**Influence Estimation and Maximization in Continuous-Time Diffusion Network**

**Abstract**

本文提出了连续时间扩散网络中影响估计的两种算法。第一个使用连续时间马尔可夫链来估计影响完全在网络指数,或者,更普遍的是,phase-type传输功能,但不扩展到大规模网络,和第二个是一种高效的随机算法,估计网络中每个节点的影响与一般的传输函数。、

**1.introduction**

**影响估计问题**的目的是估计一组节点在一个时间段内可以影响的节点的期望数量，

**影响最大化问题**的目标是找到一组节点，这些节点在一段时间内，可以影响最多的节点

这两个问题都是具有挑战性的，因为我们首先需要精确地建模级联数据中的时间信息，其次需要设计高效的算法，可以扩展到具有数百万节点的网络。然而，在大多数情况下，必须估计或在给定的时间内使影响最大化，也就是说，必须考虑有限的时间窗。

作者认为选择连续时间的传播模型可以得到更好的效果。这种建模选择背后有两个基本原理。首先，由于后续操作是异步进行的，所以连续变量似乎更适合表示它们。其次，离散时间模型可能过于局限于捕捉数据中丰富的时间动态。

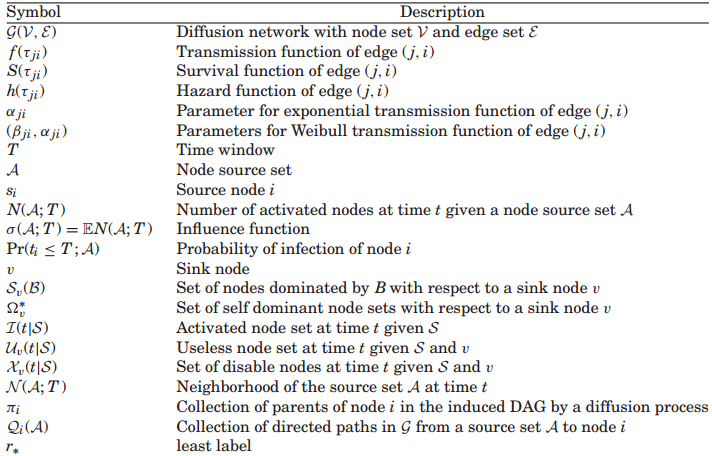
然而，基于连续时间扩散模型的影响估计和最大化也面临着巨大的挑战。首先，该设置中的影响估计问题是一个复杂的图形模型推理问题，即在循环图形模型中计算连续变量的边际密度，其次，目前还不清楚如何将基于连续时间扩散模型的影响估计和最大化算法扩展到数百万个节点。

**1.1 related work**

**1.2 作者的方法**

作者建立了一个连续时间的扩散模型，可以应用于带时间的异构网络中。

符号标记



**2.传播模型**

**2.1 连续时间的传播过程**

我们考虑发生在静态网络上的扩散和传播过程。对于每条边j和i，用一个传输函数连起来，这个数值是随时间变化的。当种子节点集A在t = 0时受到网络外部源的作用而激活时，开始扩散过程，然后，种子节点试图感染它们的邻居节点，当邻居节点i在时间ti被激活时，他又会去试图激活它的邻居节点。每一次通过边缘的传输都需要一个随机传输时间，它是从传输函数中得到的。我们认为传输时间时互相独立的。对于有些传输函数，有可能会有为无穷大，这样的话他永远无法激活邻居节点。在这里，我们假设一个渐进模型，其中一个节点i在其第一个父节点要感染它时即被感染，之后的激活都不会有作用。在这个假设下，在一个时刻t，对于一个特定节点v，网络中会存在着一些节点和边对于激活v是无用的。这里有一个性质：所有边上都给定一个传输时间。当节点i在ti时被激活，ti是G中从源到节点i的最短路径的长度，其中边权值对应于相关的传输时间。作者认为连续时间模型是IC模型的升级。

**2.2 成对传输的可能性**

我们用函数表示一条有向边j到i中结点i被节点j在时间ti被激活的条件概率。

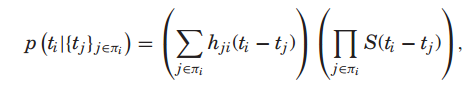
。

在以往的工作中，大多采用简单的参数传递函数，如指数分布，瑞利分布。但是在现实世界中，在节点之间的信息传递可以是异构的，并且等待时间可能不同于上述的分布。例如，在病毒式营销中，**活跃**的消费者可以立即更新他们的状态，而不活跃的用户可能只是登录并每天回复一次，因此，活跃用户和他的朋友之间的传输函数可能与不活跃用户和他的朋友之间的传输函数非常不同。此外，根据扩散信息的内容，扩散过程也可能是不同的。作为对这些复杂场景建模的一种尝试，非参数和主题相关的传输函数最近得到了考虑[Du et al. 2012, 2013; Barbieri et al. 2013]，在这种情况下定义了survival function和hazard functioin。

其中survival function被定义为：，hazard function被定义为：。有如下两个结论：

**2.3 A Graphical Model Perspective of Continuous-Time Diffusion Processes**

在本节中，我们将从图形模型的角度研究连续时间独立级联模型，其中，随机变量集合包括节点感染时间ti。我们假定图G是一个有向无环的图。我们可以用一个有向的图形模型来模拟感染时间ti的密度：其中每个πi表示节点i的父母的集合，代表了节点i被其父母节点激活的时间ti的条件密度。这是因为给定节点i的父节点的感染时间，ti与其他感染时间无关，满足有向图模型的局部马尔可夫性，我们注意到，独立级联模型只明确指定了每条有向边的成对传输函数，而没有直接定义条件密度。然而，这些条件密度可以由基于独立感染特性的成对传播函数得到



这是节点i被每个父节点j感染的概率之和

**3. 影响估计**

**4.影响力最大化**

**4.1 连续时间的影响力最大化问题**

给一个传播网络G=（V,E），一个传输函数，目标是找到一个种子集合A，使得影响函数最大。用数学公式表达为：

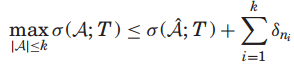
**4.2 IM问题的难点和结构**

定理4.1：给定一个网络G，一个种子集合的数量K，一个时间限制T，连续时间点的影响力最大化问题是NP-难的。用集合覆盖问题进行证明。

定理4.2：给定一个网络G，一个种子集合的数量K，一个时间限制T，影响函数是单调的和子模的。

**4.3 有效的影响力最大化算法**

作者通过执行延迟求值来加速贪婪影响最大化算法，可以显著降低计算边际收益时迭代的数量。作者还使用影响函数的子模性来获得一个限制

定理4.3：对于一个由k个节点的种子集合A和一个不在种子集合中的节点n，作者用表示边际收益，并且将每个节点的边际收益按照降序排列。

**5. 实验结果**

**5.2 在真实数据中的实验**

5.2.1 实验建立：作者使用的是MemeTracker数据集，包含超过1.72亿篇来自100万个主流媒体网站和博客的新闻文章和博客文章。作者从排名前600位的媒体网站和博客中抽取了10967个超链接级联，并将所有级联随机分为80%的训练集和20%的测试集，共5次。

影响力最大化问题的表现：

