使用基于集合的滤波识别由计数数据驱动的霍克斯过程影响网络

N. Santitissadeekorn[1],S. Delahaies[1],和 D. J. B. Lloyd[1,2]

[1]萨里大学数学系，英国吉尔福德 GU2 7XH[2]萨里大学犯罪学中心，英国吉尔福德 GU2 7XH

2023年2月2日

arXiv:2302.00647v1 [cs.SI] 1 Feb 2023

摘要

许多网络具有事件驱动动力学（例如通信、社交媒体和犯罪网络），在这些网络中，网络某个节点上事件的平均发生率会随着网络中其他事件的发生而变化。具体而言，与某个网络节点相关的事件可能会提高其他节点上事件的发生率，具体取决于节点间的影响关系。因此，人们希望利用时间数据揭示给定网络中具有方向性且随时间变化的影响结构，同时量化不确定性，即使缺乏物理网络知识也是如此。通常，推断网络影响结构的方法需要了解物理网络，或者只能推断小型网络结构。本文使用多维霍克斯过程对网络中的事件驱动动力学建模。随后，我们针对计数数据时间序列（即提供网络中每个节点单位时间内事件数量的数据）开发了一种新颖的基于集合的滤波方法。该方法不仅能跟踪影响网络结构随时间的变化，还能通过集合离散度近似量化不确定性。该方法克服了现有方法的若干缺陷，例如，现有的多维霍克斯过程推断方法速度过慢，无法实际应用于节点数超过 ~ 50 的网络；只能处理时间戳数据（即仅记录事件发生时间、不记录每个节点事件数量的数据）；而且我们的方法无需预先掌握物理网络。我们的方法具有高度可并行化的特点，可用于推断大型网络（$\sim 10,000$个节点）的影响结构。我们使用合成数据和真实世界电子邮件通信数据，展示了该方法在大型网络中的应用。

1 引言

识别并建模网络中的影响结构，是网络分析中的一项重要任务。“影响”的解释可能取决于观测数据或网络类型。对影响的一种常见定义是：将有效传播信息的节点视为影响者。衡量信息传播需要了解网络连接之外的更多信息。例如，在通信网络中，如果 A 与 B 交谈，这一行为是否会提高 B 与网络中其他人交谈的概率?

实际上，在量化影响时，可能必须考虑通信内容。然而，大多数数据中可能无法获取这些内容（例如，手机或电子邮件数据可能只包含

时间戳、发送方和接收方）。我们的讨论将仅聚焦于无内容数据。识别和建模影响的一种常见方法，是基于信息级联的影响力最大化思想[5, 15]。具体而言，该方法会找出一个 K 节点子集，使选定节点传播信息后最终触达的节点数的期望值最大化。这是一个 NP难问题，目前的研究重点是提升该算法在大规模网络中的可扩展性[2, 8, 16]。近期的进展还将信息从一个节点传播到另一个节点所需的时间纳入考量[1] ，或处理随机布尔网络等特殊类型的网络[21]。这种方法的主要缺陷之一是，其算法必须以物理网络作为输入；而我们下面将要讨论的工作，只需要计数数据时间序列，就能构建出边权重表示影响力强度的网络。研究网络影响结构的另一种知名方法是最小生成时序树[13, 20]。这种方法可用于发现时序网络中的持久级联模式。这里的时序网络由观测数据构建而成，其中包含“发送方”“接收方”、时间戳以及事件持续时间，这是一种常见的数据格式，适用于手机通信等多种数据。级联模式为推断影响结构提供了重要信息。这种方法纳入了信息传播的时序知识（例如，“突发”事件通常由影响者的活动驱动，并对信息传播产生显著影响[18]）。它基于与流行病传播类似的概念，旨在发现信息扩散中持久的级联模式。

另一种基于霍克斯过程 [11] 的方法也已被提出，用于回答有关识别和建模影响网络结构的类似问题。不过，该方法主要关注如何根据个体某些可观测事件的时间序列构建影响网络。例如，数据可能只包含 A 发出的电子邮件的时间戳，而没有收件人信息。然而，如果 A 对 B 有很强的影响，我们可能会观察到 B 在 A 活动之后立即出现异常频繁的活动。（多元）霍克斯模型被广泛用于对此类影响进行建模，并应用于许多场景（例如社会科学中的意见网络[4], 、地震网络 [28], 、电子邮件网络 [29], 或医疗保健网络 [27]）。该方法通过霍克斯模型的某些参数推断网络影响结构，而无需了解物理网络；因此，这一问题转化为参数估计问题。请注意，多维霍克斯过程的参数数量通常为 $O(N^2)$，其中 N 是网络中的节点数量。这给大规模网络推断带来了重大挑战。通常采用最大似然原理，并已针对这一问题开发了各种基于期望最大化（EM）的算法[10, 19, 25, 29]。在 [17], 中，研究者提出了一种完全贝叶斯方法，通过马尔可夫链蒙特卡罗（MCMC）估计非参数霍克斯过程的参数，重点在于潜在网络重建。该方法不仅可以为参数纳入先验模型，也可以为网络结构纳入先验模型。当节点之间存在“距离”这一概念时，先验网络结构可能会很有用（例如，节点表示空间位置，且预期不存在远程影响）。 [14]中提出了一个类似的贝叶斯框架，通过随机期望最大化估计后验众数。这些现有方法仅适用于时间戳数据（即每个事件的发生时间），而且通常可能不适用于大规模网络。

在某些情况下，时间戳数据可能不可用，只能将给定时间间隔内的事件数量汇总为单位时间内的计数数据。本文的主要目标是提出一种适用于由计数

数据驱动的多维霍克斯过程的网络推断方法。计数数据可涵盖时间间隔内任意数量的事件，能够大规模并行处理、量化不确定性，且无需了解物理网络。图1展示了本文的核心思想：利用固定时间间隔内观测到的 6 个数据源的计数数据序列，构建有向网络，其中连边表示节点之间的影响。揭示方向性影响，有助于深入了解关键影响者引发活动时网络动态可能如何演变。例如，在所谓的隐蔽网络中（如参与犯罪或恐怖活动的人之间的通信），有影响力的人往往会避免产生过多通信或活动，以免暴露身份。因此，尽管与这些影响者直接相关的活动可能很少，但其他节点会迅速对这些活动作出响应，从而在影响者活动发生后立即引发事件级联，并随着时间推移逐渐消退。

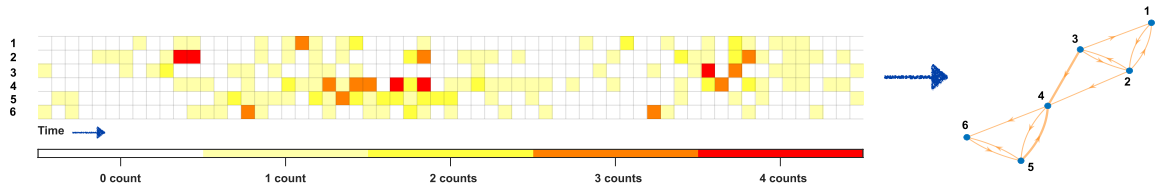


图1： 根据计数数据序列重构影响网络的示意图。在本例中，有 *5* 个数据源，计数数据是在均匀的时间间隔内收集的。

本研究采用由计数数据驱动的多维霍克斯过程，该过程允许两个或更多一维霍克斯过程相互作用，用以描述网络的影响结构。网络中的每个节点都会生成有限数量的事件，其条件发生率会根据整个网络中事件的发生情况而变化。具体而言，有向网络中的加权边表示节点之间的定向影响，这些影响由多维霍克斯模型参数化。因此，本文的网络推断可转化为多维霍克斯过程的参数估计问题。需要注意的是，多维霍克斯过程的参数数量通常为 $O(N^2)$，其中 N 是网络中的节点数。这给大规模网络推断带来了重大挑战。

大规模影响网络的重构方法需要在最优性与计算效率之间权衡。我们采用了 [23] 中近期提出的一种集合方法，用于计数数据的序贯数据同化算法。与大多数基于最大似然估计、使用大量时间戳数据进行推断的现有方法不同，这种基于集合的滤波是一种在线跟踪工具，无需对大批量数据的似然函数进行计算密集型评估。因此，它有望提升大规模问题的计算可扩展性。此外，将模型参数应用于真实世界数据时，参数往往会发生变化，而滤波系统能够跟踪这些变化 [22, 23]。通过集合离散度，所提出的方法还将提供条件强度和网络结构不确定性的近似，这有助于决策过程。

本文结构安排如下：第2节介绍用于描述影响的多维霍克斯过程模型，第3 节介绍集合滤波方法，用于依次推断多维霍克斯过程的参数和影响网络结构。第4节介绍将集合滤波方法应用于合成网络的情况，其中包括不完美模型情形。随后，我们在第5节将集合滤波方法应用于真实世界数据。

并最终在第6节得出一些结论。

2 由计数数据驱动的 Hawkes 过程与网络

本节介绍一种由计数数据驱动的多维霍克斯过程，它在某种程度上是传统连续时间霍克斯过程的离散时间版本。不过，我们强调，选择这一模型主要是出于其由计数数据驱动的动态特性，这与本研究的主要目标相契合。

我们考虑一个包含 m 个节点的网络，并令 $\{\Delta N_k^j\}$ 在 $j = 1, \ldots, m$ 和 $k = 0, 1, \ldots$ 表示第 j 个节点在长度为 δt 的第 $k-$ 个时间间隔内产生的事件总数，即 $[(k-1)\delta t, k\delta t)$。截至第 $k-$ 个时间间隔的事件数历史记录记为 $\mathcal{H}_k := \{\Delta N_1^j, \ldots, \Delta N_k^j\}$。我们假设，时间间隔 δt 内的事件数服从均值为 $\lambda_k^i \delta t$ 的泊松分布，其中 λ_k^i 表示第 $i-$ 个节点 $\lambda^i(t)$ 在 $t \in [(k-1)\delta t, k\delta t)$ 的（常数）条件强度。我们用双随机过程 [3] 对条件强度建模，其条件强度如下：

$$\lambda_{k+1}^i = \mu^i + (\lambda_k^i - \mu^i)(1 - \beta^i \delta t) + \sum_{j=1}^{m} \alpha^{ij} \Delta N_k^j, \tag{2.1}$$

其中，我们假设对所有节点均有 $\lambda_0^i = \mu^i$，且 μ^i，β^i，$\alpha^{ij} > 0$。参数 μ_i 表示基准强度（即强度过程的下界），β_i 是衰减率。α^{ij} 用于量化节点 j 对节点 i的影响力强度。因此，α^{ij} 的值就是网络中边的权重，表示第 $j-$ 个节点对第 $i-$ 个节点的影响强度。换言之，α^{ij} 的值越大，意味着在第 j 个节点的一代或多代事件发生后，第 i 个节点立即产生事件的倾向越高。请注意，该模型是具有指数衰减核 [11] 的多维霍克斯过程的离散版本。

3 集合滤波器与网络不确定性

我们的主要关注点是一种序贯数据同化方法，在该方法中，事件数量被依次用于联合估计条件强度和模型参数。模型（2.1）将作为数据同化的预报模型。诸如粒子滤波器（PF）之类的通用序贯数据同化算法 [6] 在高维问题中可能难以实施，而在尝试对大型网络进行推断时，通常会遇到这类问题。作为替代方案，集合卡尔曼滤波（EnKF）被开发为一种次优滤波算法，并已广泛应用于许多领域 [9]。然而，由于泊松似然固有的异方差，在泊松计数数据中的应用并不简单；相关讨论和数值示例参见 [23]。我们采用近期在 [23]中提出的集合泊松-伽马滤波器（EnPGF）。下面简要回顾 EnPGF 的基本思想，一些关键公式的详细推导见附录。

简而言之，EnPGF 算法主要分为两个阶段。第一阶段，根据下文的随机方程（3.2）调整条件强度的先验集合成员。此过程

中的集合重新分布（3.2）符合泊松-伽马共轭性。随后，更新后的条件强度集合将用作下一阶段的“观测值”。第二阶段，将条件强度方面的创新项对相应参数的集合进行线性回归。该算法在若干条件强度过程模型中表现良好，参见 [23]。下文将介绍 EnPGF 的数学描述及其在多维霍克斯过程中的应用。

为避免记号过于繁杂，我们暂时省略时间和位置索引，并使用ΔN 和 λ 表示给定节点在给定推断时刻的观测事件数。 λ 的经验分布由一个集合表示。我们还为$s = 1, \ldots, M$引入上标（s），以表示集合成员，其中 M为集合规模。假设先验集合由 $\lambda^{(s)}$ 组成，其集合均值为 $\bar{\lambda}$。令 P 为集合方差，$P_r = P/\bar{\lambda}^2$ 为集合相对方差。如果 $\lambda^{(s)}$ 服从具有上述均值和相对方差的伽马分布，且 ΔN 具有均值为 $\lambda\delta t$的泊松似然函数，则记为 $\lambda^{(s),a}$的后验样本服从伽马分布，其均值为 $\bar{\lambda^a}$ ，相对方差为 P_r^a，并满足以下条件：

$$\begin{aligned} \bar{\lambda^a} &= \bar{\lambda} + \frac{\bar{\lambda}}{P_r^{-1} + \bar{\lambda}\delta t}(\Delta N - \bar{\lambda}\delta t) \\ (P_r^a)^{-1} &= P_r^{-1} + \Delta N. \end{aligned} \tag{3.1}$$

如 [23],所示，与（3.1）一致的后验样本满足以下公式：

$$\frac{\lambda^{(s),a} - \bar{\lambda^a}}{\bar{\lambda^a}} = \frac{\lambda^{(s)} - \bar{\lambda}}{\bar{\lambda}} + \frac{P_r}{P_r + (\Delta N)^{-1}}\left[\frac{\Delta N_e^{(s)} - \Delta\bar{N}_e}{\Delta\bar{N}_e} - \frac{\lambda^{(s)} - \bar{\lambda}}{\bar{\lambda}}\right], \tag{3.2}$$

其中 $\Delta N_e^{(s)}$称为扰动观测，独立地从伽马分布中抽取¯其均值为 ΔN，方差为 **1**，适用于 $i = 1 \ldots, M$其样本均值记为 ΔN_e。令$A^{(s),\ a} = \left(\frac{\lambda^{(s),a} - \bar{\lambda^a}}{\bar{\lambda^a}}\right)$由式（3.2）右侧计算得出。因此，我们可以（随机地）按以下方式更新每个集合成员

$$\lambda^{(s),a} = \lambda^{(s)} + A^{(s),a}\bar{\lambda^a}. \tag{3.3}$$

为阐明下文的算法，我们现在补全所有节点和时间索引。我们定义向量$\mathbf{q}_k^i$ ，它收集第 $i-$ 个节点的所有参数估计（即 μ^i， β^i， α^{ij} ，对所有 j 而言），在时间步k处。参数向量的第 $s-$ 个集合成员记为 $\mathbf{q}_k^{i\ (s)}$，其中 $s = 1$，.....， M。类似地，我们定义 $\lambda^{i,\ (s)}_{\ k}$ 表示条件强度的第 $s-$ 个集合成员。我们还定义向量 $\Delta N_k := [\Delta N_k^1, \ldots, \Delta N_k^m]$。随后，可针对每个节点独立实施所提出的算法，简要总结如下。

步骤 1 给定第 $k-$ 步的集合 $\left(\lambda_k^{i,(s)}, \mathbf{q}_k^{i\ (s)}\right)$ ，以及 ΔN_k 我们通过将 $\lambda^{i,\ (s)\ f}_{\ k+1}$ ，代入 $\lambda^{i,\ (s)}_{\ k}$ 和 $\mathbf{q}_k^{i\ (s)}$ ，得到（2.1）。

步骤 2 一旦 ΔN_{k+1} 可用，我们便实施（3.2），从而得到 $\lambda^{i,\ (s),\ a}_{\ k+1}$。

步骤 3 随后将条件强度的新集合 $\lambda^{i,\ (s),\ a}_{\ k+1}$ 用作观测值，以更新第 $i-$ 个节点的模型参数。此操作通过集合卡尔曼滤波器（EnKF）完成，具体将在下文说明；结果是得到一组新的模型参数集合$\mathbf{q}^{i,\ (s),\ a}_{\ k+1}$ 。

步骤 4 设置 $\lambda^{i,\ (s)}_{k+1} = \lambda^{i,\ (s),\ k+1\ a}$ 以及 $\mathbf{q}^{i\ (s)}_{k+1} = , \mathbf{q}^{i,\ (s),\ k+1\ a}$。使用集合 $\left(\lambda^{i,(s)}_{k+1}, \mathbf{q}^{i\ (s)}_{k+1}\right),$ 和 ΔN_{k+1} 重复步骤 1–3。

现在我们描述上述步骤 3 的实现，该步骤同样在给定时间针对每个节点分别执行。因此，为方便起见，我们暂时省略节点和时间索引。令 $\bar{}\mathbf{q}_e$ 为所有$\mathbf{q}^{(s)}$的集合均值，即 $\bar{}\mathbf{q}_e = M^{-1}\sum_{s=1}^{M}\mathbf{q}^{(s)}$。对于每个 s，我们可以计算偏差$\mathbf{q}^{(s)} - \bar{}\mathbf{q}_e$，并将它们按列堆叠，形成缩放偏差矩阵：

$$\mathbf{X} = \frac{1}{\sqrt{M-1}}[\mathbf{q}^{(1)} - \bar{\mathbf{q}}_e, \ldots, \mathbf{q}^{(M)} - \bar{\mathbf{q}}_e] \tag{3.4}$$

类似地，我们可以计算 $\lambda^{(s)}$ 和 $\lambda^{(s)}$的缩放偏差矩阵^a，分别记为 M–维行向量 $\mathbf{Y}$ 和 $\mathbf{Y}_a$。随后，可以更新样本 $\mathbf{q}^{(s)}$，从而获得新样本 $\mathbf{q}^{(s),\ a}$ 其形式类似于“随机”EnKF 公式 [9]:

$$\mathbf{q}^{(s),a} = \mathbf{q}^{(s)} + \mathbf{X}\mathbf{Y}^\top(\mathbf{Y}\mathbf{Y}^\top + \mathbf{Y}_a\mathbf{Y}_a^\top)^{-1}(\lambda^{(s),a} - \lambda^{(s),f}). \tag{3.5}$$

注意，(3.5) 所需的求逆非常简单，因为它是一个标量。通过$\mathbf{X}$ (3.5) 进行左乘后，新集合被限制在由$\mathbf{X}$的列所张成的子空间内。换言之，对于任何在$\mathbf{X}$的列空间正交空间上具有非零投影的项，该算法缺乏进行更新所需的信息。在高维问题中，这一问题可能难以避免，因为集合规模往往远小于问题的维度。尽管如此，上述更新在由$\mathbf{X}$张成的子空间内给出了最优解，但通常在参数空间中并非最优。在参数彼此不相关的特殊情况下，$\mathbf{X}$ 的非对角项可以降为零，并且强度创新可以分别对每个参数进行回归。

4 合成实验

4.1 实验 1：完美模型

在本实验中，我们展示了 EnPGF 在估计（2.1）参数方面的应用，尤其关注网络参数 α^{ij}。数据生成模型为离散时间霍克斯过程（2.1）；因此，正如第 3 节所述，该模型与用于传播集合的模型相同。除 δt外，我们假设所有参数均未知。我们关注模拟隐蔽网络的参数设定，即网络中的某些单元具有相对较低的基线率，但具有较大的 α^{ij}，对应某个 $j \neq i$。为此，我们将节点数 m 设为 6，因此共有 48 个未知参数。

我们将真实的基准速率设为： $j \in \{1,2,3,5,6\}$ 对应 $\mu^j = 2s_1$， $\mu^4 = 0.75s_1$ 对应常数$s_1 > 0$，并将所有 j的衰减率设为 $\beta^j = 5$。激发矩阵 α^{ij} 为

$$\alpha = s_2 \begin{bmatrix} 1 & 0.5 & 0.5 & 0 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 & 0 \\ 0 & 1 & 1/5 & 0 & 0 & 0 \\ 0 & 1 & 2.5 & 0.5 & 2.5 & 0 \\ 0 & 0 & 0 & 0.4 & 1.5 & 0.5 \\ 0 & 0 & 0 & 0.4 & 0.5 & 1.5. \end{bmatrix}$$

对应常数 $s_2 > 0$。为考察不同情形，我们考虑 4 组参数： $s_1 = s_2 = 1.5$,$s_1 = 1.5$， $s_2 = 0.5$, $s_1 = 0.5$， $s_2 = 1.5$ 和 $s_1 = 0.5$， $s_2 = 0.5$。 $s_1 = s_2 = 1.5$ 这一情形将产生最高的事件速率。另一方面， $s_1 = 0.5$， $s_2 = 0.5$ 的事件速率将最低。情形 $s_1 = 1.5$， $s_2 = 0.5$ 将考察滤波器在激发率较低时估计网络结构的能力；情形 $s_1 = 0.5$， $s_2 = 1.5$ 则用于考察基线较低时的情况，这可能会影响网络重构，也可能不会。

使用上述参数生成数据，并通过 $\delta t = 0.1$ 运行 2000 步。实现 EnPGF 需要一个初始系综。我们选取的集合规模为 $M = 500$，并从均值为 $4s_1$ 、方差为 8 的伽马分布中，为每个节点独立抽取基线和衰减率。参数 α^{ij} 则从均值为 s_2 、方差为 $1/4$的伽马分布中独立抽取。

估计参数的集合均值所带来的误差降低情况如图2所示。具体而言，计算每个参数的均方根误差（rmse），并以初始系综均值的均方根误差进行归一化。因此，数值为 1 或更大表示误差没有降低，而远小于 1 的数值则表示误差显著降低。对于每个 i个节点，激发率的误差报告为 α^{ij}在 $j = 1$、.....、6 上的平均误差。同样，图3展示了每个参数的方差缩减。图4还展示了激发矩阵的集合均值。

如图 2所示，只有在情景 $s_1 = s_2 = 1.5$下，所有节点的误差降低幅度都很大。不过，所有情景下基线率的误差降低幅度都相当相似。就图 3所示的方差缩减而言，情景 $s_1 = s_2 = 1.5$的缩减速率明显慢于其他情形，尽管其误差降低幅度很大。然而，这是意料之中的，因为在观测到没有事件时，泊松-伽马推断总是会降低方差，但在其他情况下，方差则可能增大。因此，在 λ^j 方面，基线率和激发率较高时，集合离散度预计会更大，因为这会比其他情形产生更多非零观测值。因此，尽管在（3.5）中基于集合的回归总能实现方差缩减，该回归使用 λ^j 的更新作为新息项来估计参数，但在 $s_1 = s_2 = 1.5$的情况下，方差缩减速率可能较低。图 4 展示了使用集合均值重构的网络；显然，只有当基线率和激发率都足够高时，重构结果才可靠。不过，在所有情景下，仍然都能捕捉到α^{ij} 的较大取值。

在图5中，我们通过固定$s_1 = 1.5$ 并改变 s_2，考察激发率对网络重构的影响。激发矩阵的误差使用弗罗贝尼乌斯范数计算，并按 $1/s_2$进行缩放。显然，随着网络总体激发率下降，重构效果会变差。我们还通过固定 $s_2 = 1.5$ 并改变 s_1 ，考察基线率对网络重构的影响，如图6所示。同样可以观察到性能下降的趋势，但误差略小，对较低基线率的容忍度也略高。

我们对一个包含 300 个节点的大型网络进行了类似实验。根据“真实”参数值生成的合成数据长度为 150,000 步，步长为 0。1。所有节点的真实衰减率均为 7，基线率则从均值为$20/3$ 、方差为 $200/9$的伽马分布中随机抽取。真实激发矩阵如图7所示，同时图中还展示了网络估计，见图7。

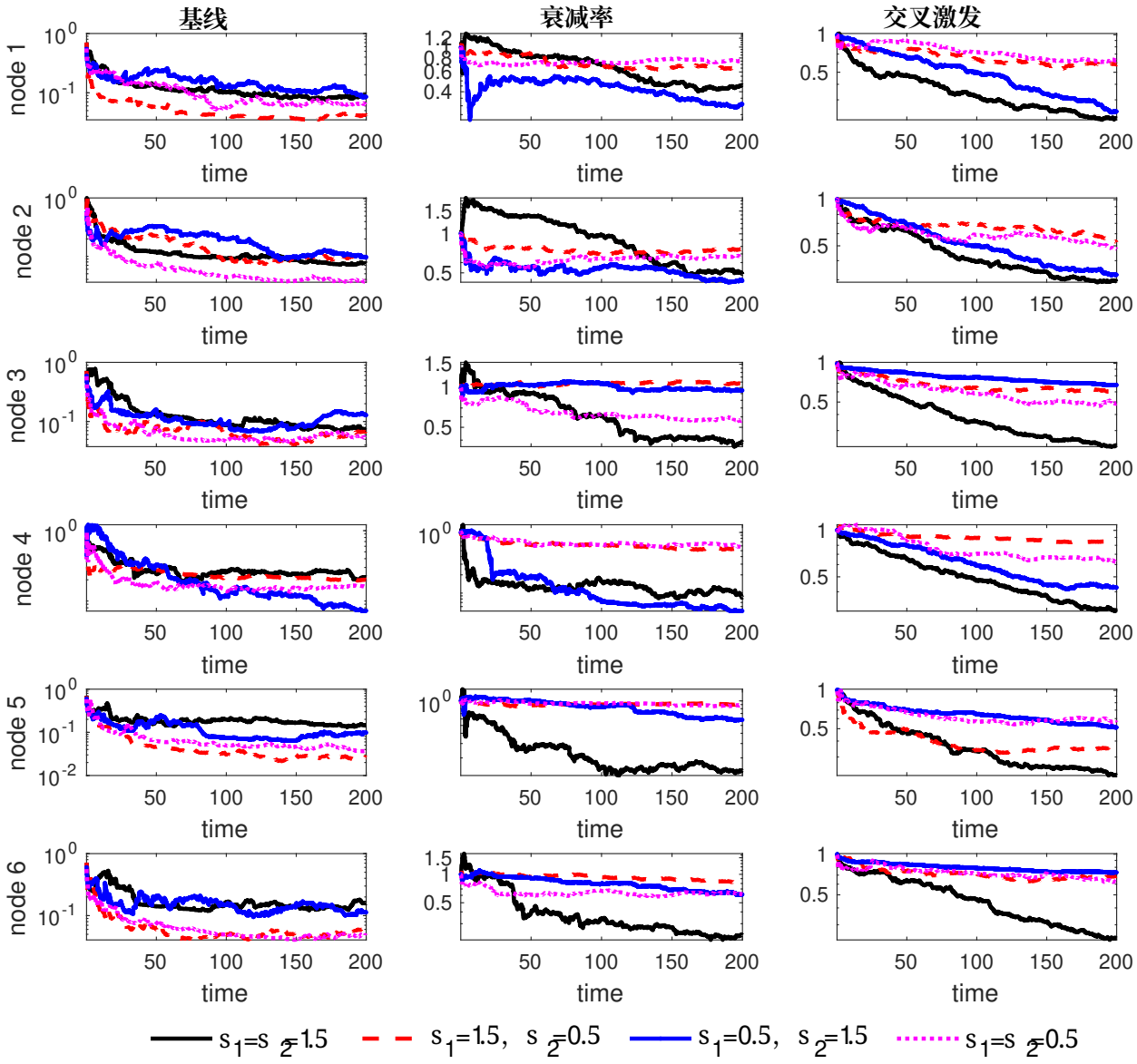


图2： 完美模型合成实验中所有 *4* 种情况的误差降低传播，以对数刻度绘制。1 以上的值表示初始误差未降低，而低于 *1* 的值则表示误差的降低程度。

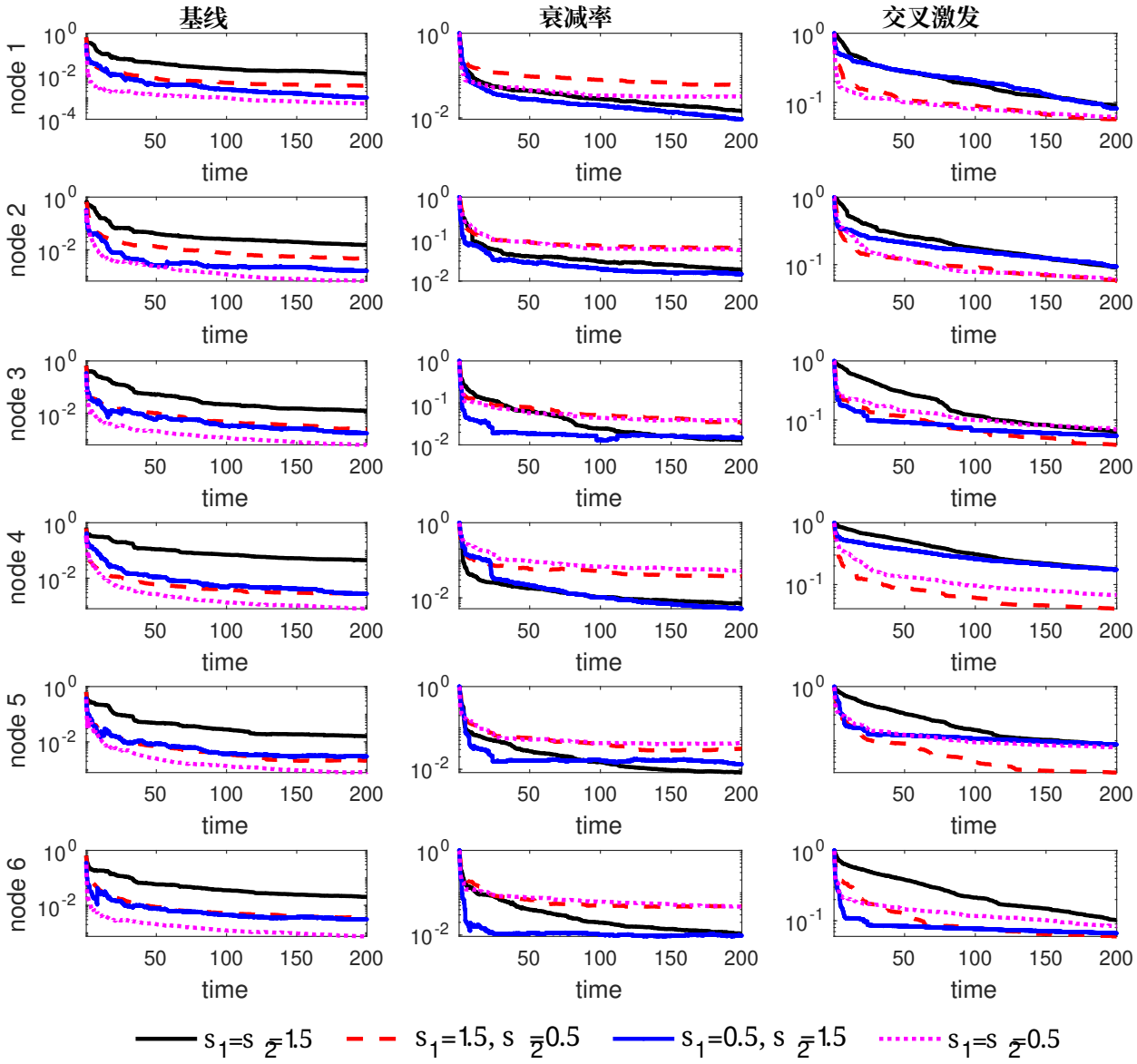


图3: 完美模型合成实验中所有*4*种情形的方差缩减传播情况，以对数刻度绘制。大于1 的值表示初始方差未缩减，而小于*1*的值表示方差缩减的程度。

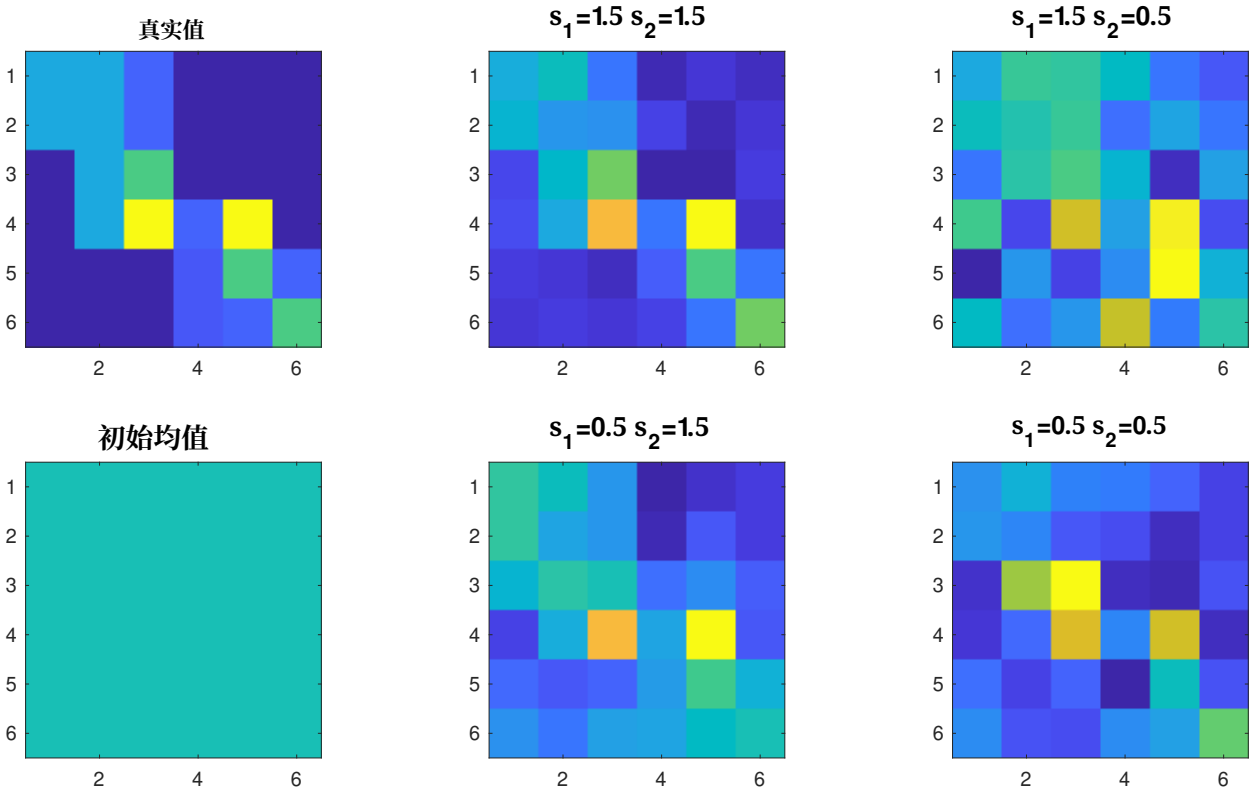


图4： 完美模型合成实验中估计的网络与真实网络结构及初始条件的比较。

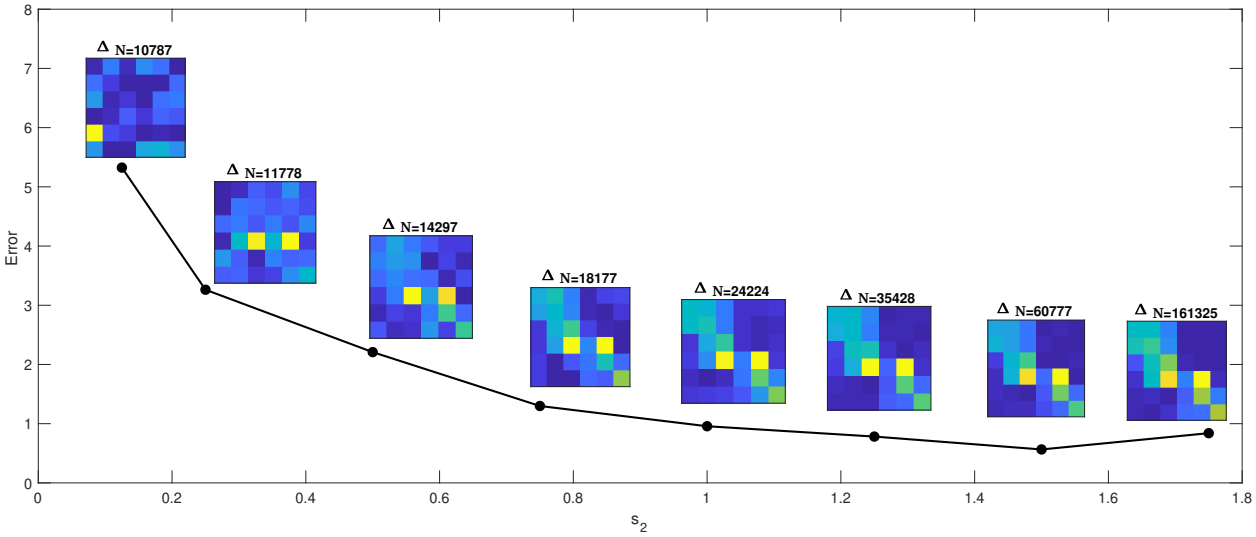


图5： 完美模型合成实验中，不同 s_2 值对应的激发矩阵误差，以缩放的弗罗贝尼乌斯范数衡量。同时报告测试数据中的计数数量 ΔN。

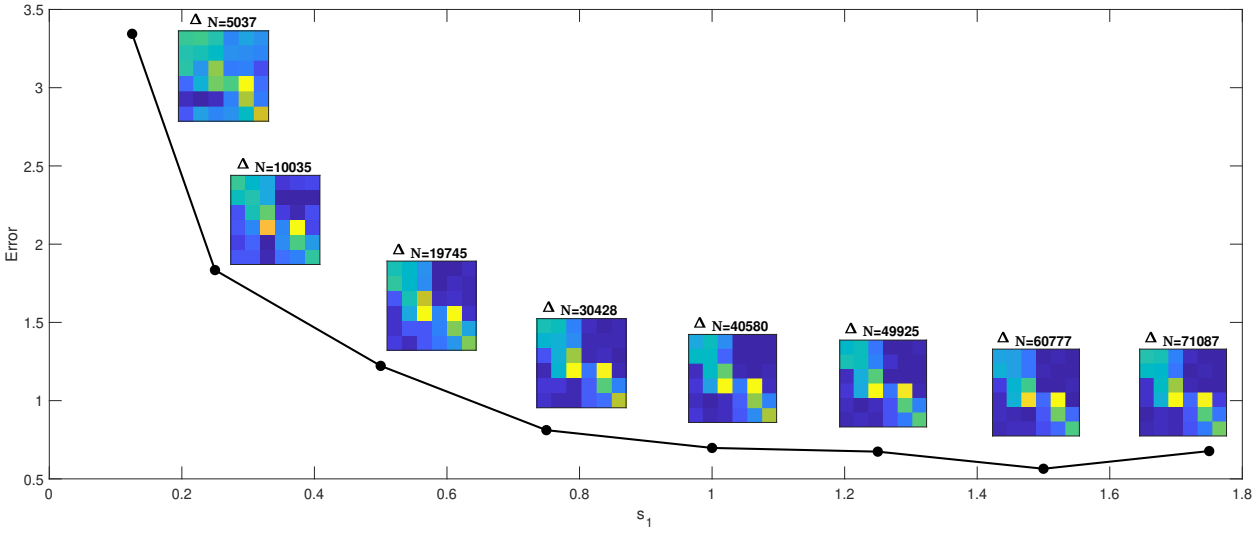


图6： 在完美模型合成实验中，对于不同的 s_1 值，使用缩放的弗罗贝尼乌斯范数衡量激发矩阵的误差。另报告测试数据中的计数数量 ΔN。

4.2 实验2：非完美模型

在本实验中，我们使用基于主体的模型（ABM）生成数据，但假设预测模型为（2.1）；因此，这是非完美模型情形。因此，本实验中 ABM 使用的参数不能视为（2.1）模型参数的真实值。我们采用了文献 [24]，中开发的 ABM，该模型允许事件通过激发过程在空间网络中产生扩散效应，并用于模拟城市犯罪。具体而言，某个节点发生事件后，不仅会提高该节点未来事件的强度，也会提高其邻居未来事件的强度，且这种关系不随时间变化。邻居节点之间的激发程度可视为一个加权有向网络。我们希望通过（2.1）揭示这一潜在网络，因此本实验主要关注参数 α^{ij} 的估计。

我们简要讨论 [24] 中引入的有限位置集（或节点）上的基于主体的模型（ABM）。此处的 ABM 由两个主要部分组成：位置 s 和主体，主体会随机生成一个“事件”，并从一个位置移动到另一个位置。主体根据以下规则，决定是否在时间 t 和 $t+\delta t$ 之间于位置 s 生成事件：

$$p_s(t) = 1 - \exp(-A_s(t))\delta t,$$

其中， $A_s(t)$ 称为时间 t 时位置 s 的“吸引力”。假设该概率具有以下形式：

$$A_s(t+\delta t) = \mu_s + \left[(1-\eta)B_s(t) + \frac{\eta}{z}\sum_{s'\sim s} B_{s'}(t)\right](1-\omega\delta t) + \sum_{s'\sim s} w(s,s')E'_s(t).$$

参数 μ_s、 ω 和 η 是静态参数，分别表示基线、衰减率、扩散率、邻居中的节点数以及激发率。项 $B_s(t)$ 是一个动态分量，其中 $B_s(0) = 0$，而 $E_s(t)$ 是位置 s 在 t 和 $t+\delta t$ 之间的事件数。参数 $w(s, s')$ 定义了位置 s' 上的事件对位置 s 的影响，以及当 $w(s, s') > 0$ 时的影响。参数 z 是给定 s 时 $w(s, s') > 0$ 的数量。因此，吸引力场的动态与多维霍克斯过程非常相似。两者的主要区别在于霍克斯过程中没有扩散。我们定义一个

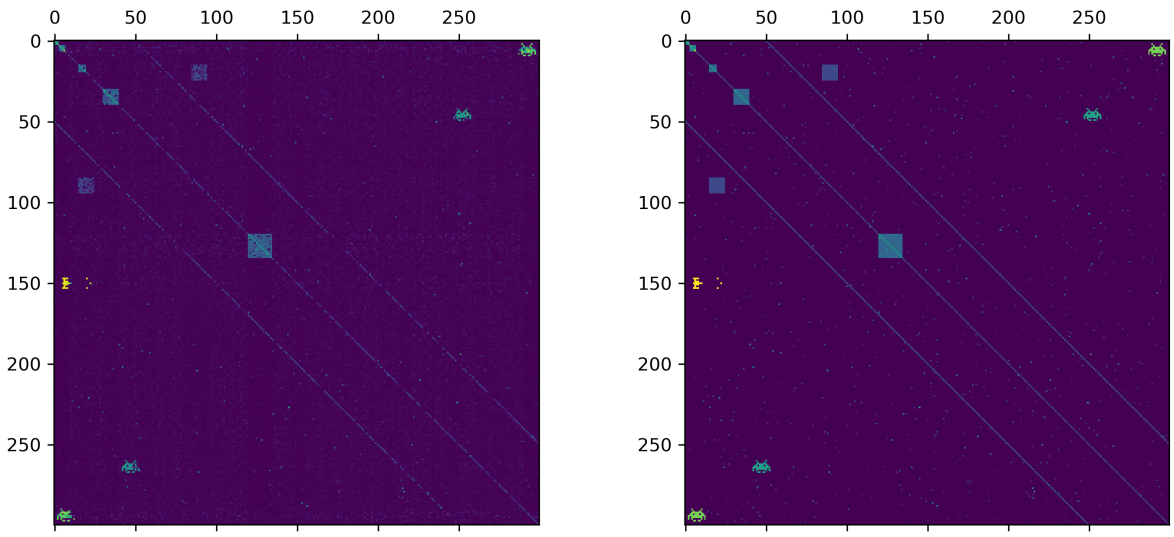


图7: （左）包含*300*个节点的网络的估计激发参数。（右）包含*300*个节点的网络的真实参数

矩阵 $\mathbf{W}$ ，其中 $w(s,\ s')$ 位于行 s 和列 s'，可将其视为该基于主体的模型的底层激发结构。

移　　动概率，即主体从位置 s 移动到 s'' （位于 s 的邻居位置之一）的概率，为

$$q(s,s'';t)=\frac{A_{s''}(t)}{\sum_{s'\sim s}A_{s'}(t)}.$$

如果位于位置 s 的主体在时间 t 决定不生成事件，那么在时间 $t+\delta t$ ，该主体将根据上述概率规则移动到 s 的一个相邻位置。因此，网络上的移动会偏向吸引力更高的位置。另一方面，如果主体生成了事件，该主体就会从仿真中移除。每个位置还会以速率 Γ 生成一个新主体（即均值为 Γ的均匀分布）。

我们测试了一个网络规模为 $m=6$ 的问题，并在测试实验中使用以下参数值： $\omega=5$， $\eta=1/4$, $\delta t=1/10$， $\Gamma=3$。除 μ_s外，所有位置 $s=1$、.....、6 的基线率均为 2。集合规模和初始系综的设置与完美模型情形中完全相同，见第4.1节。激发结构 $\mathbf{W}$ 如下所示：

$$\mathbf{W}=\begin{bmatrix}3&3&0&0&0&0\\3&3&0&0&0&0\\0&0&3&3&0&0\\0&6&6&1.5&6&0\\0&0&0&0&3&3\\0&0&0&0&3&3\end{bmatrix}. \tag{4.1}$$

通过上述参数值生成的 39964 个时间步测试数据如图8所示。请注意，该网络可视为隐蔽网络的一个玩具示例，其中位置$s=4$ 的事件频率相对较低，但它是网络中主要的影响位置。尽管基于主体的模型中的激发结构$\mathbf{W}$ 不能被准确地视为

激发参数 α^{ij} （2.1）的真实值，但预期二者具有相似的结构并非不合理。如图9所示，尽管 ABM 与霍克斯模型的机制不同，估计得到的激发网络结构仍然相似。结构中唯一显著的误识别是激发项 $w(5,\ 3)$。

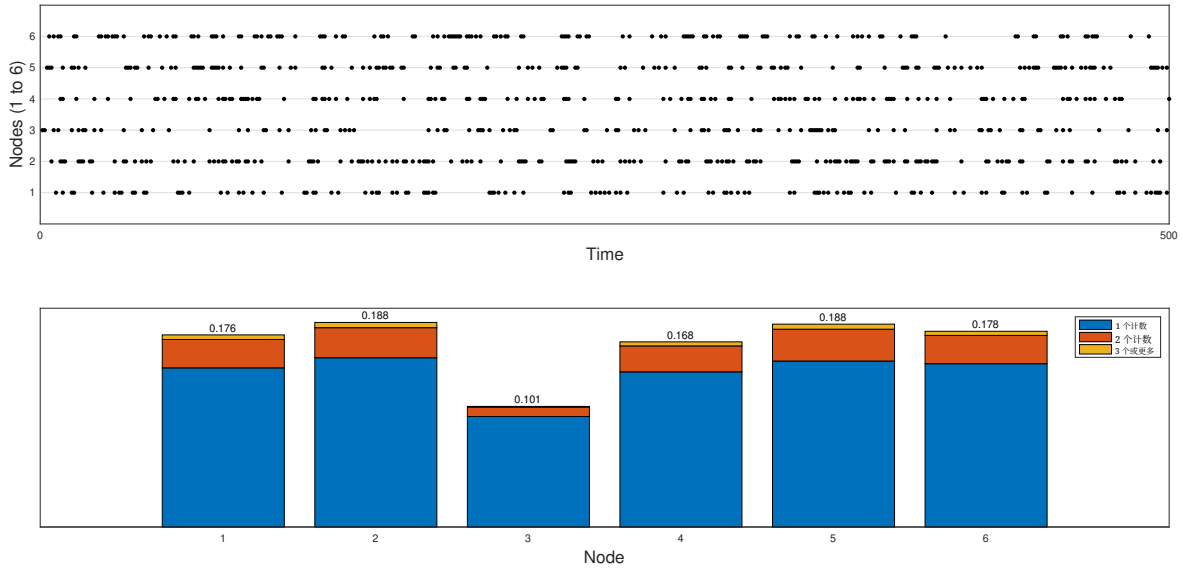


图8：*ABM*测试数据的探索。（上）展示了仅发出邮件的计数数据时间序列。圆点表示计数非零的事件。（下）柱子的高度与每个节点发出邮件的百分比成正比，该百分比也标注在柱顶。每根柱中堆叠部分的高度与图中标示的事件数量成正比（即*1*次计数、*2*次计数以及*3*次或更多次计数）。

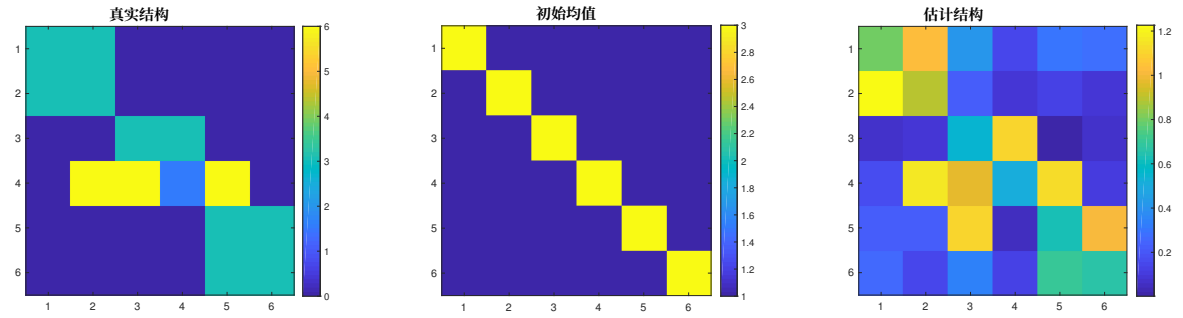


图9：非完美模型实验：（左）*ABM*网络结构。（中）矩阵 α^{ij}的初始集合均值。（右）在最后一个数据同化步骤中 α^{ij} 的集合均值。

5 真实世界数据

5.1 Ikenet 数据

Ikenet 数据集包含 22 位匿名志愿者（以 $\{1,\ldots,22\}$ 标记）在自 2010年5月 起的一年期间内电子邮件往来的时间戳。共有 6678 次电子邮件往来，不包括发给自己的邮件。数据可以用网络形式进行可视化，其中节点代表志愿者。具体而言，我们绘制了一个无向网络，任意两位志愿者之间的连边权重由他们之间电子邮件往来的总量决定；

有向网络中的有向连边则以发给指定志愿者的电子邮件数量作为权重。为便于清晰地可视化，我们仅绘制一个子网络，其中只包含出现次数高于所有边平均出现次数两倍的边，见图 11。显然，(9,18) 这一对是最重要的一对，且两个方向上的权重相近。

Fox 等人 [10] 研究了 Ikenet 数据，其中每个人的发送或接收活动均采用具有周期性基线率的霍克斯过程建模。结果表明，与齐次泊松模型相比，该模型的赤池信息准则（AIC）更优。Fox 等人 [10] 使用最大似然估计拟合模型参数。Zipkin 等人 [29]，使用霍克斯过程对 i, $j \in \{1, \ldots, 22\}$ 的交换对（i, j）进行建模，以解决缺失数据问题（部分现有边中的发送方或接收方信息丢失）。他们的实验表明，对于 Ikenet 数据，核函数（指数衰减或幂律衰减）的选择对拟合优度影响不显著。他们的工作主要关注缺失数据会如何影响参数估计，以及如何利用霍克斯过程填补这些缺失数据。

在我们的实验中，我们仅使用发送者数据，并将时间戳数据转换为计数数据，所用的固定时间间隔为 $\delta t = 0.1$ 小时，因此数据同化包含 78832 个时间步。计数数据时间序列和计数的条形图见图 10。

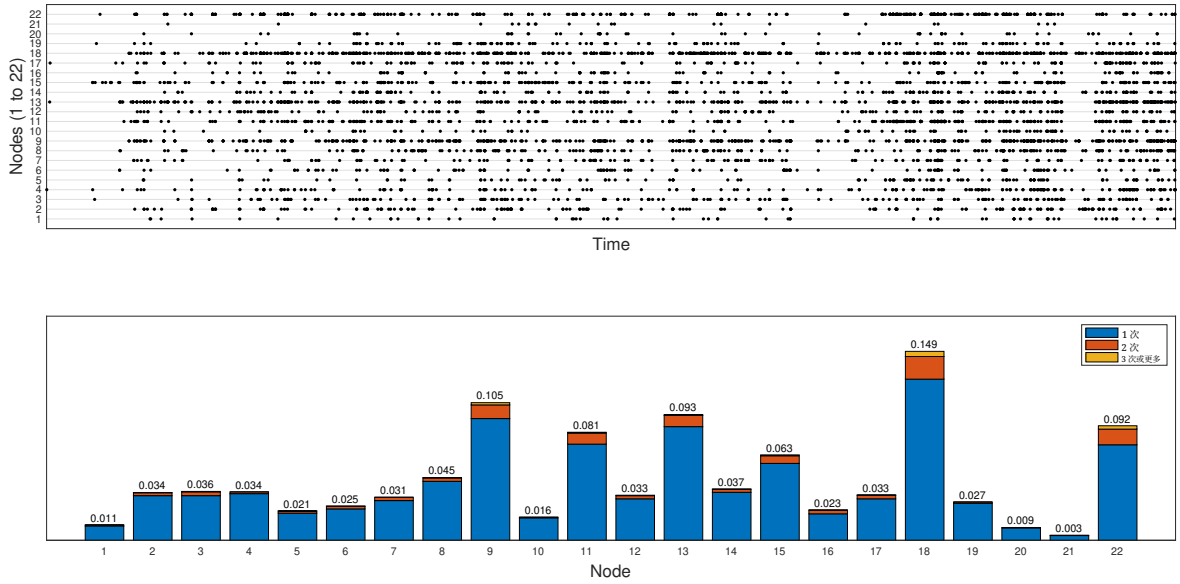


图 10：*IKEnet* 数据的基础探索。（上）展示了时间步0 至 *500* 之间仅针对发出邮件的计数数据时间序列。圆点表示计数非零的事件。（下）条形的高度与各节点发出邮件的百分比成正比，具体数值也标注在条形顶端。每个条形中堆叠部分的高度与图中所示的事件数量成正比（即 *1* 次计数、*2* 次计数和 *3* 次或更多次计数）。

目标是估计（2.1）中的所有参数。参数集合的集合规模为$M = 200$。最终数据同化步骤中的集合用于网络推断。结果将展示哪些志愿者最具影响力，能够促使其他人比“通常”更频繁地发送电子邮件（“通常”由未知的基线率表示）。

现在我们讨论在上述设定下得到的结果。基线率、衰减率和自激发率（即 α_{ii} ）的箱线图总结了集合离散度，见

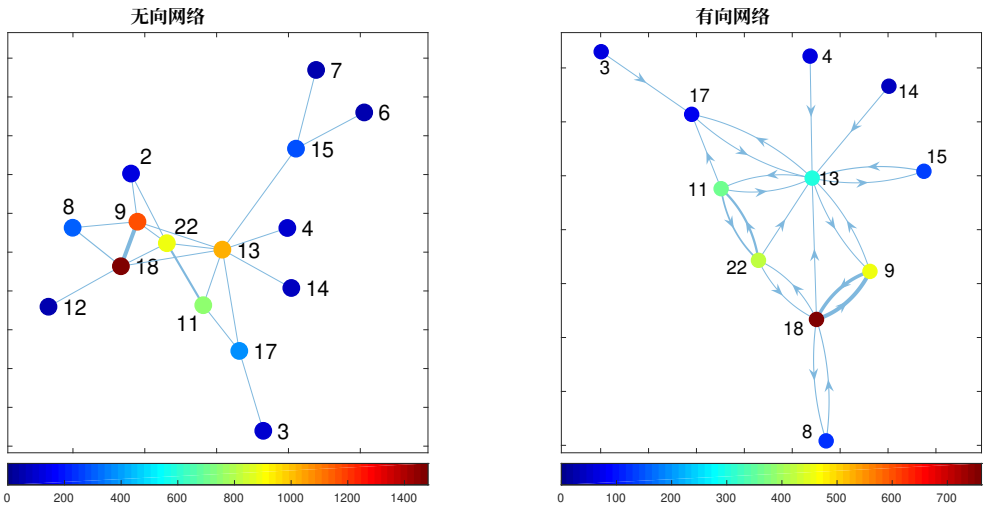


图 11： （左）无向子网络。节点颜色随节点的度而变化，边大小随对应节点对之间电子邮件通信的数量而变化。（右）有向子网络。网络图仅包含权重大于平均权重两倍的边。

图12。节点 18 和 9 的基准速率及平均强度率最高，这与完整数据集的可视化结果一致，见图11。然而，自激发程度最高的节点是节点 22。同样，我们主要关注的是其连边由 α_{ij} （针对） $i \neq j$ 加权的影响网络。为了便于清晰地展示，我们这里只绘制子网络，见图13。通过可视化影响网络，可以获得一些新的见解。例如，尽管节点 18 与节点 9 之间的电子邮件交换量相近，但节点 18 对节点 9 的影响显著大于反方向，参见图11。节点 11 和 22 这一对节点也能观察到类似现象。需要强调的是，这里仅使用发件人信息的计数时间序列。因此，了解影响网络与接收方信息是否一致很有意思。

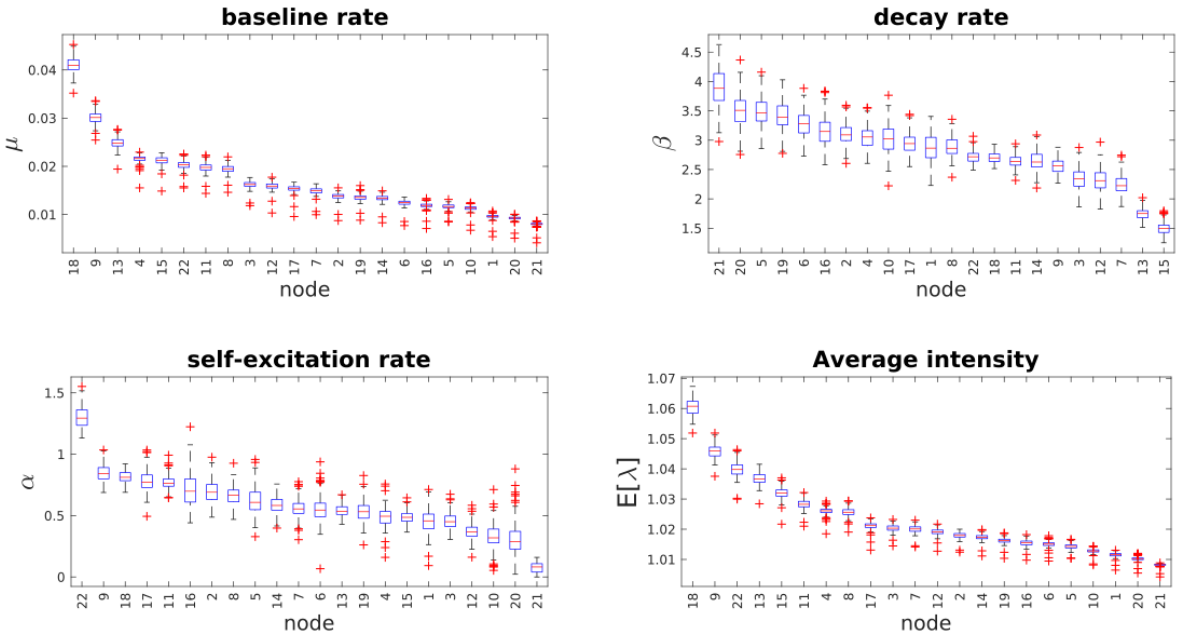


图12： 基线率和衰减率集合的箱线图。

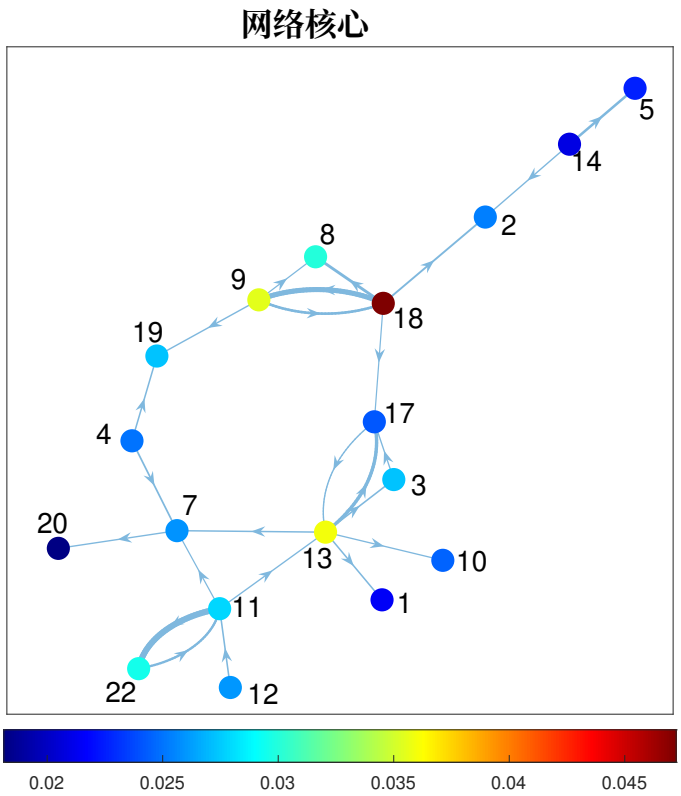


图 13： 移除一些不显著的连边和节点后的（均值）激发网络结构。每个节点根据其自激发率的集合均值着色。

我们还可以通过集合的经验分布来探索网络的不确定性。为此，我们考察了网络各种中心性度量的不确定性。我们计算了从每个集合成员推导出的影响网络的基本中心性度量，包括出度、入度和介数中心性度量，如图14所示。具体而言，对于给定的中心性度量，我们考察集合中 $i-$ 个节点排名第 $j-$ 的比例，即排名的经验分布。结果表明，某些排名相当确定，例如，节点 11 和 18 分别取得了出度中心性的前两名。出度中心性可作为衡量节点整体激发能力的指标，即有影响力的节点发出电子邮件后，其他节点随即发送电子邮件的能力。因此，节点 18 在网络中影响力最大。不过，入度中心性仍有很大的不确定性。尽管如此，节点 21 具有较高的趋势，呈现较大的入度中心性但较低的出度中心性，因此它是影响力最小的节点。发送者网络的介数中心性似乎高度不确定，只有主要由节点 21 占据的末位排名例外。

5.2 欧洲电子邮件数据

我们展示了 EnPGF 处理由欧洲电子邮件数据生成的相对大型网络的能力。欧洲电子邮件数据包含 818 位匿名通信联系人的电子邮件时间戳时间序列，时间跨度为 731 天。我们通过剔除几乎没有电子邮件活动的天数（例如周末、节假日以及数据收集受干扰的时段），以及在 731 天内收发邮件总数少于 40 封的节点，对数据进行“清理”。最终共得到 514 个节点，时间戳数据按一小时为单位进行聚合，以便

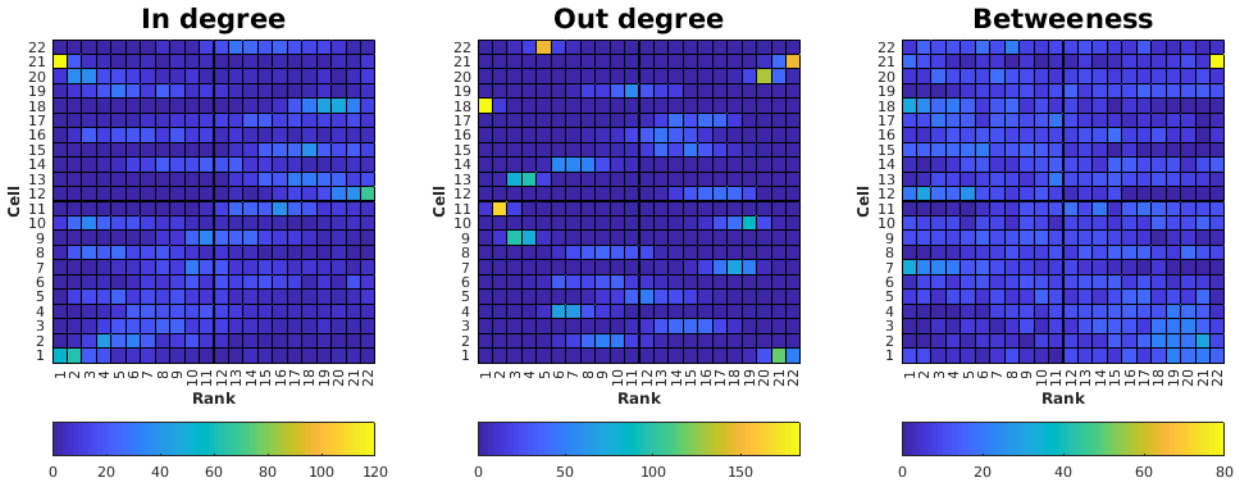


图 14：发送者网络中心性度量的不确定性。色彩图的 (i, j) 元素表示在 *200* 次中，第 j 个节点被赋予第 i 名的次数。

测试。再次，我们只考虑发出邮件；因此，收件人未知。本次分析中，我们使用由 300 个成员组成的集合。图 15 展示了算法的估计结果。显然，发现的网络是稀疏的，因为绝大多数边的权重都不显著（但不为零）。网络核心仅包含权重大于 0.4 的边，清楚地显示了节点 420 与 518 之间存在很强的相互影响；这两个节点恰好是发送电子邮件数量最多的节点。估计结果还检测到节点 419 对节点 122 有显著影响。

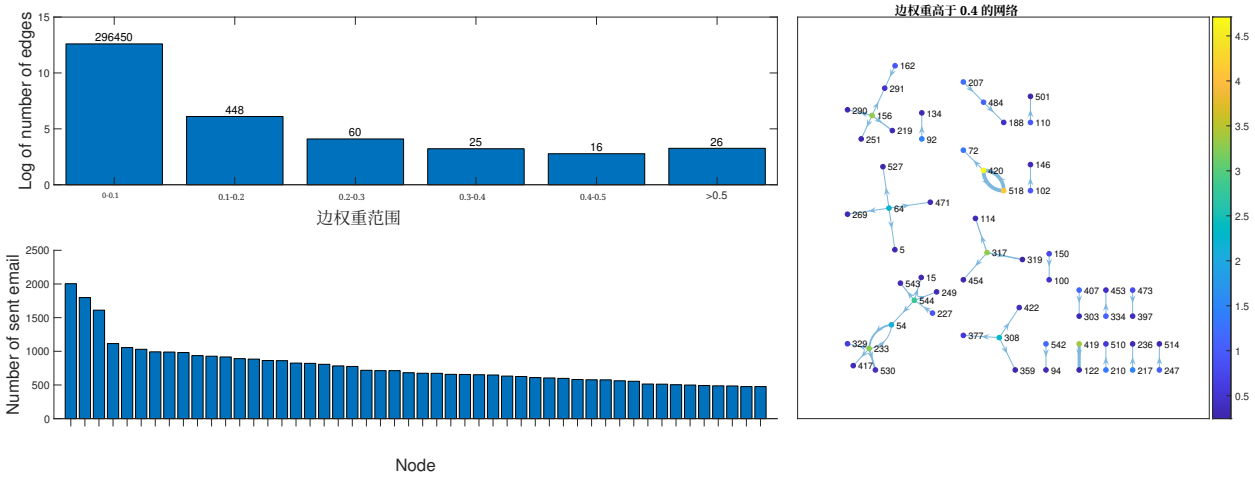


图15：（左上）*EnPGF* 算法得到的边权重直方图，以对数尺度显示。大多数边接近于零，表明该网络是稀疏的。（左下）发送电子邮件数量最多的排名前50的节点。（右）仅包含权重高于 0.4 的边的网络“核心”。节点颜色对应于该节点的总出度（即当前语境下的影响力强度）。

6 结论

我们使用基于集成的框架，利用由计数数据驱动的底层多维霍克斯过程模型来推断影响结构。所提出的方法在完美模型情景和基于 ABM 的模型（即非完美模型）中均展现出良好的效果。我们还仅使用发送信息，将该方法应用于 Ikenet 电子邮件往来数据。图 13 所示结果揭示了霍克斯模型所解释的有趣定向影响结构（2.1）。将这些结果与数据集中完整的信息（即发送方和接收方信息）进行比较后，可以直观地理解其中一些主要影响连边，见图 11。例如，节点 18 对节点 9 的强影响，很可能是因为两者之间有大量电子邮件往来。然而，一些影响连边则较难解释，例如，尽管节点对 (11,22) 和 (11,17) 之间的电子邮件往来次数相近，为什么节点 11 对节点 22 的影响却远强于对节点 17 的影响。本可行性研究不涉及探究霍克斯模型所解释的影响网络是否符合现实世界中的实际情况。我们还指出，所提出的方法可用于分析网络中心性度量的不确定性。在这一应用场景中，结果有助于理解对有影响力的节点进行排名时存在的不确定性。

本文所用方法的主要局限在于推断影响网络所需的数据量（我们预计，所需数据量会随网络规模呈指数级增长），以及多元霍克斯过程能否充分刻画影响。任何采用多元霍克斯过程进行影响网络推断的方法都会面临这些局限。不过，与现有方法相比，本文提出的方法具有若干优势。我们能够处理计数数据，而不仅仅是数据集中常见的时间戳数据；我们不需要了解物理网络；并且得益于并行化，该方法能够对大型网络和数据集进行推断。尽管本文的演示仅限于指数衰减霍克斯过程，但我们的方法论为扩展至多种多维霍克斯模型提供了可能。例如，将本研究扩展到反应性点过程，将过去事件的自激效应与过去事件历史记录或另一类事件历史记录所产生的自抑制效应结合起来，将会很有意义。在犯罪建模中，犯罪强度可视为受到以往犯罪历史的自激作用，并受到警察巡逻的自抑制作用（见 [12] 对抗性巡逻）；在电网故障背景下，故障强度会受到网络先前故障的自激作用，并受到检查或维护的自抑制作用（电网维护 [7]，水管维护 [26]）。然而，在一般情形下，这些霍克斯模型的模型参数与强度动态之间若存在强非线性，可能会给（基于集合的）回归步骤带来问题。此外，尽管我们假设节点之间的参数相互独立，但这一限制可以取消。不过，这会妨碍算法的并行化，从而降低该方法用于大规模推断的可行性。尽管如此，或许可以采用 EnKF 中广泛使用的局部化技术，以提高计算效率并避免参数的虚假相关。在地球物理背景下，虚假相关往往表现为相距较远位置的参数之间存在长程相关。

代码可用性

本文用于生成结果的代码可在以下位置获取

https://github.com/naratips/EnPGF

数据可用性声明

本文所用的所有数据均可在提出合理请求后提供。

致谢

NS 衷心感谢英国工程与物理科学研究委员会对项目资助 grantEP/P030882/1 的支持。

附录

我们基于泊松-伽马共轭性构造更新规则（3.1）。假设先验分布服从伽马分布，其尺度参数和形状参数分别为 a 和 b。给定一个 δN，它是从均值为 $\lambda\delta t$的泊松分布中抽取的观测值，则 λ 的后验分布将具有尺度参数和形状参数 $a+\delta N$ 和 $b+\delta t$。然而，由于我们需要根据集合均值和均值偏差来更新集合，因此不再使用尺度参数和形状参数。为此，我们将伽马分布改用均值和相对方差进行参数化。回顾一下，尺度参数为 a 且形状参数为 b 的伽马分布，其均值为 a/b，方差为 a/b^2。因此，我们可以写成

$$\begin{aligned} E[\lambda \mid \Delta N] &= \frac{a+\delta N}{b+\delta t} = \frac{a}{b} + \frac{(a/b)}{a+(a/b)\delta t}\left(\Delta N - (a/b)\delta t\right) \\ &= E[\lambda] + \frac{E[\lambda]}{\frac{Var[\lambda]}{E[\lambda]^2} + E[\lambda]\delta t}\left(\Delta N - E[\lambda]\delta t\right). \end{aligned} \tag{6.1}$$

将 $E[\lambda]$ 替换为集合均值，并将$Var[\lambda]$ 替换为集合方差，即可得到（3.1）中的均值更新方程。（3.1）中相对方差的更新规则也可以类似地验证。如（3.3）所示，更新均值后，我们需要计算所有集合成员相对于均值的偏差，以满足（3.1）中的相对方差更新方程。现在我们将证明，可以使用（3.2）的右侧随机更新每个集合成员，从而使更新后的集合满足（3.1）中的相对方差。

为验证（3.2），我们首先将（3.2）中的每一项简写为

$$w^{(s)} := \frac{\lambda^{(s),a} - \bar{\lambda^a}}{\bar{\lambda^a}}, \quad u^{(s)} := \frac{\lambda^{(s)} - \bar{\lambda}}{\bar{\lambda}}, \quad t^{(s)} := \frac{\Delta N_e^{(s)} - \Delta\bar{N}_e}{\Delta\bar{N}_e}, \quad c := \frac{P_r}{P_r + (\Delta N)^{-1}}. \tag{6.2}$$

我们作如下近似：

$$\begin{aligned}\frac{Var[\lambda^a \mid \Delta N]}{E[\lambda^a \mid \Delta N]^2} &\approx \frac{\frac{1}{M-1}\sum_{s=1}^{M}(\lambda^{(s),a}-\bar{\lambda^a})^2}{\bar{\lambda^a}^2} = \frac{1}{M-1}\sum_{s=1}^{M} w^{(s)} =: P_r^a \\ \frac{Var[\lambda]}{E[\lambda]^2} &\approx \frac{\frac{1}{M-1}\sum_{s=1}^{M}(\lambda^{(s)}-\bar{\lambda})^2}{\bar{\lambda}^2} = \frac{1}{M-1}\sum_{s=1}^{M} u^{(s)} =: P_r \\ \frac{Var[\Delta N_e \mid \Delta N]}{E[\Delta N_e \mid \Delta N]^2} = \Delta N^{-1} &\approx \frac{\frac{1}{M-1}\sum_{s=1}^{M}(\Delta N_e^{(s)}-\Delta\bar{N}_e)^2}{\Delta\bar{N}_e{}^2} = \frac{1}{M-1}\sum_{s=1}^{M} t^{(s)}.\end{aligned} \tag{6.3}$$

我们可以假设，给定 ΔN 的观测值后，扰动观测 ΔN_e 在条件上与 λ 独立。因此，基于大数弱定律，我们作出如下近似

$$\frac{1}{M}\sum_{s=1}^{M} u^{(s)}t^{(s)} \approx E[ut] = E[u]E[t] \approx \frac{1}{M}\sum_{s=1}^{M} u^{(s)}\frac{1}{M}\sum_{s=1}^{M} t^{(s)} = 0. \tag{6.4}$$

将（3.2）两边平方，然后求和并除以 $M-1$。将上述所有近似应用于所得结果，可得

$$\begin{aligned}P_r^a &= E\left[\left(\frac{\lambda^{(s),a}-\bar{\lambda^a}}{\bar{\lambda^a}}\right)^2\right] = P_r - 2\frac{P_r^2}{P_r+(\Delta N)^{-1}} + \frac{P_r^2}{P_r+(\Delta N)^{-1}}, \\ &= P_r - \frac{P_r^2}{P_r+(\Delta N)^{-1}}.\end{aligned} \tag{6.5}$$

参考文献

[1] W. Chen、W. Lu 和 N. Zhang。在社交网络中基于时滞扩散过程的时间关键型影响力最大化。载于*AAAI*人工智能会议论文集，第26卷，第591–598页，2021年。

[2] W. Chen、Y. Yuan 和 L. Zhang。线性阈值模型下社交网络中可扩展影响力最大化。载于*2013 IEEE*第*13*届国际数据挖掘会议，第88–97页，2010年。

[3] D. R. Cox 和 P. A. Lewis。事件序列的统计分析。Methuen，伦敦，1966年。

[4] A. De、I. Valera、N. Ganguly、S. Bhattacharya 和 M. Gomez-Rodriguez。学习并预测社交网络中的观点动态。载于神经信息处理系统进展 *29*：*2016*年神经信息处理系统年会，*2016*年*12*月*5*日至*10*日，西班牙巴塞罗那，第 397–405 页，2016年。

[5] Pedro Domingos 和 Matt Richardson。挖掘客户的网络价值。载于*KDD '01*：第七届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第 57–66 页，纽约，美国纽约州，2001年。ACM出版社。

[6] A. Doucet 和 A. M. Johansen。粒子滤波与平滑教程：十五年后。载于 D. Crisan 和 B. Rozovsky 编，《牛津非线性滤波手册》。牛津大学出版社，2011年。[7] S¸eyda Ertekin、Cynthia Rudin 和 Tyler H. McCormick。反应性点过程：一种预测地下电力系统电力故障的新方法。《应用统计年鉴》，9(1)：122–144，2015年3月。[8] PabloA . Estevez、Pablo A. Vera 和 Kazumi Saito。遴选社交网络中最具影响力的节点。*2007*年国际神经网络联合会议，第 2397–2402 页，2007年。[9] G. Evensen。《数据同化：集合卡尔曼滤波》。施普林格出版社，2006年。[10] E. W. Fox、M. B. Short、F. P. Schoenberg、K. D. Coronges 和 A. L. Bertozzi。利用自激点过程对电子邮件网络建模并推断领导力。《美国统计学会杂志》，111(514)：564–584，2016年。[11] A. G. Hawkes 和 D. Oakes。自激过程的簇过程表示。《应用概率杂志》，11(3)：493–503，1974年。[12] Bn Hefferan、Oliver M. Cliff 和 Robert Fitch。利用反应性点过程进行对抗性巡逻。载于澳大拉西亚机器人与自动化会议论文集，澳大利亚布里斯班，2016年。[13] S. Huang、A. W.C. Fu、和 R. Liu。时间图中的最小生成树。载于*2015*年 *ACM SIGMOD*数据管理国际会议，第 4 号，第 419–430 页，2015年。[14] T. Iwata、A. Shah 和 Z. Ghahramani。通过共享级联点过程发现在线社交活动中的潜在影响。载于第 *19* 届 *ACM SIGKDD*知识发现与数据挖掘国际会议，第 266–274 页，2013年。[15] David Kempe、Jon Klienberg 和 ´Eva Tardos。最大化影响力在社交网络中的传播。载于*KDD ’03*：第九届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第 137–146 页。ACM出版社，2003年。[16] Yun-Yong Ko、Kyung-Jae Cho 和 Sang-Wook Kim。社交网络中高效且有效的影响力最大化：一种混合方法。《信息科学》，465：144–161，2018年。[17] S. W. Linderman 和 R. P. Adams。从点过程数据中发现潜在网络结构。载于国际机器学习会议，第 1413–1421 页，2014年。[18] G. Miritello、E. Moro 和 R. Lara。信息传播中社会关系的动态强度。《物理评论*E*》，83(045102(R))，2011年。[19] M. B. Mohler、G. O.and Short、P. J. Brantingham、F. P. Schoenberg 和 G. E. Tita。犯罪的自激点过程建模。《美国统计学会杂志》，106(493)：100–108，2011年。[20] S. Morse、M.C. Gonzalez 和 N. Markuzon。持久级联在扩散中的作用。《物理评论*E*》，99(012323)，2019年。

[21] T. Parmer、L. M. Rocha 和 F. Radicchi。布尔网络中的影响力最大化。《自然·通讯》，13(3457)，2022年。[22] N. Santitissadeekorn、D. J. B. Lloyd、M. B. Short 和 S. Delahaies。泊松计数数据条件强度过程的近似滤波：城市犯罪应用。《计算统计与数据分析》，144：106850，2020年。[23] N. Santitissadeekorn、M. B. Short 和 D. J. B. Lloyd。用于一维自激过程的序贯数据同化及其在城市犯罪数据中的应用。《计算统计与数据分析》，128：163–183，2018年。[24] M.B. Short、M.R. D’Orsogna、V. B. Pasour、G. E. Tita、P. Jeffrey Brantingham、A. L. Bertozzi 和 L. B. Chayes。犯罪行为的统计模型。《应用科学中的数学模型与方法》，18：1249–1267，2008年。[25] B. Yuan、H. Li、A. L. Bertozzi、P. J. Brantingham 和 M. A. Potter。多变量时空霍克斯过程与网络重构。《*SIAM*数学数据科学杂志》，1(2)：356–382，2019年。[26] C. Zhang、H. Wu、R. Bie、R. Mehmood 和 A. Kos。预防性管道维护中故障事件的动态建模。《*IEEE Access*》，6：12539–12550，2018年。[27] Y. Zhao 等。使用 Klipi 多维霍克斯模型挖掘医疗记录。发表于*KDD*健康信息学研讨会，2014年。[28] J. Zhuang、Y. Ogata 和 D. Vere-Jones。利用随机重建分析地震聚集特征。《地球物理研究杂志：固体地球》，109(B5)：1–17，2004年。[29] J. R. Zipkin、F. P. Coronges 和 A. L. Bertozzi。社交网络交互的点过程模型：参数估计与缺失数据恢复。《欧洲应用数学杂志》，27(3)：502–529，2016年。