Network Reconstruction through diffusive arrival times

***Abstract*— Network reconstruction problem is one of the hot and knotty issues in the research of complex network or network science. In this paper, we use kernel density estimation technique to estimate the distribution of the time difference of arrival in such diffusion process on the basis of stochastic temporal network. We analyze the statistical property discrepancies between edges in the network, then give proof on the left deviation of the estimated survival function on the time-aggregated network. Next we design a probability threshold cutting algorithm which can be used to reconstruct the time-aggregated network of stochastic temporal network. To verify this, we run a lot of simulations on different networks which show high reconstruction speed and accuracy of our algorithm. Last, we discuss the relation between network scale and data amount of the reconstruction procedure which illustrates the compatibility with such large scale network reconstruction problem. Furthermore, a parallelization design idea is presented to speed up the algorithm.**

***Index Terms*— network reconstruction, stochastic temporal network, time-aggregated network, waiting time distribution, kernel density estimation.**

I. Introduction

复杂网络，是一种描述现实世界复杂系统的重要工具，在生物、信息、交通等各个领域都有重要的研究意义和实用价值。随着目前信息时代和数据时代的发展，网络科学越来越受到人们的关注。总体来看，网络科学是专门研究复杂网络系统的定性和定量规律的一门交叉科学 ,研究涉及到复杂网络的各种拓扑结构及其性质，比如随机图网络、无标度网络、小世界网络等著名随机网络的研究。与动力学特性 (或功能) 之间相互关系 ,包括动力学同步及其产生机制、网络上的病毒传播及免疫、链路预测、网络可控性研究、网络演化博弈等，以及工程实际所需的网络设计原理及其应用研究，其交叉研究内容十分广泛而丰富。

网络重构问题也是当今网络科学研究中的热点问题和难点问题。在很多情况下，网络的拓扑结构并不能直接被我们观察或测量出来。网络重构就是考虑这样的一种逆问题：通过一些可以观测到的网络动力学行为的数据，来逆向恢复出原始网络的拓扑结构甚至相关性质，也就是所谓的网络重构。我们特别关注了北京师范大学王文旭课题组和复旦大学李翔课题组近年来相关的研究成果。王文旭课题组在近年基于压缩感知[1]技术，在网络重构问题上取得了一系列的成果。首先，他们从进化博弈的角度，利用少量的博弈数据，用压缩感知的方法将网络重构问题转化为稀疏信号的重构问题，而且能在有一定噪声数据的情况下达到很高的重构精度[2]。随后，又从SIS和CP这两种网络传播模型出发，将极度非平凡的传播网络重构问题转化为压缩感知技术框架下的问题，实现了基于二进制时序数据的网络重构[3]。不久后，他们又从个体结点出发，利用Lasso方法，同样将重构问题转化为稀疏信号的恢复问题，将每个结点和其他所有结点之间的连接视为一个稀疏信号的重构，最后整合所有结点的领域信息，从而重构整个网络[4]。Lasso技术中，惩罚项保证了重构的鲁棒性，L1-norm则保证了信号的稀疏性，也就是只需要较少的观测数据。上述基于压缩感知的网络重构都具有一定的鲁棒性。复旦大学李翔课题组近期也在重构时效网络研究取得了一定的突破，考虑了非泊松条件下的传播过程，利用传播过程的到达时间数据实现了重构随机时效网络的有效推断[5]。

网络重构问题虽然取得了一定的发展，但仍然面临着高难度的挑战。如何从不同种类数据中精确恢复完整的网络信息，甚至包括连边方向、权重等，仍然亟待研究。本文研究的内容主要是：以随机时效网络模型为框架，利用扩散过程中的首达时间数据进行随机时效网络模型中的时间累积网络的重构。这一研究的出发点是，在某些情况下，我们并不关心具体的网络交互细节，反而对网络的统计特征更感兴趣，而随机时效网络既包含了底层网络拓扑信息——时间累积网络，又含有点对层面上的时间特性——等待时间分布。相对于传统的基于静态网络模型的重构方法，这种方法更适用于从含有时间特性的数据中恢复出网络拓扑关系。更具体地说，我们的算法只需要知道在扩散过程中，节

点的首达时间信息，就可以重构出网络节点之间的连边关系。通过这篇论文的研究，我们可以清晰地刻画随机时效网络模型和其上发生的扩散过程，以及如何利用扩散过程数据的特殊性质进行时间累积网络的重构，阈值剪枝操作和扩散过程数据量又是怎样影响重构精度的。

II. Diffusion process on stochastic temporal network

本章介绍随机时效网络模型，以及其上发生的扩散过程是如何定义及模拟的。

A. Derivation of Stochastic Temporal Network

为了引入随机时效网络，首先我们介绍一般的时效网络$\mathcal{N}=(\mathcal{V},\mathcal{E})$。其中$\mathcal{V}$是时效网络节点的集合，$\mathcal{E}$是节点之间的时效交互事件$\epsilon=(u,v,t,\delta t)$的集合，$(u,v)$表示事件发生的相关节点对，$t\in\left[0,T^W\right]$表示事件发生的时间，$T^W$表示观测窗口的长度，$\delta t$表示事件的持续时间，也就表明该事件发生的时间段是$(t,t+\delta t)$。我们假设事件的发生间隔$\delta t$趋近于0。因此，在同一个时间点不可能有两个交互事件同时发生。从而我们把随机时效网络中的事件定义为(u,v,t)，抹去了事件的持续时间这一特征，认为时间的发生是在极短的时间内完成的。这个假设主要对应的是信息扩散过程或病毒传播过程。

接着我们构造随机时效网络中的时间累积网络。首先通过映射：$P_{V^2}:V^2\times\left[0,T^W\right]\mapsto V^2,(u,v,t)\mapsto(u,v)$将所有的时效连边投影成为静态的时间累积网络连边，从而构造出了一个时间累积网络：

$$\mathcal{G}=P_{V^2}(E)=\{(u,v)|(u,v,t)\in E\}$$

其次，对时间累积网络中的任意一条连边$(u,v)\in\mathcal{G}$，假设其发生事件的事件间隔是服从一个实证分布$\psi(t)$的。具体做法是，记$\{(u,v,t_{uv}^i)\}_{i=1,2,\ldots,M}=P_{V^2}^{-1}[(u,v)]$为连边$(u,v)\in\mathcal{G}$上所有记录下来的交互事件，按照事件的升序排列，$t_{uv}^1<t_{uv}^2<\ldots<t_{uv}^M$。则$\psi_{uv}(t)$是事件间隔时间$\{\Delta t_{uv}^i=t_{uv}^i-t_{uv}^{i-1}\}$的实证分布。在数据量有限的情况下，往往可以采用核密度估计的方法来估计出相应的事件间隔分布：

$$\psi_{uv}(t)=\frac{1}{M}\sum_{i=1}^{M}K_h(t-\Delta t_{uv}^i)$$

式中，M表示数据个数，$K_h(\cdot)$表示带宽为h的核函数。下面我们根据事件间隔分布$\psi_{uv}(t)$引入连边上的等待时间分布$\rho_{uv}(\tau)$。首先，等待时间τ_{uv}定义为在扩散过程中，从节点u被首次通知到(u,v)上首次出现时效边使得v被u通知的中继时间。假设不同连边上的等待时间分布是相互独立的，则等待时间τ_{uv}服从一个依据时间间隔分布$\psi_{uv}(t)$进行长度偏差采样到的概率密度分布，写作

$$\rho_{uv}(\tau)=\frac{1}{m_{uv}}\int_{\tau}^{\infty}\psi_{uv}(t)dt\Theta(\tau)$$

其中$m_{uv}=\int_0^\infty t\psi_{uv}(t)dt$表示事件间隔分布$\rho_{uv}(\tau)$的均值，$\Theta(\tau)$表示单位阶跃函数。由此我们可以得到所有时间累积网络上的等待时间分布。从上面的步骤中，我们总结推导出了随机时效网络模型下的时间累积网络$\mathcal{G}$和等待时间分布$\rho_{uv}(\tau)$，则随机时效网络可以表示为

$$\mathcal{N}_{\mathcal{S}}=(\mathcal{G},\boldsymbol{\rho}),\boldsymbol{\rho}=\{\rho_{uv}(\tau)\}_{(u,v)\in\mathcal{G}}$$

接着我们使用生存分析的方法来描述随机时效网络上的扩散过程。首先，给定一个任意的随机时效网络 $\mathcal{N}_{\mathcal{S}}=(\mathcal{G},\boldsymbol{\rho})$。我们注意其等待时间分布$\rho_{uv}(\tau)$，规定$\rho_{uv}(\tau)=0,\tau<0$。因为对于扩散过程而言，不存在等待时间$\tau_{uv}<0$的情况。这样一来，就可以写出等待时间分布的规范条件

$$\int_0^\infty\rho_{uv}(\tau)d\tau=1$$

在生存分析中，还有一个概念是生存函数(Survival Function)，定义为

$$\Phi_{uv}(\tau)=1-F_{uv}(\tau)=1-\int_0^\tau\rho_{uv}(\tau)d\tau=\int_\tau^\infty\rho_{uv}(\tau)d\tau$$

$F_{uv}(\tau)$表示等待时间分布的概率密度函数，$\Phi_{uv}(\tau)$表示连边(u,v)在τ时刻之前没有被激活的概率。很显然，对于任意不属于时间累积网络的节点对$(u,v)\notin\mathcal{G}$，$\rho_{uv}(\tau)\equiv0$。需要特别说明的是，为了问题的规范，我们使用一个支撑集$[0,\tau_{max})$来限定等待时间分布。防止由于支撑区间过大，使得难以对其进行核密度估计。

B. Diffusion Process on Stochastic temporal network

下面介绍并模拟随机时效网络$\mathcal{N}_{\mathcal{S}}=(\mathcal{G},\boldsymbol{\rho})$上发生的扩散过程。简明起见，我们规定时间累积网络$\mathcal{G}$是无向的，即$(u,v)\in\mathcal{G}$和$(v,u)\in\mathcal{G}$的扩散规律是相同的。下面考虑扩散过程的发生，算法1给出了该过程的实现。 通过该算法，我们可以得到一个随机时效网络的

Algorithm 1: Generation of Diffusive Arrival Times

Input: $\mathcal{N}_{\mathcal{S}} = (\mathcal{G}, \boldsymbol{\rho})$, s^*

Output: $\mathcal{D} = \{t_v\}_{v\in\mathcal{G}}$

$\{w_{uv}\}_{(u,v)\in\mathcal{G}} \leftarrow 0$;

for *each* (u,v) *in* $\mathcal{G}$ **do**

  $w_{uv} \leftarrow random\_sampling(\rho_{uv}(\tau))$;

$\mathcal{D} \leftarrow dijkstra(\{w_{uv}\}, s^*)$;

return $\mathcal{D}$;

一次扩散数据$\mathcal{D} = \{t_v\}_{v\in\mathcal{G}}$，如果想得到多个扩散数据，可以从网络中随机选取节点作为扩散源，仿真扩散过程，就可以得到一组扩散过程数据$D = \{\mathcal{D}^i\}_{i=1,2,...,M}$，其中第$i$次的扩散数据表示为$\mathcal{D}^i = \{t_v^i\}_{v\in\mathcal{G}}$。$M$代表数据的总组数。

我们考虑在一个固定的随机时效网络上发生的一次信息扩散过程。随机时效网络的时间累积网络无向无自环，等待时间分布服从均质化假设。下面绘制了相应的网络图和扩散图。

III. CHARACTERISTIC DIFFERNECE ANALYSIS OF ESTIMATED DISTRIBUTION

A. Defination of Edges on Diffusion Process

在分析之前，我们特别说明文中提到的几个概念的区别：网络边、非网络边，扩散边和非扩散边。对于随机时效网络 $\mathcal{N}_{\mathcal{S}} = (\mathcal{G}, \boldsymbol{\rho})$，某一次扩散过程$\mathcal{D}$的扩散路径形成的扩散树为$\mathcal{T}$，如果$(u,v) \in \mathcal{G}$则称$(u,v)$是网络边，如果$(u,v) \notin \mathcal{G}$则称$(u,v)$是非网络边，如果$(u,v) \in \mathcal{T}$则称$(u,v)$是扩散边，如果$(u,v) \in \mathcal{G}$但$(u,v) \notin \mathcal{T}$则称$(u,v)$是网络边则称$(u,v)$是非扩散边。

接着，我们使用$d_{uv} = t_v - t_u$表示由首达时间数据得到节点对的到达时间差。对于扩散边，其首达时间差d_{uv}是依据相应的等待时间分布$\rho_{uv}(\tau)$采样得到的，在统计上等于相应的等待时间$d_{uv} = \tau_{uv}$ ；而对于非扩散边或非网络边,其首达时间差d_{uv}则是根据网络的拓扑结构和周围节点被通知情况而决定的，不服从等待时间分布ρ_{uv}。但需要我们注意的是，扩散过程中会普遍存在下述的情况。假设我们知道有这么一条连边$(u^*, v^*) \in \mathcal{G}$存在，当发生了多次信息扩散过程$D = \{\mathcal{D}^i\}_{i=1,2,...,M}$时，连边$(u^*, v^*)$受邻居节点的影响，不一定在任意的扩散过程$\mathcal{D}^i$中都是扩散边。总之，扩散边和非扩散边是针对于一次扩散过程$\mathcal{D}$而言的。在某一次扩散过程$\mathcal{D}^i$中，不是所有的网络边都会成为扩散边。在这一次传播过程$\mathcal{D}^i$的扩散树$\mathcal{T}^i$中出现的边是扩散边，而且一定属于网络边，因为只有网络连边存在，扩散过程才可能从这条连边上经过，从而成为扩散边；不属于扩散树$\mathcal{T}^i$，但属于网络边的边是非扩散边。因此，在一次扩散中，网络边既可能是扩散边，也可能是非扩散边,此外，非网络边不属于时间累积网络，在任意的扩散过程中都一定不会出现。所以，对于假设已知存在的连边$(u^*, v^*) \in \mathcal{G}$作为非扩散边出现在扩散过程中时，会干扰到我们的重构过程。因为作为非扩散边出现的情况下，连边上两个节点的首达时间差$d_{u^*v^*}$同样不服从等待时间分布$\rho_{u^*v^*}(\tau)$的采样,会影响其估计分布的精确度，降低我们认为该边是网络连边的可能性。

B. Distribution Estimation on Edges through Time Differences of Arrival

随后，我们用核密度估计的方法进行节点对首达时间差的分布估计。对于上述扩散过程D，我们考虑某一连边(u,v)上的首达时间差数据$\boldsymbol{d}_{uv} = \{d_{uv}^i\}_{i=1,2,...,M}$的概率分布进行估计，其概率密度函数估计为：

$$\hat{\rho}_{uv}(\tau) = \frac{1}{M}\sum_{i=1}^{M} K_h(\tau - d_{uv}^i)$$

$\hat{\rho}_{uv}(\tau)$表示估计出的概率密度分布。最常用的一种核函数是高斯核函数$K(\tau) = \frac{1}{\sqrt{2\pi}} exp(-\tau^2/2)$ 高斯核函数的形状和平滑核带宽将会决定平滑的效果。在我们后续实验中，重构算法并不严重依赖核函数和核带宽的选取，因此核带宽取一个很小的值如0.01，保证数据不会被过渡平滑。

利用KDE方法，我们发现网络边节点对和非网络边节点对的首达时间差估计分布是具有明显差异的。

1) Left Skew of Estimated Survival Function on True Edges:

Proposition 1: $\forall(u,v) \in \mathcal{G}, \tau \in [0, \tau_{max})$, if $M \to \infty, h \to 0, M \cdot h \to 0$, we have

$$\hat{\Phi}_{uv}(\tau) \leq \Phi_{uv}(\tau) \tag{1}$$

Proof: We consider M diffusion processes took on a STN model. The diffusive arrival times(DATs) we obtained from the M diffusive simulations are $D = \{\mathcal{D}^i\}_{i=1,2,...,M}, \mathcal{D}^i = \{t_v^i\}_{v\in\mathcal{G}}$. For each $\mathcal{D}^i \in D$,$\mathcal{T}^i$ is the diffusion path or diffusion tree. Then we consider the time differences of arrival times(TDOAs) on a

certain edge$(u,v) \in \mathcal{G}$, which takes the form $\boldsymbol{d}_{uv} = \{d_{uv}^i\}_{i=1,2,\ldots,M}$. Without loss of generality, we assume that the time-aggregated network is undirected, which means diffusions on $(u,v) \in \mathcal{G}$ and $(v,u) \in \mathcal{G}$ obey the same WTD, $\rho_{uv}(\tau) = \rho_{vu}(\tau)$. In this way $d_{uv}^i = |t_u - t_v|$ for any i. If it is a directed network, considering individually the two directions will get the same result.

To facilitate the proof, we devide $\boldsymbol{d}_{uv}$ into two parts ${}^1\boldsymbol{d}_{uv}$, ${}^2\boldsymbol{d}_{uv}$. In other words, For any $d_{uv}^i \in \boldsymbol{d_{uv}}$, if $(u,v) \in \mathcal{T}^i$, put d_{uv}^i into ${}^1\boldsymbol{d}_{uv}$, else ${}^2\boldsymbol{d}_{uv}$. Correspondingly, ${}^1\hat{\rho}_{uv}(\tau)$ and ${}^1\hat{\Phi}_{uv}(\tau)$ denote the estimation of ${}^1\boldsymbol{d}_{uv}$, similarly ${}^2\hat{\rho}_{uv}(\tau)$ and ${}^2\hat{\Phi}_{uv}(\tau)$ for ${}^2\boldsymbol{d}_{uv}$. In order to proof the proposition, we first proof the two sides of the problem, which are the consistency of ${}^1\hat{\Phi}_{uv}(\tau)$ and ${}^2\hat{\Phi}_{uv}(\tau) < \Phi_{uv}(\tau)$ as $M \to \infty$. We will see later on, that the two parts will finally make the proposition tenable.

First, it has been confirmed that under the kernel density estimation technique, ${}^1\hat{\rho}_{uv}(\tau)$ is consistent, for each $\epsilon > 0$,

$$\lim_{M\to\infty} \mathbb{P}\left(|{}^1\hat{\rho}_{uv}(\tau) - \rho_{uv}(\tau)| > \epsilon\right) = 0$$

We briefly explain why it is established. Consider the mean square error of ${}^1\hat{\rho}_{uv}(\tau)$, $MSE({}^1\hat{\rho}_{uv}(\tau)) = Var({}^1\hat{\rho}_{uv}(\tau)) + (Bias({}^1\hat{\rho}_{uv}(\tau)))^2$. The estimator ${}^1\hat{\rho}_{uv}(\tau)$ is asymptotically unbiased, $\lim_{M\to\infty}(Bias({}^1\hat{\rho}_{uv}(\tau))) = 0$. While after derivation, we can also get $\lim_{M\to\infty} Var({}^1\hat{\rho}_{uv}(\tau)) = 0$. So ${}^1\hat{\rho}_{uv}(\tau)$ is consistent in the quadratic mean (weakly consistent). In the next moment, let $\Omega_n = \{|{}^1\hat{\rho}_{uv}(\tau) - \rho_{uv}(\tau)| \le \epsilon\}, \mathbb{P} \to 1$, then on Ω_n, $\forall\tau \in [0, \tau_{max})$,we have

$$\begin{aligned}
&\left|{}^1\hat{\Phi}_{uv}(\tau) - \Phi_{uv}(\tau)\right| \\
=&\Big|1 - \int_0^\tau {}^1\hat{\rho}_{uv}(t)dt - \left[1 - \int_0^\tau \rho_{uv}(t)dt\right]\Big| \\
=&\Big|\int_0^\tau {}^1\hat{\rho}_{uv}(t)dt - \int_0^\tau \rho_{uv}(t)dt\Big| \\
\le&\int_0^\tau \left|{}^1\hat{\rho}_{uv}(t) - \rho_{uv}(t)\right|dt \\
\le&\tau_{max}\cdot\epsilon
\end{aligned}$$

So we get

$$\mathbb{P}(\left|{}^1\hat{\Phi}_{uv}(\tau) - \Phi_{uv}(\tau)\right| \le \tau_{max}\cdot\epsilon) \ge \mathbb{P}(\Omega_n) \to 1$$

Then we konw that ${}^1\hat{\Phi}_{uv}(\tau)$ is consistent in the interval $[0, \tau_{max})$.

Next, for the other part data ${}^2\boldsymbol{d}_{uv}$, it is clear ${}^2\hat{\Phi}_{uv}(\tau) < \Phi_{uv}(\tau)$ when $M \to \infty$. For any $d_{uv}^i \in {}^2\boldsymbol{d}_{uv}$, the random sampling value τ_{uv} from $\rho_{uv}(\tau)$ was cut off by the real time difference d_{uv}^i, which indicates another shorter path (e.g. $u \to k \to v$) exists compared with $u \to v$. So that $d_{uv}^i = t_v^i - t_u^i < \tau_{uv}^i$. Then we have

$$\begin{aligned}
{}^2\hat{\Phi}_{uv}(\tau) &= 1 - \int_0^\tau {}^2\hat{\rho}_{uv}(\tau)dt \\
&= 1 - \frac{1}{\#{}^2\boldsymbol{d}_{uv}} \sum_{i=0}^{\#{}^2\boldsymbol{d}_{uv}} \int_0^{tau} K_h(t - d_{uv}^i)dt \\
&< 1 - \frac{1}{\#{}^2\boldsymbol{d}_{uv}} \sum_{i=0}^{\#{}^2\boldsymbol{d}_{uv}} \int_0^{tau} K_h(t - \tau_{uv}^i)dt
\end{aligned}$$

The consistency of $1 - \frac{1}{\#{}^2\boldsymbol{d}_{uv}} \sum_{i=0}^{\#{}^2\boldsymbol{d}_{uv}} \int_0^{tau} K_h(t - \tau_{uv}^i)dt$ is obvious as the previous proof. In this case ${}^2\hat{\Phi}_{uv}(\tau) < \Phi_{uv}(\tau)$ is simultaneously established.

XX

Finally, we come to a conclusion that as $M \to \infty$, the estimator is left-skewed, $\hat{\Phi}_{uv}(\tau) \le \Phi_{uv}(\tau)$. ■

The proposition can be understood in an easy way. If $(u,v) \in \mathcal{G}$ is not a diffusive path in some diffusion tree $\mathcal{T}^i$, it must be cut off by another shorter path through other nodes. As a result, the time difference d_{uv} is smaller than the corresponding sampling time τ_{uv}, which leads to the left skew of the estimator on the whole.

2) Right-deviation of Estimated Survival Function on False Edges: 非网络边节点对上的估计分布生存函数相对于网络边节点对上的估计分布生存函数是相对右偏的，即在支撑区间$[0, \tau_max)$的右侧，$\hat{\Phi}_{(u,v)\in\mathcal{G}}(\tau) > \hat{\Phi}_{(u,v)\notin\mathcal{G}}(\tau)$大概率成立。表现在生存函数的图中，红色估计线总是在蓝色估计线的左下方。且随着横坐标数值的增大，网络边节点对的估计分布生存函数$\hat{\Phi}_{(u,v)\in\mathcal{G}}(\tau)$（红线）和非连边节点对的估计分布生存函数$\hat{\Phi}_{(u,v)\notin\mathcal{G}}(\tau)$（蓝线）的距离不断增大。这说明，非连边节点对的首达时间差的分布情况由于受到扩散过程在网络拓扑中的随机性的影响，相对于网络边节点对来说更可能分布在区间的右侧，所以其生存函数的降幅才会比网络连边节点对的生存函数的降幅缓慢。

3) Overflow of Estimated Survival Function on False Edges : 此外，非网络边节点对的估计分布$\hat{\Phi}_{(u,v)\notin\mathcal{G}}(\tau)$在支撑区间$[0, \tau_max)$右侧的降速更缓慢，且大概率存在数值溢出的情况，现象表现在生存函数图中为$\hat{\Phi}_{(u,v)\notin\mathcal{G}}(\tau) > 0$。对于非网络边节点对的首

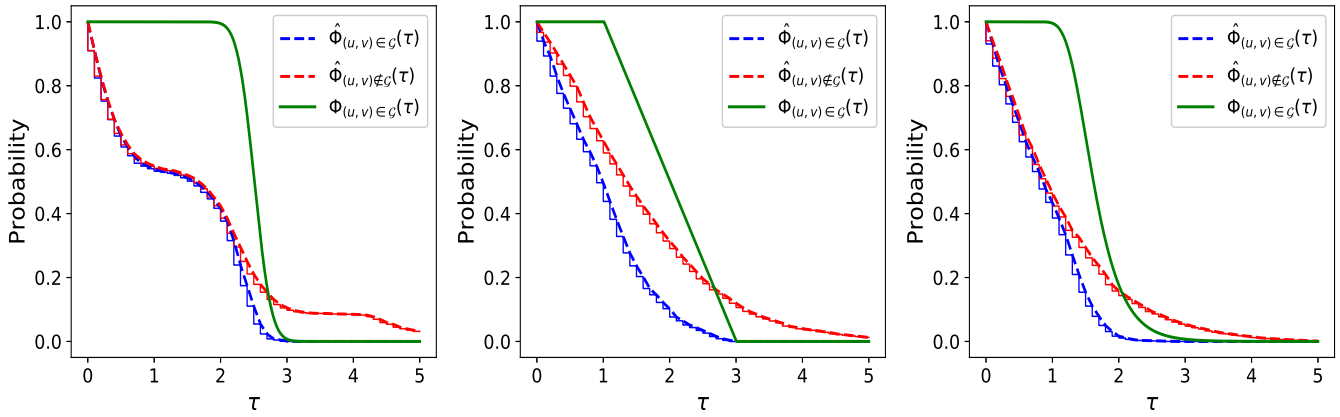

Fig. 1. Estimation on a single edge

达时间差，由于这个数值不是从等待时间分布采样得到的，不服从等待时间分布，再加上网络拓扑和扩散过程的影响，往往会出现首达时间差超出支撑区间的情况。为了方便统计，我们将首达时间差超出的数值都记录在区间的最右侧，表现在估计分布的图示中，我们可以发现非连边节点对的估计分布的概率密度函数为$\hat{\rho}_{(u,v)\notin\mathcal{G}}(\tau)$(蓝线)在支撑区间的最右端总会出现一个大值，这是所有溢出情况累加的结果；在生存函数中这种现象表现为$\hat{\Phi}_{(u,v)\notin\mathcal{G}}(\tau) > 0$，而对于而网络边节点对的估计分布$\hat{\rho}_{(u,v)\in\mathcal{G}}(\tau)$、$\hat{\Phi}_{(u,v)\in\mathcal{G}}(\tau)$(红线)和原始分布$\rho(\tau)$、$\Phi(\tau)$(紫线)是不存在这种情况的，即$\hat{\rho}_{(u,v)\in\mathcal{G}}(\tau)$、$\hat{\Phi}_{(u,v)\in\mathcal{G}}(\tau)$、$\rho(\tau)$、$\Phi(\tau) \simeq 0$

IV. RECONSTRUCTION ALGORITHM THROUGH DIFFUSIVE ARRIVAL TIMES

A. Reconsturction Algorithm

便于分析，我们在重构算法中使用均质假设，即时间累积网络中所有连边的等待时间分布都是相同的。首先，我们定义一个阈值θ和等待时间τ_θ，分别表示累积分布函数$\Phi_{uv}(\tau)$的上θ分位点处对应的原生存函数概率值和等待时间，即$\Phi_{uv}(\tau_\theta) = \theta$。从而，网络重构的问题转化为节点对首达时间差数据的估计分布的分类问题。我们认为当$\theta \to 0$时，对于一个网络边$(u,v) \in \mathcal{G}$，其节点对的首达时间数据出现$d_{uv} > \tau_\theta$的概率$P(d_{uv} > \tau_\theta) \to 0$。而对于整个时间累积网络的重构，只需要把所有可能的节点对遍历，判断每一个节点对上是否存在连边，最后对所有结果取并，即可得到重构后的时间累积网络。下面我们写出了整个重构算法的流程。

Algorithm 2: Reconstruction Algorithm through Diffusive Arrival Times

Input: $D = \{\mathcal{D}^i\}_{i=1,2,\dots,M}$, $\rho_{uv}(\tau)$
Output: $\hat{\mathcal{G}}$

$\hat{\mathcal{G}} \leftarrow EmptyGraph$;
for u *in* $\mathcal{V}$ **do**
  for v *in* $\mathcal{V}$ **do**
    $\hat{\rho}_{uv}(\tau) = \frac{1}{M}\sum_{i=1}^{M} K_h(\tau - d_{uv}^i)$;
    $\hat{\Phi}_{uv}(\tau) = \int_\tau^\infty \rho_{uv}(\tau)d\tau$;
    if $\hat{\Phi}_{uv}(\tau) > \epsilon$ **then**
      $\hat{\mathcal{G}} \leftarrow \hat{\mathcal{G}} \cup (u,v)$;
return $\hat{\mathcal{G}}$;

B. Indicators For Reconstuction Results

在我们的算法中，与重构结果最相关的参数是相对数据量（是扩散过程的样本数，是网络中节点的个数），以及剪枝阈值的选取。因此，根据数据量和剪枝阈值的变化，网络重构性能也会随之变化。怎样合理地设置这两个参数，以及在什么参数条件下重构效果最优，是我们接下来重点讨论的问题。为了标准化地衡量重构结果的好坏，我们使用两种标准曲线指标：ROC曲线和PR曲线[48][49]，以此衡量重构精度的高低。首先我们解释这两种曲线是如何绘制的，以及他们代表着什么样的物理意义。下面我们讨论一个含有正负样本的二分类问题。根据算法分类结果的不同，会出现以下几种情况:TP（True Positive）实际为正样本，被分类为正样本;FP（False Positive）实际为负样本，被分类为正样本;TN（True Negative）实际为负样本，被分类为

负样本;FN（False Negative）实际为正样本，被分类为负样本。为了不引起歧义，我们使用1表示正样本或分类为正样本，0表示负样本或分类为负样本。在我们使用的ROC曲线或PR曲线中，关注的重点是下面几个指标。

$$\mathrm{TPR}(\theta) = \mathrm{Recall}(\theta) = \frac{\mathrm{TP}(\theta)}{\mathrm{TP}(\theta) + \mathrm{FN}(\theta)}$$

$$\mathrm{FPR}(\theta) = \frac{\mathrm{FP}(\theta)}{\mathrm{FP}(\theta) + \mathrm{TN}(\theta)}$$

$$\mathrm{Precision}(\theta) = \frac{\mathrm{TP}(\theta)}{\mathrm{TP}(\theta) + \mathrm{FP}(\theta)}$$

对于ROC曲线，随着参数θ的变化，每一个θ都对应着以FPR为横坐标,TPR为纵坐标的二维坐标系上的一个点$[TPR(\theta), FPR(\theta)]$，所有点组成的轨迹就是该重构结果的ROC曲线。ROC曲线越贴近坐标轴的左上角，表示重构效果越好。同理，对于PR曲线，对应的以Recall为横坐标，Precision为纵坐标的二维坐标系上的点$[Recall(\theta), Precision(\theta)]$组成的轨迹就是该重构结果的PR曲线。PR曲线越贴近坐标轴的右上角，表示重构效果越好。

C. Simulations and Results

和上一章中的估计分布分析对应，我们同样选取了ER、SF、WS三种随机网络用于算法的验证。此外，额外加入了Football、Lattice2d、Sierpinski三种确定网络作为时间累积网络，以说明算法在一些实际社交关系网络或特殊结构网络上的重构也是有效的。等待时间分布取Gaussian、Uniform、Gumbel三种分布。在上述情况下，我们对随机时效网络上的扩散过程进行了大量仿真，并用仿真得到的扩散过程的首达时间数据进行时间累积网络的重构。在每一张图中，我们都绘制了在相对数据量$C = M/N$分别为0.2、0.4、0.6、0.8、1.0的情况下，随着剪枝阈值θ变化对应的ROC曲线和PR曲线。用F1-Score作为重构效果的最终衡量指标

$$\mathrm{F1}(\theta) = \frac{2 \cdot \mathrm{Precision}(\theta) \cdot \mathrm{Recall}(\theta)}{\mathrm{Precision}(\theta) + \mathrm{Recall}(\theta)}$$

我们记录下了在参数θ变化中最优情况下的F1值以及对应的TPR, FPR, Precision, Recall值，绘制成了相应的数据统计表。图表中每一个坐标点或数据都是在10次独立的仿真之后取均值得出的。

从曲线图可以看出，随着数据量C的，重构的ROC曲线和PR曲线分别移动至左上角和右上角，这表明重构精度是随着数据量不断提升的。另外值得一提的是，从图中可以看出，我们的算法在ROC曲线中的$\mathrm{FPR} = 0$一侧和PR曲线中的$\mathrm{Recall} = 1$一侧的取值是相对更多的，存在这个结果的原因是，我们的算法始终保证不漏掉实际存在的网络边，即使可能会引入一些非网络边。从而在所有情况下都保证了极高的精度。对于网络重构问题，我们更关心的问题是，如何找到所有可能存在的连边，即使可能会引入一些冗余边。这里需要特别说明的是，从重构算法的PR曲线来看，也许部分网络下的结果并不是完美，只有在较大数据量时才能达到很好的重构精度。这主要是因为网络重构问题的极度不平衡导致的。例如在一个100个节点的网络中，实际存在的连边只有$300 \sim 400$条，但所有节点对组成的可能的连边却有$(100 \times 99)/2 = 4950$条。因此，想从如此庞大的可能性中选取其中一小部分作为我们的连边候选，是具有很大难度的，尤其是我们只有节点首达时间信息的情况下。由于正负样本的不均衡，在重构时，为了保证所有的实际连边都能被找出，往往会增加一些错误的冗余边，这是在我们的重构问题中难以完全避免的。可以看到，尽管如此，我们的重构算法还是能在多种网络中取的很好的重构精度。

为了说明我们重构算法的简洁性和高效性，我们分析并比较了该算法与已有算法之间的性能差别。需要特别说明的是，复杂度是理论上重构问题能达到的最低时间复杂度，因为对于一个网络拓扑的重构，势必

TABLE I

RECONSTRUCTION RESULTS TABLE

	F1	TPR	FPR	Precision
		Gaussian		
$C = 0.2$	0.5335	0.8168	0.0828	0.3961
$C = 0.4$	0.8594	0.9773	0.0197	0.7669
$C = 0.6$	0.9502	0.9628	0.0041	0.9379
$C = 0.8$	0.9864	0.9954	0.0015	0.9775
$C = 1.0$	0.9893	1	0.0014	0.9788
		Uniform		
$C = 0.2$	0.4124	0.6509	0.0921	0.3019
$C = 0.4$	0.7603	0.8253	0.0210	0.7048
$C = 0.6$	0.9224	1	0.0102	0.8560
$C = 0.8$	0.9783	1	0.0027	0.9575
$C = 1.0$	0.9970	1	0.0004	0.9940
		Gumbel		
$C = 0.2$	0.5336	0.7871	0.0676	0.4270
$C = 0.4$	0.9305	0.9014	0.0173	0.7699
$C = 0.6$	0.9145	0.9488	0.0080	0.8826
$C = 0.8$	0.9441	0.9668	0.0052	0.9225
$C = 1.0$	0.9637	0.9719	0.0029	0.9558

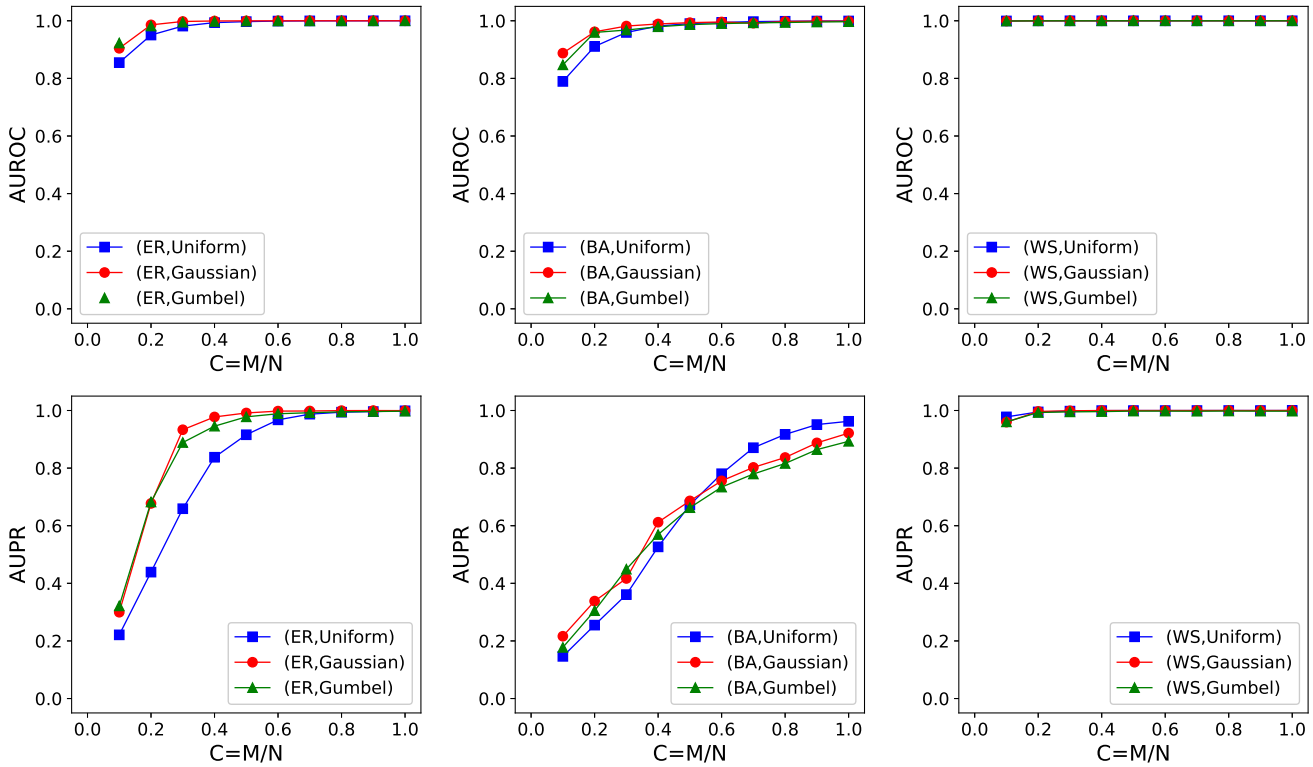

Fig. 2. ROC and PR Curve of Reconstruction Results

要遍历所有的节点对，并判断任意节点对之间是否存在连边。我们算法的最终复杂度是$O(N^2M)$，只是乘了扩散过程总个数。与Li Xun等人提出的基于MCMC（马尔可夫链蒙特卡罗方法）采样的拓扑重构算法相比较而言，显然我们的算法具有更简洁的思路和更高的运行效率，以及相似的重构结果。Li Xun等人的算法虽然和我们的算法具有相似的重构精度，但MCMC的稳态采样需要多次迭代、需要中间变量的更新和大量除法和log计算；而且MCMC迭代过程不可并行，无法设计高效的并行计算框架来重写算法。相比之下，我们的算法只需要对网络中所有可能的节点对进行一次遍历即可，涉及到的计算都是简单的乘法和加法计算。

TABLE II

RECONSTRUCTION RESULTS TABLE

	F1	TPR	FPR	Precision
		Gaussian		
$C = 0.2$	0.5335	0.8168	0.0828	0.3961
$C = 0.4$	0.8594	0.9773	0.0197	0.7669
$C = 0.6$	0.9502	0.9628	0.0041	0.9379
$C = 0.8$	0.9864	0.9954	0.0015	0.9775
$C = 1.0$	0.9893	1	0.0014	0.9788

此外，我们算法还有一个独特的优势，随着网络规模的增大，实现一定重构精度所需要的相对数据量是递减的。为了验证该结果，我们在ER作为时间累积网络，Gaussian作为等待时间分布的随机时效网络上进行了仿真实现，测量了在网络规模不断增大的情况下，使重构结果的值达到0.95所需要的最小数据量变化的情况。为了保证网络结构的相对相似性，我们取节点数为N的时间累积网络的连边数大致为3N。在这种情况下，影响重构结果的主要因素是网络规模，用节点的数

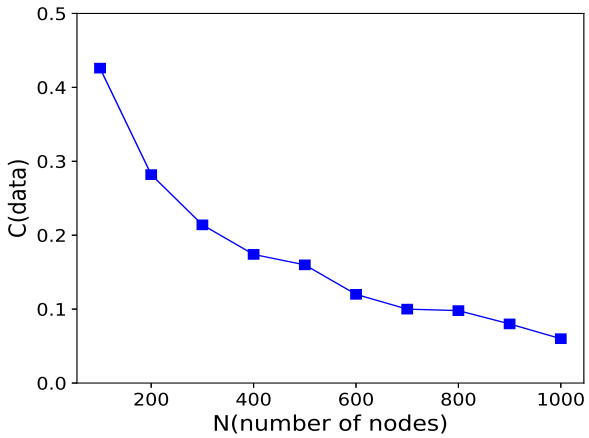

Fig. 3. Large Scale Network Reconstruction

量来衡量。从图示可以发现，随着网络规模的增大，重构需要的相对数据量是大约以负指数规律递减。在实际应用中，我们面对的很可能是超大规模的网络，而获得的数据量往往是有限的。这种情况下，我们的算法就显得尤为高效，仅通过相对于网络规模而言很小的一部分数据量，就可以达到很好的网络重构效果。

V. COCLUSION

回顾全文，我们利用随机时效网络及其上发生的扩散过程为理论模型，分析了网络边和非网络边在扩散过程中的性质差异，并以此为基础构建了一个利用首达时间数据重构随机时效网络中的时间累积网络的算法。该网络拓扑推断算法可以进行并行化处理，加倍提升大规模网络的重构速度。基于节点对层面的网络重构，可以在只知道局部信息的情况下重构出局部网络的拓扑结构，而不需要依赖所有节点的数据信息。使我们有可能精确地对目标节点或目标子网络进行重构。同时，重构需要的数据的结构很简单，仅仅是在网络中节点的首达时间信息。时间标签是我们在实际生活中最普遍、最易获得的信息，也不会涉及到个人隐私，这与大数据时代下的隐私保护是相符的。利用仿真结果，我们展示了算法在大规模网络中的优秀表现，以及对数据量的弱依赖。总体来看，我们的工作为基于时间标签数据的网络重构提供了一个完整的解决思路。算法的简洁性和高效性也保证了能够轻松移植到实际的网络重构或关系挖掘问题中。难以避免地，我们这种基于模型的重构算法有很强的假设条件。首先，随机时效网络零模型本身就是时效网络的一种简化，论文的研究中使用的又是一阶零模型，没有考虑网络中相邻连边的相关性。同时，我们为等待时间分布做了均质化假设，这要求个体之间的交互行为是相同的或服从同一规律。在此基础上，后续研究也许可以延伸讨论异质网络的重构问题。此外，如果考虑噪声数据的影响、或者数据的部分缺失，这些情况下怎样进行高精度的网络重构，都是需要进一步讨论的问题。因此，不论是模型设计、重构算法设计，还是实际因素对算法的影响、是否能应用于真实问题中，都需要继续进行更加细致和深入的研究和分析。网络重构问题仍然任重道远。

APPENDIX I
PROOF OF THE FIRST ZONKLAR EQUATION

Appendix one text goes here.

APPENDIX II

Appendix two text goes here.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.