$ and $\max_{j \notin S_i} F_{ij}$ and using conveniently the elementary inequality $|a_j - b_j| \geq a_j - b_j$, we get

$$\min_{j \in S_i} \widehat{F}_{ij}^{(h,R)} - \max_{j \notin S_i} \widehat{F}_{ij}^{(h,R)} \geq \min_{j \in S_i} F_{ij}^{(h,R)} - \max_{j \notin S_i} F_{ij}^{(h,R)} - 2C_{\mathrm{prob}} \sqrt{\frac{(h^{-1} + R^2 h) \log d}{n}}.$$

By Proposition A.5, the theoretical screening gap is at least $c_{\mathrm{scr}} \alpha h$. Thus if

$$2C_{\mathrm{prob}} \sqrt{\frac{(h^{-1} + R^2 h) \log d}{n}} \leq \frac{1}{2} c_{\mathrm{scr}} \alpha h,$$

then every support index outranks every non-support index. Since $m \geq k \geq |S_i|$, the top-m empirical screening set contains S_i. This is implied by (65) after enlarging the constant. □

Proposition A.10. *Under the assumptions in Lemma A.4, Proposition A.6, and Lemma A.12, there exists a constant $C_{ols} > 0$ such that if*

$$n \geq C_{ols} \max\left\{ \frac{R^2}{h}, \frac{h^{-1} + R^2 h}{\alpha^2 h^2} \right\} \log d, \tag{66}$$

then with probability at least $1 - d^{-c}$ the following holds simultaneously for all rows i and all candidate sets $C \supset S_i$ with $|C| \leq m$:

$$\|\widehat{y}_{i,C}^{(h,R)} - y_{i,C}^{(h,R)}\|_\infty \leq C_{ols} \sqrt{\frac{(h^{-1} + R^2 h) \log d}{n}}. \tag{67}$$

In particular, on the sure-screening event, the same bound holds for $C = \widehat{C}_i$ for every i.

Proof. Fix (i, C). On the event (63), Lemma A.4 leads to $\|\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)}\|_{\mathrm{op}} \leq \kappa/2$. Then using Weyl's inequality $\widehat{\Sigma}_C^{(R)}$ is invertible and $\|(\widehat{\Sigma}_C^{(R)})^{-1}\|_{\mathrm{op}} \leq 2/\kappa$. Moreover,

$$y_{i,C}^{(h,R)} - y_{i,C}^{(h,R)} = (\widehat{\Sigma}_C^{(R)})^{-1} \Big[\widehat{g}_{i,C}^{(h,R)} - g_{i,C}^{(h,R)} + \Sigma_C^{(R)} y_{i,C}^{(h,R)} - \widehat{\Sigma}_C^{(R)} y_{i,C}^{(h,R)} \Big],$$

then rearranging:

$$\widehat{y}_{i,C}^{(h,R)} - y_{i,C}^{(h,R)} = (\widehat{\Sigma}_C^{(R)})^{-1} \Big[(\widehat{g}_{i,C}^{(h,R)} - g_{i,C}^{(h,R)}) - (\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)}) y_{i,C}^{(h,R)} \Big].$$

Proposition A.6 implies $\|y_{i,C}^{(h,R)}\|_\infty \leq C\alpha h$ uniformly, hence $\|y_{i,C}^{(h,R)}\|_2 \leq C_m \alpha h$. Therefore (63) and (64) give

$$\|\widehat{y}_{i,C}^{(h,R)} - y_{i,C}^{(h,R)}\|_\infty \leq \frac{2}{\kappa} \Big(\|\widehat{g}_{i,C}^{(h,R)} - g_{i,C}^{(h,R)}\|_\infty + C_m \alpha h \|\widehat{\Sigma}_C^{(R)} - \Sigma_C^{(R)}\|_{\mathrm{op}} \Big) \leq C_{ols} \sqrt{\frac{(h^{-1} + R^2 h) \log d}{n}}.$$

A union bound over all row/model pairs is already built into Lemma A.12. The requirement (66) guarantees both the invertibility condition and the final bound above. □

We can now proof the upper bound in Theorem 3.1 from the main text.

Proof of Theorem 3.1. Choose $\alpha_0 > 0$ small enough that Propositions A.5, A.6, and Lemma A.4 all apply for every $0 < \alpha \leq \alpha_0$. From the proofs of Proposition A.4, Lemma A.4, and Proposition A.6, choose

$$A_R := \max\left\{ \frac{16\beta C_{\mathrm{clip}}}{\mu_- w_-}, \frac{16\beta C_m}{\mu_-}, \frac{8C_y}{w_-}, 1 \right\},$$

and set $R(\alpha) := A_R\alpha^{-1}$. Next choose

$$A_h := \frac{1}{2}\min\left\{1, \frac{\mu_- w_-}{32\beta C_{\mathrm{bin}} A_R}, \frac{w_-}{8C_y A_R}\right\},$$

and set $h(\alpha) := A_h\alpha^2$. With this pair, the screening gap (26), the theoretical least-squares slope separation (29), and the Gram nondegeneracy all hold simultaneously.

Let $n = \lfloor T/h(\alpha)\rfloor$. Once α is fixed, the substituted values $h(\alpha)$ and $R(\alpha)$ are also fixed, so the constants from Lemma A.12, Proposition A.9, and Proposition A.10 become finite numbers depending on α. Besides the screening and OLS requirements from Propositions A.9 and A.10, we must also satisfy the admissible-range condition (61). For $\alpha \le \alpha_0$ we have $h(\alpha) \le h_0 = 1$ and $R(\alpha) \ge 1$, so $u_0(h(\alpha), R(\alpha)) \ge ch(\alpha)^2/R(\alpha)$. Therefore, for this fixed α, all remaining requirements are implied by

$$n \ge A_\alpha \log d,$$

where

$$A_\alpha := C_0 \max\left\{\frac{h(\alpha)^{-1} + R(\alpha)^2 h(\alpha)}{\alpha^2 h(\alpha)^2}, \frac{R(\alpha)^2}{h(\alpha)}, \frac{\big(h(\alpha)^{-1} + R(\alpha)^2 h(\alpha)\big)R(\alpha)^2}{h(\alpha)^4}, \frac{R(\alpha)^4}{h(\alpha)^5}\right\}.$$

Now choose $C_* > 0$, depending only on $(\beta, \mu_\pm, w_\pm, k, m, \alpha)$, large enough that, for every integer $d \ge 2$,

$$T \ge C_* \log d$$

implies both $T \ge 1$ and $n = \lfloor T/h(\alpha)\rfloor \ge A_\alpha \log d$. Hence Lemma A.12, Proposition A.9, and Proposition A.10 apply simultaneously. Now, with probability at least $1-d^{-c_*}$ one has simultaneously

$$S_i \subset \widehat{C}_i \qquad \text{for all } i,$$

and

$$\|\widehat{y}^{(h,R)}_{i,\widehat{C}_i} - y^{(h,R)}_{i,\widehat{C}_i}\|_\infty \le \frac{1}{8}\alpha w_- h.$$

Then for every row i, because $\widehat{C}_i \supset S_i$ and (29) applies with $C = \widehat{C}_i$,

$$\min_{j\in S_i} \widehat{y}_{ij} \ge \min_{j\in S_i} y^{(h,R)}_{ij,\widehat{C}_i} - \frac{1}{8}\alpha w_- h \ge \frac{5}{8}\alpha w_- h > \frac{1}{2}\alpha w_- h = \tau_{\alpha,h},$$

while for $j \in \widehat{C}_i \setminus S_i$,

$$|\widehat{y}_{ij}| \le |y^{(h,R)}_{ij,\widehat{C}_i}| + \frac{1}{8}\alpha w_- h \le \frac{3}{8}\alpha w_- h < \frac{1}{2}\alpha w_- h = \tau_{\alpha,h}.$$

Hence thresholding at $\tau_{\alpha,h}$ recovers S_i exactly for every row. The overall success probability is at least $1 - d^{-c_*}$ for some $c_* > 0$. □

Remark A.1. *The main goal of this work is to characterize the scaling of the observation time, which we prove to scale logarithmically in the dimension of the network. Naturally, the dependence of the multiplicative constant on other parameters may be loose. In particular, a formal power counting of the coupling parameter α suggests a scaling of α^{-12}, which is at great variance to the scaling of α^{-2} of the lower bound, leaving an optimality gap of α^{-10}. As the proof requires α to be small, this may seem to limit the practical applicability of the result. First, we would like to emphasize that the proposed procedure is first and foremost a proof tool and is not intended to be a practical algorithm to replace established methods. Second, we will show in the following analysis that the α^{-12} scaling is more an artifact of the proof rather than the intrinsic scaling of the problem.*

A.6 On the intrinsic α-scaling of the problem

Here, we show that replacing Lemma A.2, which requires one new Lemma and an updated Corollary, already improves the α-dependence to α^{-4} polylog$(1/\alpha)$, thereby improving the optimality gap from α^{-10} to α^{-2} polylog$(1/\alpha)$. This does not yet reveal the intrinsic α-dependence of the problem, but may help future work to characterize it in full detail, which is why we report it here.

The starting point is again the cluster representation, in which we previously defined the intensity measure $\nu(dv, dt, d\mathcal{C}) = \mu_v\, dt\, \mathbb{P}_v(d\mathcal{C})$ for the Poisson point process $(v, t, \mathcal{C})$. Recall also the definition given before for $N^{win}_{J,s}(\mathcal{C})$, i.e. the number of type-J events in the truncated unit window $[s, s+1) \cap [0, \infty)$, and consider the result at Lemma A.7. The main changes needed in order to obtain the stated power counting are the following:

Lemma A.13. *There exists $\eta_X > 0$ and $C_X < \infty$ depending only on $(\beta, \mu_+, k, w_+, \gamma_0)$ such that*

$$\sup_{d,B,j} \mathbb{E}[e^{\eta_X X_j(0)}] \le C_X$$

Proof. Following the usual cluster representation and identifying a as the age of the offspring w.r.t. the root v, write

$$X_j(0) = \sum_{(v,t,\mathcal{C})} f_j(t,\mathcal{C}) \quad , \qquad f_j(t,\mathcal{C}) = \sum_{(j,a)\in\mathcal{C},\, t+a\le 0} e^{\beta(t+a)}$$

Apply Campbell's theorem for the Laplace functional with the function $\eta f_j(t,\mathcal{C}) > 0$ (with $\eta > 0$) in order to obtain

$$\log \mathbb{E}[e^{\eta X_j(0)}] = \sum_v \mu_v \int_0^\infty \mathbb{E}_v[e^{\eta f_j(-u,\mathcal{C})} - 1]du$$

where $t = -u$ is substituted in the integral. Using $e^x - 1 \le xe^x$ for $x \ge 0$ leads to

$$\log \mathbb{E}[e^{\eta X_j(0)}] \le \eta \sum_v \mu_v \int_0^\infty \mathbb{E}_v[f_j e^{\eta f_j}]du. \tag{68}$$

In order to bound $\sum_v \mu_v \mathbb{E}_v[fe^{\eta f}]$, consider again the cluster representation: fixing a root time $t = -u \le 0$, partition the contributing ages in $a \in [u-m-1, u-m)$ (with $m \ge 0$) and $a = u$. For the former, the contributing weights in the representation are $e^{-\beta(u-a)} < e^{-\beta m}$ and their number is at most $N^{win}_{\{j\},\,u-m-1}$; in fact $-u + a \in [-m-1, -m)$ and therefore $u - a > m$. For the latter, the weight is 1 and the number is $N^{win}_{\{j\},\,u}$. Thus, since $[0,u] = \{u\} \cup \bigcup_{m\ge 0}[u-m-1, u-m)$, the pathwise bound is:

$$f \le 1 \cdot N^{win}_{\{j\},\,u} + \sum_{m\ge 0} e^{-\beta m} \cdot N^{win}_{\{j\},\,u-m-1} = \sum_{r\ge 0} w_r A_r$$

and, taking $w_0 = 1$ and $w_{m+1} = e^{-\beta m}$, the total weight is:

$$W = \sum_{r\ge 0} w_r = 1 + \sum_{m\ge 0} e^{-\beta m} = 1 + \frac{1}{1-e^{-\beta}} =: C_\beta < \infty.$$

Defining $p_r = w_r/W$, apply Jensen in the second inequality (noting that a function as $\phi(y) = y\exp\{\eta W y\}$ is increasing and convex on $[0,\infty)$):

$$fe^{\eta f} \le W\left(\sum_{r\ge 0} p_r A_r\right) e^{\eta W \sum_{r\ge 0} p_r A_r} \le W \sum_{r\ge 0} p_r A_r e^{\eta W A_r} = \sum_{r\ge 0} w_r A_r e^{\eta W A_r}$$

Then apply $\mathbb{E}_v[\cdot]$ and sum against μ_v to have:

$$\sum_v \mu_v \mathbb{E}_v[fe^{\eta f}] \le \sum_{r\ge 0} w_r \sum_v \mu_v \mathbb{E}_v[A_r e^{\eta W A_r}]. \tag{69}$$

Choose $\eta := \frac{\eta_C}{C_\beta}$ so that $\eta W = \eta_C$. It trivially holds that $A_r \le A_r^2$ (in fact A_r takes values in $\{0,1,2,...\}$), then also $A_r e^{\eta_c A_r} \le A_r^2 e^{\eta_c A_r}$ pointwise. Using this inequality and Lemma A.7 with the convention that $N^{win}_{J,s} = 0$ if $s < -1$, we get:

$$\sum_v \mu_v \mathbb{E}_v[A_r e^{\eta W A_r}] \le \sum_v \mu_v \mathbb{E}_v[A_r^2 e^{\eta_C A_r}] \le C_{cl} e^{-c_{cl}(s_r)_+}$$

with $s_0 = u$ and $s_{m+1} = u - m - 1$. Substituting this into the inequality (69) leads to:

$$\sum_v \mu_v \mathbb{E}_v[fe^{\eta f}] \le C_{cl}[e^{-c_{cl}u} + \sum_{m\ge 0} e^{-\beta m} e^{-c_{cl}(u-m-1)_+}].$$

Finally, integrate in $u \in [0,\infty)$ term by term; then substituting the result in (68) gives

$$\log \mathbb{E}[e^{\eta X_j(0)}] \le \eta\, C_{\mathrm{cl}}\Big[\frac{1}{c_{\mathrm{cl}}} + \sum_{m\ge 0} e^{-\beta m}\Big(m + 1 + \frac{1}{c_{\mathrm{cl}}}\Big)\Big] =: \eta C' < \infty.$$

□

Notice that for every integer $p \geq 1$ and $x \geq 0$, $x^p \leq p! e^x$; hence $\mathbb{E}[X_j(0)^p] = \eta_X^{-p} \mathbb{E}[(\eta_X X_j(0))^p] \leq p! \eta_X^{-p} \mathbb{E}[e^{\eta_X X_j(0)}] \leq p! \eta_X^{-p} C_X =: M_p'$ uniformly over (d, B, j). This translates to the fact that Proposition A.1 can be substituted with known constants only depending on (β, μ_+, k, w_+), without any threshold on α (i.e. without any α_P).
Now, using this Lemma, we can state a new clipping corollary.

Corollary A.1. *There exist $c_0 > 0$, $C_0 < \infty$ such that, for all $R \geq 1$, uniformly over (d, B, i, j), with $\delta_j^{(R)} := (X_j(0) - R)_+$,*

$$\mathbb{E}\big[\delta_j^{(R)}\big] \;\leq\; C_0 e^{-c_0 R}, \qquad \mathbb{E}\big[(\delta_j^{(R)})^2\big]^{1/2} \;\leq\; C_0 e^{-c_0 R}, \qquad \mathbb{E}\big[\delta_j^{(R)}\, \lambda_i(0)\big] \;\leq\; C_0 e^{-c_0 R}.$$

Proof. For $x > 0$, using $(x - R)_+ \leq x\, \mathbf{1}\{x > R\} \leq x\, e^{\eta_X (x-R)/2}$ and $y e^y \leq e^{2y}$, we get the chain of inequalities:

$$\mathbb{E}[\delta^{(R)}] \leq e^{-\eta_X R/2}\, \mathbb{E}\big[X e^{\eta_X X/2}\big] \leq e^{-\eta_X R/2} \cdot \frac{2}{\eta_X}\, \mathbb{E}\big[e^{\eta_X X}\big] \leq \frac{2 C_X}{\eta_X}\, e^{-\eta_X R/2}$$

where in the last inequality the previous Lemma is used. Similarly for the second: use $(x - R)_+^2 \leq x^2 e^{\eta_X (x-R)/2}$ and $y^2 e^y \leq C e^{2y}$; this gives $\mathbb{E}[(\delta^(R))^2] \leq C C_X e^{-\eta_X R/2}$. The third follows by Cauchy–Schwarz from the second and $\sup \mathbb{E}[\lambda_i(0)^2] < \infty$ (see how this is done in the proof for Lemma A.2 as a reference). Take $c_0 := \eta_X/4$ and C_0 the largest of the constants. □

The following Lemma would now substitute Lemma A.2:

Lemma A.14 (Lemma A.2 changed)**.** *For all $R \geq 1$, uniformly over (d, B, i, j, ℓ):*

$$\mathbb{E}[\delta_j^{(R)}] \leq C_0 e^{-c_0 R}, \qquad \mathbb{E}[(\delta_j^{(R)})^2] \leq C_0^2 e^{-2 c_0 R},$$

$$\big|\operatorname{Cov}\big(Z_j^{(R)}, \delta_\ell^{(R)}\big)\big| \;\vee\; \big|\operatorname{Cov}\big(\delta_j^{(R)}, \lambda_i(0)\big)\big| \leq C_1 e^{-c_0 R}.$$

Proof. The first two are already proven. For the third: $|\operatorname{Cov}(Z_j, \delta_\ell)| \leq \mathbb{E}[Z_j \delta_\ell] + \mathbb{E}[Z_j]\mathbb{E}[\delta_\ell] \leq \mathbb{E}[X_j \delta_\ell] + M_1' \mathbb{E}[\delta_\ell]$ and $\mathbb{E}[X_j \delta_\ell] \leq (M_2')^{1/2} \mathbb{E}[\delta_\ell^2]^{1/2} \leq C e^{-c_0 R}$ by Cauchy-Schwarz and the previous result. Similarly, using some results from the proof for Lemma A.1 and similarly to some considerations in the proof for Lemma A.2, we get $|\operatorname{Cov}(\delta_j, \lambda_i)| \leq \mathbb{E}[\delta_j \lambda_i] + \mathbb{E}[\delta_j]\mathbb{E}[\lambda_i] \leq C e^{-c_0 R} + C_0 e^{-c_0 R} \beta \bar{m}$ with $\bar{m} := \sup_j m_j \leq \mu_+/(\beta(1-\gamma))$. Since $e^{-c_0 R} \leq C_{c_0}/R$, any results whose proof uses Lemma A.2 remain true. □

Choose now $R = A_R \log(1/\alpha)$ (see Proposition A.4 and Proposition A.6 revisited via Lemma A.14) and consequently $h = A_h\, \alpha/\log(1/\alpha)$ (see Proposition A.5 and Proposition A.6). The same formal power counting as before, now results in α^{-4} polylog(α), which reduces the optimality gap w.r.t. the lower bound to α^{-2} polylog(α). This does not require any restriction on the model class considered.

B An information-theoretic view of support recovery

We briefly introduce the information-theoretic perspective of support recovery that is used in the lower bound. We do not derive any new statements, but instead aim to explain why entropy, mutual information, KL divergence, and Fano's inequality naturally appear in exact support recovery.

Information theory studies how much can be learned about an unknown object from noisy observations. Let V be a discrete random variable taking values in a finite set $\mathcal{V}$, and suppose that, conditional on $V = v$, the observation Y has law P_v. In communication language, V is the message, the stochastic mechanism $v \mapsto P_v$ is a noisy channel, and an estimator $\widehat{V} = \widehat{V}(Y)$ is a decoder. In a statistical problem, the same viewpoint applies when V indexes a finite collection of hypotheses or models.

Support recovery fits naturally into this framework because the target is discrete. In the hard subclass used below, the unknown object is the parent set

$$S \in \mathcal{S}_{d,k}(i_\star), \qquad |\mathcal{S}_{d,k}(i_\star)| = \binom{d-1}{k}.$$

Thus S can be viewed as a message chosen from $\mathcal{S}_{d,k}(i_\star)$. The Hawkes process acts as the noisy channel: conditional on S, it produces the observation

$$Y_T = (X(0), N_{[0,T]})$$

with law P_S^T. A support estimator $\widehat{S} = \widehat{S}(Y_T)$ is a decoder trying to infer which message was sent.

The entropy of a discrete random variable V is

$$H(V) := -\sum_{v \in \mathcal{V}} \mathbb{P}(V = v) \log \mathbb{P}(V = v),$$

where logarithms are natural, so information is measured in nats. Entropy measures the uncertainty in V before observing any data. If V is uniform on $\mathcal{V}$, then $H(V) = \log |\mathcal{V}|$. Hence choosing one object among M equally likely alternatives requires $\log M$ nats of information. In our support-recovery problem, a uniformly chosen parent set satisfies

$$H(S) = \log |\mathcal{S}_{d,k}(i_\star)| = \log \binom{d-1}{k}.$$

For fixed k, this quantity is of order $\log d$. This is the basic combinatorial uncertainty that the data must resolve.

After observing Y, the remaining uncertainty about V is measured by the conditional entropy $H(V \mid Y)$. The reduction in uncertainty caused by the observation is the mutual information

$$I(V;Y) := H(V) - H(V \mid Y).$$

If Y is independent of V, then $I(V;Y) = 0$. If Y determines V exactly, then $I(V;Y) = H(V)$. Thus mutual information measures how much of the initial uncertainty about the message is actually transmitted through the noisy channel.

Mutual information can also be expressed using KL divergence. For probability measures P and Q,

$$\mathrm{KL}(P\|Q) := \mathbb{E}_P \left[\log \frac{dP}{dQ} \right],$$

whenever P is absolutely continuous with respect to Q. KL divergence is not a metric, but it measures statistical distinguishability: small KL divergence means that observations from P are hard to distinguish from observations from Q. If V has prior distribution π, and

$$\bar{P} := \sum_{v \in \mathcal{V}} \pi_v P_v$$

is the mixture law of Y, then

$$I(V;Y) = \sum_{v \in \mathcal{V}} \pi_v \, \mathrm{KL}(P_v \| \bar{P}).$$

Thus the information carried by Y about V is the average distinguishability between the law generated by a fixed message and the overall mixture law. To get a lower-bound, it is sufficeint to compare against a convenient reference law Q:

$$I(V;Y) \le \sum_{v \in \mathcal{V}} \pi_v \, \mathrm{KL}(P_v \| Q).$$

Fano's inequality is the step that turns an information bound into an error lower bound. The idea is the following. Suppose V is uniform on a finite set $\mathcal{V}$ with $M = |\mathcal{V}|$, and let $\widehat{V} = \widehat{V}(Y)$ be any estimator. If the estimator is usually correct, then after seeing Y there cannot be much uncertainty left about V: the value $\widehat{V}(Y)$ almost identifies the true message. The only substantial uncertainty comes from the event $\{\widehat{V} \neq V\}$. On the event of no error, the message is known from the decoder's output. On the error event, the true message may still be one of many alternatives. Therefore a small error probability forces the conditional entropy $H(V \mid Y)$ to be small.

Fano's inequality formalizes the converse implication. If the observation leaves large conditional entropy, then no estimator can have small error probability. Equivalently, since

$$H(V \mid Y) = H(V) - I(V;Y),$$

if the mutual information $I(V;Y)$ is much smaller than the initial uncertainty $H(V) = \log M$, then the observation has not conveyed enough information to identify the message. The finite-hypothesis form used here is

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V;Y) + \log 2}{\log M}.$$

The additive $\log 2$ term accounts for the binary uncertainty of whether an error occurred, and is negligible when M is large. The main message of the inequality is that reliable identification among M nearly hidden possibilities requires nearly $\log M$ nats of mutual information.

This average-error statement immediately gives a minimax lower bound. For any prior π,

$$\sup_{v \in \mathcal{V}} \mathbb{P}_v(\widehat{V} \neq v) \geq \sum_{v \in \mathcal{V}} \pi_v \mathbb{P}_v(\widehat{V} \neq v).$$

Hence, if Fano's inequality shows that the prior-average error is bounded away from zero, then every estimator must fail with non-negligible probability for at least one hypothesis in the class.

In our Hawkes setting, the message set is

$$\mathcal{V} = \mathcal{S}_{d,k}(i_\star), \qquad M = \binom{d-1}{k}$$

and for fixed k, $\log M \asymp \log d$. We show that the observation up to time T carries at most

$$C_{\text{init}} + C_{\text{path}} T$$

nats of information about this support, with constants independent of d and T. Fano's inequality then compares the information available from the data to the information needed to identify the support. If $T = o(\log d)$, the data contains too little information to distinguish among the possible parent sets, and the probability of an error is bounded away from zero.

C Lower bound proof details

Here, we provide the full proof of Theorem 4.1. As before, the unknown parent set is $S \in \mathcal{S}_{d,k}(i_\star)$, the active edge weight in the subclass is θ_-, and the target and source background rates are $\bar{\mu}_\star$ and $\bar{\mu}$. Let P_S^T be the law of the stationary observation $Y_T = (X(0), N_{[0,T]})$ under support S, and let P_0^T denote the law of the reference model with no interactions and the same background rates. We write π_S and π_0 for the corresponding stationary laws of $X(0)$, and $K_S^T(x,\cdot)$ and $K_0^T(x,\cdot)$ for the conditional laws of the path $N_{[0,T]}$ given $X(0) = x$.

The proof is structured in four parts. In Section C.1, we collect two standard entropy facts and properties of one-dimensional Poisson shot-noise density function needed for the stationary initial condition. In Section C.2, we reduce the initial KL term to a fixed $(k+1)$-dimensional problem. In Section C.3, we bound the dynamic path KL by a second moment of the source block. Finally, in Section C.4 we apply Fano's inequality to prove Theorem 4.1.

C.1 Preparatory lemmas

First, we prove the chain-rule for the Kullback–Leibler divergence.

Lemma C.1. *For every $S \in \mathcal{S}_{d,k}(i_\star)$,*

$$\mathrm{KL}(P_S^T \| P_0^T) = \mathrm{KL}(\pi_S \| \pi_0) + \int \mathrm{KL}\big(K_S^T(x,\cdot) \| K_0^T(x,\cdot)\big)\, \pi_S(dx),$$

with equality in $[0,\infty]$.

Proof. This is the standard chain rule for relative entropy of probability kernels. When $P_S^T \ll P_0^T$, write

$$P_S^T(dx, d\eta) = \pi_S(dx) K_S^T(x, d\eta), \qquad P_0^T(dx, d\eta) = \pi_0(dx) K_0^T(x, d\eta),$$

and expand the Radon–Nikodym derivative. The extended-real case follows by the usual approximation argument for relative entropy of kernels. □

The next lemma bounds the mutual information by the average KL to a baseline.

Lemma C.2. *Let V be uniform on a finite set $\mathcal{V}$, and let Y be an observation whose law under $V = v$ is P_v. Then for any probability measure Q,*

$$I(V;Y) \le \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathrm{KL}(P_v \| Q).$$

Proof. Let $\bar{P} := |\mathcal{V}|^{-1} \sum_{v \in \mathcal{V}} P_v$ be the mixture law. Then

$$\frac{1}{|\mathcal{V}|} \sum_{v} \mathrm{KL}(P_v \| Q) = \frac{1}{|\mathcal{V}|} \sum_{v} \mathrm{KL}(P_v \| \bar{P}) + \mathrm{KL}(\bar{P} \| Q) = I(V;Y) + \mathrm{KL}(\bar{P} \| Q),$$

follows by writing $\log(dP_v/dQ) = \log(dP_v/d\bar{P}) + \log(d\bar{P}/dQ)$, integrating under P_v, and averaging over v. The claim follows from $\mathrm{KL}(\bar{P} \| Q) \ge 0$. □

The next result collects several properties of the density of the stationary law of one-dimensional Poisson shot noise. It is later used to show that translated stationary densities have finite KL cost.

For $r > 0$, let π_r^{P} be the stationary law of $dU(t) = -\beta U(t)\, dt + dM(t)$, where M is Poisson with rate r. Equivalently,

$$U_r \overset{d}{=} \int_{(-\infty,0]} e^{\beta s}\, dM_s \overset{d}{=} \int_0^\infty e^{-\beta t}\, dM_t$$

and its characteristic function is

$$\mathbb{E}\big[e^{itU_r}\big] = \exp\!\left(\frac{r}{\beta} \int_0^1 \frac{e^{itu} - 1}{u}\, du \right), \qquad t \in \mathbb{R}. \tag{70}$$

Thus U_r is infinitely divisible with Lévy density $\nu_r(du) = (r/\beta)\mathbf{1}_{(0,1)}(u) u^{-1}\, du$.

Lemma C.3. *Fix $r > 0$ and set $\alpha = r/\beta$. The law π_r^{P} has a density q_r on $(0,\infty)$ such that:*

(i) *q_r is strictly positive on $(0,\infty)$ and unimodal.*

(ii) *There exists a constant $c_r > 0$ such that*

$$q_r(x) = c_r x^{\alpha - 1} \quad \text{for Lebesgue-a.e. } x \in (0,1).$$

(iii) *There exists a finite constant C_r such that for Lebesgue-a.e. $x > 0$,*

$$-\log q_r(x) \le C_r\big(1 + x \log(1+x) + |\log x|\big). \tag{71}$$

(iv) *U_r has exponential moments of every order,*

$$\mathbb{E}[e^{\eta U_r}] = \exp\!\left(\frac{r}{\beta} \int_0^1 \frac{e^{\eta u} - 1}{u}\, du \right) < \infty, \qquad \eta > 0, \tag{72}$$

and hence $\mathbb{E}[U_r \log(1+U_r)] < \infty$, $\mathbb{E}[|\log U_r|] < \infty$, and $\int q_r |\log q_r| < \infty$.

Proof. The Lévy measure has density $k(u)u^{-1}\, du$ with $k(u) = \alpha \mathbf{1}_{(0,1)}(u)$ nonincreasing. Sato [55, Theorem 15.10 and Corollary 15.11] therefore implies that U_r is self-decomposable. Absolute continuity, strict positivity on $(0,\infty)$, and unimodality then follow from Sato [55, Theorem 27.13, Lemma 53.1, and Theorem 53.1]. This proves (i).

For (ii), stationarity gives, for every $f \in C_c^1(0,\infty)$,

$$\int_0^\infty \{-\beta x f'(x) + r(f(x+1) - f(x))\} q_r(x)\, dx = 0.$$

Taking $f \in C_c^1(0,1)$ removes the jump-in term and yields $\beta(xq_r)' - rq_r = 0$ in $\mathcal{D}'((0,1))$. Hence $(x^{1-\alpha} q_r)' = 0$ distributionally, so $q_r(x) = c_r x^{\alpha-1}$ for a.e. $x \in (0,1)$; positivity gives $c_r > 0$. After redefining q_r on a null set, we may take this identity to hold pointwise on $(0,1)$.

For (iii), use a unimodal version of q_r, and choose a point $m_r \geq 0$ such that q_r is nonincreasing on $[m_r, \infty)$. Fix $\delta \in (0, \min\{1/(2\beta), 1\})$. For $n \geq 1$, decompose $U_r = U_r^{(1,n)} + U_r^{(2,n)} + U_r^{(3)}$ over the time intervals $[0, \delta/n]$, $(\delta/n, 1]$, and $(1, \infty)$ in the representation $U_r = \int_0^\infty e^{-\beta t}\, dM_t$. Since $U_r^{(3)} \stackrel{d}{=} e^{-\beta} U_r'$ with U_r' an independent copy of U_r, strict positivity of the density gives $p_\delta := \mathbb{P}(U_r^{(3)} \in [\beta\delta, 1]) > 0$. On the event

$$A_n := \{M([0, \delta/n]) = n\}, \quad B_n := \{M((\delta/n, 1]) = 0\}, \quad C := \{U_r^{(3)} \in [\beta\delta, 1]\},$$

one has $U_r \in [n, n+1]$: indeed $U_r^{(1,n)} \geq n e^{-\beta\delta/n} \geq n - \beta\delta$, while $U_r^{(2,n)} = 0$ and $U_r^{(3)} \in [\beta\delta, 1]$. Independence and Stirling's upper bound give constants $c_1, c_2 > 0$, depending only on (r, β), such that

$$\mathbb{P}(U_r \in [n, n+1]) \geq e^{-r\delta/n} \frac{(r\delta/n)^n}{n!}\, e^{-r(1-\delta/n)} p_\delta \geq c_1 e^{-c_2 n \log n}.$$

For integers $n \geq \lceil m_r \rceil$, monotonicity gives $q_r(n) \geq \int_n^{n+1} q_r(x)\, dx$, hence $q_r(n) \geq c_1 e^{-c_2 n \log n}$. If $x \geq m_r + 1$ and $n = \lceil x \rceil$, then $q_r(x) \geq q_r(n)$, and therefore

$$-\log q_r(x) \leq C\{1 + x \log(1+x)\}, \qquad x \geq m_r + 1.$$

Near zero, part (ii) gives $-\log q_r(x) \leq C(1 + |\log x|)$ for a.e. $x \in (0, 1)$. As before, use a unimodal version so that q_r is nondecreasing on $(0, m_r]$ and nonincreasing on $[m_r, \infty)$. On the remaining compact interval $[1, m_r + 1]$, this version is strictly positive, and by unimodality its infimum is attained at one of the endpoints. Hence the infimum is positive. Combining these three bounds proves (71).

Finally, (iv) follows from the Poisson shot-noise formula

$$\log \mathbb{E}[e^{\eta U_r}] = r \int_0^\infty (e^{\eta e^{-\beta t}} - 1)\, dt = \frac{r}{\beta} \int_0^1 \frac{e^{\eta u} - 1}{u}\, du < \infty, \qquad \eta > 0,$$

because $e^{\eta u} - 1 \leq u(e^\eta - 1)$ on $[0, 1]$. The exponential moments imply $\mathbb{E}[U_r \log(1 + U_r)] < \infty$ and $\mathbb{E}[\log_+ U_r] < \infty$. For the negative logarithmic moment, (ii) gives

$$\int_0^1 |\log x| q_r(x)\, dx = c_r \int_0^1 |\log x| x^{\alpha - 1}\, dx < \infty,$$

so $\mathbb{E}[|\log U_r|] < \infty$. Moreover, (ii) gives integrability of $q_r |\log q_r|$ near zero. On compact subintervals of $(0, \infty)$, the chosen unimodal version is locally bounded by monotonicity on the two sides of its mode, after redefining the value at the mode if necessary; hence $q_r \log_+ q_r$ is integrable there. On the tail $[m_r + 1, \infty)$, unimodality gives $q_r(x) \leq q_r(m_r + 1)$, so the positive part is bounded by a constant times q_r. The negative part is controlled by (71) and the moment bounds above. Therefore $\int q_r |\log q_r| < \infty$. □

We now use the properties from Lemma C.3 to control the KL of translating the target's baseline stationary law by a nonnegative random amount.

Lemma C.4. *Let $q_0 := q_{\bar\mu_\star}$ be the density of $\pi^{\mathrm{P}}_{\bar\mu_\star}$. For $w \geq 0$, define the translated density*

$$q_{0,w}(z) := q_0(z - w)\, \mathbf{1}_{\{z > w\}}, \qquad z > 0,$$

that is, the law of $U + w$ when $U \sim \pi^{\mathrm{P}}_{\bar\mu_\star}$. Then there exists a finite constant C_{tr}, depending only on $(\beta, \bar\mu_\star)$, such that

$$g(w) := \mathrm{KL}(q_{0,w}\, dz \| q_0\, dz) \leq C_{\mathrm{tr}}\big(1 + w \log(1 + w)\big), \qquad w \geq 0.$$

Proof. Let $U \sim \pi^{\mathrm{P}}_{\bar\mu_\star}$. By a change of variables,

$$g(w) = \mathbb{E}[\log q_0(U) - \log q_0(U + w)].$$

Lemma C.3(iv) gives $J_0 := \int q_0 |\log q_0| < \infty$. Since $U + w$ is absolutely continuous, (71) applies to it a.s., and

$$g(w) \leq J_0 + C\mathbb{E}\{1 + (U + w)\log(1 + U + w) + |\log(U + w)|\}.$$

The elementary bound $1 + U + w \le (1 + U)(1 + w)$ gives

$$(U + w)\log(1 + U + w) \le U\log(1 + U) + w\log(1 + U) + (U + w)\log(1 + w),$$

whose expectation is bounded by $C(1 + w\log(1 + w))$. Also,

$$|\log(U + w)| \le |\log U| + \log(1 + U) + \log(1 + w),$$

because the negative part decreases when w is added and the positive part is bounded by $\log(1+U+w)$. Lemma C.3(iv) makes all constants finite, and the claim follows. □

C.2 KL of the stationary initial law

We now control the KL contribution from the state $X(0)$. The point of the lower-bound subclass is that all irrelevant coordinates factor out, leaving a fixed-dimensional comparison involving only the target coordinate and its k active parents. Let $\bar{\pi}_k$ denote the stationary law of the $(k+1)$-dimensional block $(Z, \mathbf{Y})$ in which $Y^{(1)}, \dots, Y^{(k)}$ are independent Poisson shot-noise coordinates with rate $\bar{\mu}$, and Z has intensity $\bar{\mu}_\star + \theta_- \sum_{m=1}^k Y_{t-}^{(m)}$.

Proposition C.1. *For every $d \ge k + 2$ and every $S \in \mathcal{S}_{d,k}(i_\star)$,*

$$\mathrm{KL}(\pi_S \| \pi_0) = \mathrm{KL}\big(\bar{\pi}_k \| \pi^{\mathrm{P}}_{\bar{\mu}_\star} \otimes (\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes k}\big) =: C_{\mathrm{init}}.$$

The constant C_{init} is finite and depends only on $(\beta, \bar{\mu}, \bar{\mu}_\star, \theta_-, k)$; in particular it is independent of both S and d.

Proof. Fix $S = \{s_1, \dots, s_k\} \in \mathcal{S}_{d,k}(i_\star)$ and order the remaining coordinates in any fixed way. Let R_S be the coordinate permutation that puts $i_\star$ first, then the elements of S, and then all other coordinates. KL is invariant under this bijection.

Under $\Theta^{(S)}$, all coordinates outside $\{i_\star\} \cup S$ are independent Poisson shot-noise coordinates with stationary law $\pi^{\mathrm{P}}_{\bar{\mu}}$, and they are independent of the target/source block. The nontrivial block $(Z_t, \mathbf{Y}_t) := (X_{i_\star}(t), (X_{s_1}(t), \dots, X_{s_k}(t)))$ satisfies

$$\begin{aligned} dY_t^{(m)} &= -\beta Y_t^{(m)}\, dt + dN_t^{(m)}, \qquad m = 1, \dots, k, \\ dZ_t &= -\beta Z_t\, dt + dM_t, \end{aligned}$$

where $N^{(1)}, \dots, N^{(k)}$ are independent Poisson processes of rate $\bar{\mu}$ and M has predictable intensity $\bar{\mu}_\star + \theta_- \sum_{m=1}^k Y_{t-}^{(m)}$. If $\bar{\pi}_k$ denotes the stationary law of this block, then

$$R_S \# \pi_S = \bar{\pi}_k \otimes (\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes(d-k-1)}.$$

Under the baseline model,

$$R_S \# \pi_0 = \pi^{\mathrm{P}}_{\bar{\mu}_\star} \otimes (\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes k} \otimes (\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes(d-k-1)}.$$

Product additivity of relative entropy gives

$$\mathrm{KL}(\pi_S \| \pi_0) = \mathrm{KL}\big(\bar{\pi}_k \| \pi^{\mathrm{P}}_{\bar{\mu}_\star} \otimes (\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes k}\big),$$

and the right-hand side is independent of the particular set S by exchangeability of the source coordinates.

It remains only to prove finiteness. Let $\mathcal{H} = \sigma(Y_s^{(m)} : s \le 0, 1 \le m \le k)$ be the stationary source history. Conditional on $\mathcal{H}$, the target process on $(-\infty, 0]$ is an inhomogeneous Poisson process with intensity $\bar{\mu}_\star + \theta_- \sum_m Y_{t-}^{(m)}$. Decompose it as $M = M^{(0)} + M^{(1)}$, where, conditional on $\mathcal{H}$, $M^{(0)}$ is homogeneous Poisson of rate $\bar{\mu}_\star$, $M^{(1)}$ is Poisson with intensity $\theta_- \sum_m Y_{t-}^{(m)}\, dt$, and the two are independent. Then

$$Z_0 = U + W, \qquad U := \int_{(-\infty,0]} e^{\beta s}\, dM_s^{(0)}, \qquad W := \int_{(-\infty,0]} e^{\beta s}\, dM_s^{(1)}.$$

Here $U \sim \pi^{\mathrm{P}}_{\bar{\mu}_\star}$ and U is independent of $(\mathcal{H}, W)$.

The $\mathbf{Y}_0$-marginal of $\bar{\pi}_k$ is $(\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes k}$. Hence the entropy chain rule and conditional Jensen's inequality give

$$\mathrm{KL}\big(\bar{\pi}_k \| \pi^{\mathrm{P}}_{\bar{\mu}_\star} \otimes (\pi^{\mathrm{P}}_{\bar{\mu}})^{\otimes k}\big) = \mathbb{E}\big[\mathrm{KL}(P_{Z_0|\mathbf{Y}_0} \| \pi^{\mathrm{P}}_{\bar{\mu}_\star})\big] \le \mathbb{E}\big[\mathrm{KL}(P_{Z_0|\mathcal{H}} \| \pi^{\mathrm{P}}_{\bar{\mu}_\star})\big].$$

Indeed, since $\mathbf{Y}_0$ is $\mathcal{H}$-measurable, $P_{Z_0|\mathbf{Y}_0}$ is obtained by averaging $P_{Z_0|\mathcal{H}}$ over the conditional law of $\mathcal{H}$ given $\mathbf{Y}_0$, and relative entropy is convex in its first argument. Given $\mathcal{H}$, the conditional law of Z_0 is the mixture of translates $q_{0,w}(z)\,dz$ with mixing law $P_{W|\mathcal{H}}$. Convexity of KL and Lemma C.4 therefore yield

$$\mathrm{KL}(P_{Z_0|\mathcal{H}} \| \pi^{\mathrm{P}}_{\bar{\mu}_\star}) \le \mathbb{E}[g(W) \mid \mathcal{H}] \le C\,\mathbb{E}[1 + W\log(1+W) \mid \mathcal{H}].$$

Thus finiteness follows once $\mathbb{E}[W\log(1+W)] < \infty$.

In fact W has exponential moments of all orders. For $\eta > 0$, the conditional Laplace functional gives

$$\mathbb{E}[e^{\eta W} \mid \mathcal{H}] = \exp\left(\theta_- \int_{-\infty}^{0} (e^{\eta e^{\beta s}} - 1) \sum_{m=1}^{k} Y_{s-}^{(m)}\, ds\right) \le \exp(\theta_-(e^{\eta} - 1)V)\,,$$

where

$$V := \int_{-\infty}^{0} e^{\beta s} \sum_{m=1}^{k} Y_{s-}^{(m)}\, ds.$$

Using the stationary shot-noise representation of the sources and Tonelli's theorem,

$$V = \sum_{m=1}^{k} \int_{(-\infty,0]} (-u) e^{\beta u}\, dN_u^{(m)} = \int_{(-\infty,0]} (-u) e^{\beta u}\, d\widetilde{N}_u,$$

where $\widetilde{N}$ is homogeneous Poisson with rate $k\bar{\mu}$. The kernel $u \mapsto (-u)e^{\beta u}\mathbf{1}_{\{u \le 0\}}$ is bounded and integrable, so Campbell's formula implies $\mathbb{E}[e^{cV}] < \infty$ for every $c > 0$. Consequently $\mathbb{E}[e^{\eta W}] < \infty$ for every $\eta > 0$, and in particular $\mathbb{E}[W\log(1+W)] < \infty$. This proves $C_{\mathrm{init}} < \infty$ and completes the proof. □

C.3 Dynamic KL bound

It remains to control the entropy accumulated over the observation window. We start by computing the stationary second moment of the aggregated source block.

Lemma C.5. *For $S \in \mathcal{S}_{d,k}(i_\star)$, set $\Sigma_S(t) := \sum_{j\in S} X_j(t)$. Under $\Theta^{(S)}$, Σ_S is a stationary Poisson shot-noise OU process driven by a homogeneous Poisson process of rate $k\bar{\mu}$, and*

$$\mathbb{E}[\Sigma_S(t)^2] = \frac{k^2\bar{\mu}^2}{\beta^2} + \frac{k\bar{\mu}}{2\beta} =: C_{\mathrm{path}}, \qquad t \in \mathbb{R}.$$

Proof. The k parent coordinates are independent Poisson shot-noise OU processes with rate $\bar{\mu}$, so their sum is driven by the superposed Poisson process of rate $k\bar{\mu}$. Campbell's formula gives $\mathbb{E}[\Sigma_S] = k\bar{\mu}/\beta$ and $\mathrm{Var}(\Sigma_S) = k\bar{\mu}/(2\beta)$, which proves the statement for $\mathbb{E}[\Sigma_S^2]$. □

In the following proposition, we derive the bound on the dynamic KL term.

Proposition C.2. *For every $S \in \mathcal{S}_{d,k}(i_\star)$,*

$$\int \mathrm{KL}\big(K_S^T(x,\cdot) \| K_0^T(x,\cdot)\big)\, \pi_S(dx) \le \frac{\theta_-^2}{\bar{\mu}_\star}\, C_{\mathrm{path}}\, T.$$

Consequently,

$$\mathrm{KL}(P_S^T \| P_0^T) \le C_{\mathrm{init}} + \frac{\theta_-^2}{\bar{\mu}_\star}\, C_{\mathrm{path}}\, T.$$

Proof. Fix $x \in \mathbb{R}^d$. Under model S, the intensity is

$$\lambda_{i_\star}^{(S)}(t) = \bar{\mu}_\star + \theta_- \sum_{j\in S} X_j(t-), \qquad \lambda_i^{(S)}(t) \equiv \bar{\mu} \quad (i \ne i_\star),$$

whereas under the baseline model

$$\lambda_{i_\star}^{(0)}(t) \equiv \bar{\mu}_\star, \qquad \lambda_i^{(0)}(t) \equiv \bar{\mu} \quad (i \neq i_\star).$$

For π_S-a.e. initial state x, both point-process systems are nonexplosive on $[0,T]$, and the only differing intensity is the $i_\star$-coordinate. Since both target intensities are bounded below by $\bar{\mu}_\star > 0$, the conditional laws satisfy $K_S^T(x,\cdot) \ll K_0^T(x,\cdot)$. Hence, we can apply Jacod's [29] Girsanov formula for multivariate point processes ([11, Chapter VI] and [31] for more details) to the single differing coordinate $i_\star$,

$$\mathrm{KL}\big(K_S^T(x,\cdot)\|K_0^T(x,\cdot)\big) = \mathbb{E}_S\left[\int_0^T \left(\lambda_{i_\star}^{(S)}(t)\log\frac{\lambda_{i_\star}^{(S)}(t)}{\bar{\mu}_\star} - \lambda_{i_\star}^{(S)}(t) + \bar{\mu}_\star\right)dt \,\middle|\, X(0)=x\right]. \quad (73)$$

For $a \geq 0$ and $b > 0$, $a\log(a/b) - a + b \leq (a-b)^2/b$ and the resulting entropy term is finite after averaging over π_S with $b = \bar{\mu}_\star$ and Lemma C.5. Applying this with $a = \bar{\mu}_\star + \theta_- \sum_{j\in S} X_j(t-)$ and $b = \bar{\mu}_\star$ gives

$$\mathrm{KL}\big(K_S^T(x,\cdot)\|K_0^T(x,\cdot)\big) \leq \frac{\theta_-^2}{\bar{\mu}_\star}\int_0^T \mathbb{E}_S\Big[\Big(\sum_{j\in S} X_j(t)\Big)^2 \,\Big|\, X(0)=x\Big]\,dt.$$

Integrating over $\pi_S(dx)$ and applying stationarity yields

$$\int \mathrm{KL}\big(K_S^T(x,\cdot)\|K_0^T(x,\cdot)\big)\,\pi_S(dx) \leq \frac{\theta_-^2}{\bar{\mu}_\star}\int_0^T \mathbb{E}_S[\Sigma_S(t)^2]\,dt = \frac{\theta_-^2}{\bar{\mu}_\star}\,C_{\mathrm{path}}\,T.$$

The last claim follows from Lemma C.1 and Proposition C.1. □

C.4 Proof of Theorem 4.1

We now turn the KL bound into a lower bound over the hard subclass using Fano's inequality[23]. Let

$$M := |\mathcal{S}_{d,k}(i_\star)| = \binom{d-1}{k},$$

and let $S^\star$ be uniform on $\mathcal{S}_{d,k}(i_\star)$. The following finite-$d$ bound is the quantitative form used in the main text.

Proposition C.3. *For every $d \geq k+2$ and every $T > 0$,*

$$\inf_{\widehat{S}} \sup_{S\in\mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S(\widehat{S}(Y_T) \neq S) \geq 1 - \frac{C_{\mathrm{init}} + \frac{\theta_-^2}{\bar{\mu}_\star}C_{\mathrm{path}}T + \log 2}{\log\binom{d-1}{k}}.$$

Proof. For any estimator $\widehat{S} = \widehat{S}(Y_T)$, Fano's inequality gives

$$\sup_{S\in\mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S(\widehat{S} \neq S) \geq 1 - \frac{I(S^\star;Y_T) + \log 2}{\log M}.$$

The baseline mutual-information inequality in Lemma C.2, with $Q = P_0^T$, yields

$$I(S^\star;Y_T) \leq \frac{1}{M}\sum_{S\in\mathcal{S}_{d,k}(i_\star)} \mathrm{KL}(P_S^T\|P_0^T).$$

By Lemma C.1, Proposition C.1, and Proposition C.2, each summand satisfies

$$\mathrm{KL}(P_S^T\|P_0^T) \leq C_{\mathrm{init}} + \frac{\theta_-^2}{\bar{\mu}_\star}C_{\mathrm{path}}T.$$

Substituting this into Fano's inequality proves the claim.

□

We now prove the information-theoretic lower bound.

Proof of Theorem 4.1. Let C_{path} be the constant from Lemma C.5, and let C_{init} be the constant from Proposition C.1. Since k is fixed,

$$\log \binom{d-1}{k} = k \log d + O(1), \qquad d \to \infty.$$

Hence there are constants $a_k > 0$ and $d_1 \geq k+2$, depending only on k, such that

$$\log \binom{d-1}{k} \geq a_k \log d, \qquad d \geq d_1.$$

Choose $d_0 \geq d_1$ large enough that

$$\frac{C_{\mathrm{init}} + \log 2}{\log \binom{d-1}{k}} \leq \frac{1}{4}, \qquad d \geq d_0,$$

and set

$$c := \frac{a_k \bar{\mu}_\star}{4\theta_-^2 C_{\mathrm{path}}}, \qquad \varepsilon := \frac{1}{2}.$$

If $d \geq d_0$ and $T \leq c \log d$, then

$$\frac{\theta_-^2}{\bar{\mu}_\star} C_{\mathrm{path}} T \leq \frac{1}{4} \log \binom{d-1}{k}.$$

The finite bound in Proposition C.3 therefore gives

$$\inf_{\widehat{S}} \sup_{S \in \mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S(\widehat{S}(Y_T) \neq S) \geq \frac{1}{2}.$$

This proves the theorem with the above constants. The same argument also gives the corresponding full-class impossibility statement, since $\mathcal{G}_{d,k}^{\mathrm{sub}}(i_\star) \subseteq \mathcal{G}_{d,k}$. Indeed, any full support estimator $\widehat{G}(Y_T)$ induces a row estimator $\widehat{S}(Y_T)$ by taking its estimated support in row $i_\star$. On the subclass, the event $\{\widehat{S}(Y_T) \neq S\}$ is contained in the event that the full support is estimated incorrectly, and therefore

$$\inf_{\widehat{G}} \sup_{(\mu,\Theta) \in \mathcal{G}_{d,k}} \mathbb{P}_{\mu,\Theta}\Big(\widehat{G}(Y_T) \neq \mathrm{supp}(\Theta)\Big) \geq \inf_{\widehat{S}} \sup_{S \in \mathcal{S}_{d,k}(i_\star)} \mathbb{P}_S\Big(\widehat{S}(Y_T) \neq S\Big).$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by the theoretical analysis in Section 3 and Section 4 and the corresponding proof details in Appendices A and C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We state the assumptions needed for the theoretical results in each lemma, proposition, and theorem. Additionally, the conclusion discusses the main limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We state the assumptions needed for the theoretical results in each lemma, proposition, and theorem. Appendices A and C provide the full proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper is compliant with the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Due to the theoretical nature of the paper, there is no direct broad impact to declare.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use any existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The paper does not involve LLMs in a non-standard way.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.