复杂网络上疾病传播溯源算法综述

黄春林[1),2)] 刘兴武[1)] 邓明华[3),4),5)] 周杨[6),7)] 卜东波[1)]*

[1)](中国科学院计算技术研究所，北京市 100190)

[2)](中国科学院大学，北京市 100049)

[3)](北京大学定量生物学中心，北京市 100871)

[4)](北京大学数学科学学院，北京市 100871)

[5)](北京大学统计科学中心，北京市 100871)

[6)](中国疾病预防控制中心，北京市 102206)

[7)](内梅亨大学，内梅亨 荷兰 6525)

摘 要 流感、肺结核等呼吸道传染病严重威胁人类的健康，因此当疫情爆发时，快速、准确地推断疾病起源，对于疾病防控具有重要的理论意义和应用价值。和社交网络上的谣言传播以及计算机网络上的病毒传播不同，呼吸道疾病依赖于人际物理接触，而且具有更为复杂的疾病传播模型。在本篇综述里，我们首先介绍了人际接触网络、疾病传播模型和疾病传播溯源问题的形式化定义，以及溯源问题在传播时间、快照覆盖程度、传播源数量和传播源候选节点四个层面上的推广，给出了溯源算法的评价指标（准确率和错误距离）和基于贝叶斯极大似然估计的设计脉络；然后分别分析了现有的溯源算法，包括基于传染源中心性的算法、基于置信传播的算法、基于蒙特卡洛的算法、以及基于最小描述长度的算法。在这 4 类算法中，基于传染源中心性的算法最多，使用了包括传播中心性、Jordan 中心性、动态年龄和无偏中介中心性共 4 种中心性指标，并且基于传播中心性和 Jordan 中心性的算法被推广到更为一般的情形，如多个传播源、快照信息不完全等。我们分别在四种理想网络和两种真实人际接触网络下，实现并比较了常用溯源算法的性能。评估结果（包括准确率、错误距离、运行时间）表明：（1）溯源算法普遍对网络结构较为敏感；（2）多数算法对疾病传播参数具有鲁棒性；（3）相对于其他算法而言，动态消息传递算法尽管耗时几乎最长，但具有最高的准确度；（4）在耗时较短的算法中，无偏中介中心性具有相对较小的误差距离。根据实验结果，我们根据不同的使用场景推荐了不同的算法：（1） 当运行时间不重要时，我们推荐动态消息传递算法；（2） 相反，当我们希望快速溯源时，应该考虑基于无偏中介中心性的算法，当网络是随机树时 Jordan 中心估计算法更优；（3）反向贪心算法和动态年龄算法分别在随机网络和无标度网络上兼顾了准确率和运行时间。最后，我们总结了本文中介绍的所有溯源算法的适用性和时间空间复杂度，讨论了它们的实际应用以及后续的免疫措施，并提出未来的研究趋势，包括研究更准确的极大似然估计算法以提高算法的准确度、挖掘并利用传播过程中的信息以提高现有溯源算法的效率，以及考虑动态人际接触网络以提高算法的实用性等。

关键词 复杂网络；疾病溯源；极大似然；置信传播；蒙特卡洛

中图法分类号 TP18

论文引用格式：

黄春林,刘兴武,邓明华,周杨,卜东波, 复杂网络上疾病传播溯源算法综述,2017，Vol.40,在线出版号 No.5

HUANG Chun-Lin, LIU Xing-Wu, DENG Ming-Hua, ZHOU Yang, BU Dong-Bo, A survey on algorithms for epidemic source identification on complex networks, 2017,Vol.40,Online Publishing No.5

本课题得到国家科技重大专项（No.2008ZX10003009-005）、国家“九七三”重点基础研究发展规划基金项目（No.2012CB316502）、国家自然科学基金项目（No.11175224, No.11121403, No.31270834, No.61272318）、中国科学院理论物理研究所理论物理国家重点实验室开放工程项目（No.Y4KF171CJ1）资助. 黄春林，男，1987年生，博士研究生，计算机学会(CCF)会员（41640G）,主要研究领域为复杂网络上的疾病传播. 刘兴武，男，1976年生，副研究员，硕士生导师，主要研究方向为图算法、分布式算法. 邓明华，男，1969年生，教授，博士生导师，主要研究方向为生物信息学、系统生物学.. 周杨，女，1980年生，博士研究生，主要研究方向为结核病的分子流行病学. 卜东波（**通讯作者**），男，1973年生，研究员，博士生导师，主要研究方向为算法设计与分析、生物信息学.

A survey on algorithms for epidemic source identification on complex networks

HUANG Chun-Lin[1),2)] LIU Xing-Wu[1)] DENG Ming-Hua[3),4),5)] ZHOU Yang[6),7)] BU Dong-Bo[1)]

[1)](Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

[2)](University of Chinese Academy of Sciences, Beijing 100049)

[3)](Centre for Quantitative Biology, Peking University, Beijing 100871)

[4)](School of Mathematical Sciences, Peking Univerisity, Beijing 100871)

[5)](Center for Statistical Sciences, Peking University, Beijing 100871)

[6)](Chinese Centre for Disease Control and Prevention, Beijing 102206)

[7)](Radboud University Nijmegen, Nijmegen 6525, Netherlands)

Abstract Respiratory infectious diseases, such as influenza and tuberculosis, are among big threats to human health, making quick and accurate identification of epidemic source of the infectious disease an important issue theoretically and practically in the field of disease spreading and control. Different from the spreading of rumors in social networks or computer viruses in computer networks, epidemic spreading of respiratory infectious diseases relies on human physical interactions, and more complex spreading models. In this review, we first present the formal definition of human contact networks, disease spreading models, and the epidemic source identification problem together with generalization in four dimensions (spreading time, snapshot coverage, quantity of epidemic sources, and candidates of the epidemic source). We also present two evaluation indicators (accuracy and error distance) of source identification algorithms, and algorithm design principles based on Bayesian maximal likelihood evaluation. Consequently we summarize currently existing algorithms, which are based on centrality of epidemic source, belief propagation, Monte-Carlo technique, and minimum description length. The first one of these four categories, whose basis is centrality of epidemic source, employs different kinds of centrality measurements such as rumor centrality, Jordan centrality, dynamical age, and unbiased betweenness centrality. Moreover, algorithms with rumor centrality and Jordan centrality are generalized to more general scenarios, where disease spreads from several sources, or information of the snapshot is incomplete. After the introduction of all these algorithms, we implement and evaluate them on four idealized networks (random tree, scale-free network, small-world network, and random network) as well as two realistic human contact networks (one in a French primary school, and the other one in a Chinese university), with various transmission rates. The evaluation results, including accuracy, error distance, and running time, indicate the following four facts: (1) Most source identification algorithms are sensitive to network structures, showing different accuracies, error distances, and running times on different networks; (2) Most algorithms are robust to epidemic parameters; (3) The algorithms based on dynamic message-passing, though time-consuming, locate the epidemic source much more accurately than other algorithms; (4) Among all fast algorithms, the algorithms based on unbiased betweenness centrality have relatively small error distance. Based on experiment results, we recommend different algorithms in different realistic applications: (1) Dynamic message-passing is recommended in the case where running time is not cared; (2) On the contrary, the algorithm based on unbiased betweenness shall be considered if a fast source identification is higly valued, and the Jordan Center Estimation algorithm is much better on random trees; (3) Reverse Greedy and Dynamical Age take both accuracy and running time into consideration on random networks and small-world networks, respectively. Finally, we summarize basic settings of these algorithms, and compare their time and space complexities. The summaries are followed by the discussion of their practical applications, as well as the consequent vaccination strategies. We list research directions of epidemic source identification in the future, including developing more advanced methods

of estimating likelihoods to improve the accuracy of source identification, utilizing more information in the spreading process to accelerate existing source identification algorithms, and designing algorithms on dynamic networks to adapt them for realistic scenarios.

Key words complex network; epidemic source identification; maximum likelihood; belief propagation; Monte-Carlo

1 引言

流感、肺结核等呼吸道传染病具有高致死率和传染性，严重威胁人类健康[1-3]。因此，当疫情爆发时，快速、准确地推断疾病起源，对于疾病防控具有重要的现实意义。简要地说，所谓疾病溯源[4, 5]，是指基于流行病学中的疾病传播模型[1, 6-9]和人际接触网络[10-12]，根据已观察到的所有或部分患病个体情况（常称为快照，snapshot），推断该疾病的源头。

与传染病溯源问题类似的问题有计算机网络上的病毒溯源问题，以及社交网络上的谣言溯源问题[13]。上述问题具有相近的目的，但是在如下两点上存在着显著的不同：

（1）网络结构不同：呼吸道传染病的传播依赖于人际近距离接触，这种接触是真实的物理接触，与社交网络和计算机网络相比具有不同的特性，显著影响疾病传播；

（2）传播模型不同：在信息传播过程中，每个个体只有两种可能的状态，即激活态（active）和未激活态（inactive）[13]；其中激活态代表某个个体在传播过程中接受到信息，否则处于未激活态。相比而言，疾病传播模型能够包括多达 4 个状态，即易感态（Susceptible，记为 S）、潜伏态（Exposed，记为 E）、感染态（Infected，记为 I），以及恢复态（Recovered，记为 R）[6, 7, 9]。

即便在只采用两个状态的情况下，信息传播与疾病传播依然存在显著的不同：信息传播通常使用独立级联（Independent Cascade, IC）模型[14]和线性阈值（Linear Threshold, LT）模型[14]，而疾病传播则有 SI（Susceptible-Infected）和 SIS（Susceptible- Infected-Susceptible）两种情形[7]。在独立级联模型中，每个处于激活态的个体仅在刚被激活时传播一次它接受的信息；在线性阈值模型中，每个处于未激活态的个体仅在它周围已处于激活态个体的数量超过某个阈值时才会接受信息；SI 模型与独立级联模型类似，但是每个处于被感染状态个体在不停地传播病毒而不是仅传播一次，SIS 模型则在 SI 模型的基础上考虑了个体患病后再次恢复健康的情况。由于信息传播与疾病传播存在上述明显的不同，信息传播溯源算法不能完全适用于疾病传播溯源问题[14-21]。简要地说，疾病传播溯源问题提出了如下挑战：

（1）与信息传播模型中只有两个状态不同，疾病传播模型通常会出现多个状态的情形（如 SEIR 模型，Susceptible-Exposed-Infected-Recovered），甚至被感染节点还会恢复到健康状态（如 SIS 模型），形成了复杂的状态转移逻辑；

（2）在推断传播源头的过程中，通常需要计算指定传播源时观察到特定快照状态的概率，而这种概率计算往往具有指数级的时间复杂度；

（3）在实际疫情发展过程中，通常只能获得部分个体的状态信息，很难获得所有人的健康状态信息。这种信息的不完整性进一步加剧了溯源问题的难度。

本文综述了在疾病传播过程中根据观察到的状态信息，来推断疾病源头的溯源算法。本文首先介绍了溯源问题的预备知识以及形式化定义（第 2 节）；然后分类介绍了基于传染源中心性的溯源算法（第 3 节）、基于置信传播的溯源算法（第 4 节）、基于蒙特卡洛的溯源算法（第 5 节）、以及基于最小描述长度的溯源算法（第 6 节）；最后，本文通过实验比较了这些溯源算法的性能，对它们做了分析与总结，并提出了溯源问题的发展趋势（第 6 节）。

2 疾病传播溯源问题简介

疾病传播依赖于两个关键因素，即人际接触网络和疾病传播模型，分别介绍如下。

2.1 人际接触网络

传染病的传播依赖于支持病原体在个体之间传播的人际接触网络，其中网络节点代表处于健康状态或者被感染状态的个体，而节点之间的连边代表人与人之间的近距离物理接触。

人际接触网络可以形式化定义为图 $G(V, E)$，其中 V 为节点集合，E 是定义在 V 上的边集。在图 G 中，某个节点 i 的邻居定义为与之直接接触的节点集合，记为 ∂i；节点 i 的度定义为 ∂i 中节点数目，即和节点 i 有物理接触的人数。图 G 中任意两个节点之间可定义测地距离，即节点间最短路径中的边数。目前溯源问题的研究通常在无权无向的网络上进行，即网络中边上都不附带权重和方向信息。

溯源算法性能与网络的拓扑结构密切相关[22]。如果网络中不存在环形结构，则称该网络为树（tree）；如果树上每个节点的度都相同，则称之为正则树（regular tree）；如果正则树上每个节点的度都为 2，则称之为线状图（line graph）。

在非正则树中，Shah 等人[4, 5] 进一步定义了几何树（geometric tree）。几何树为按多项式数量级增长的树，即所有与根节点（或传播源）的测地距离为 d 的节点，其数量 $n(d)$ 满足 $bd^{\alpha} \le n(d) \le cd^{\alpha}$，其中 α 是几何树的参数，b 和 c 是常数。Luo 等人[23-25] 进一步定义了双源头几何树。简单地说，若 $\rho(s_1, s_2)$ 代表在图 G 上两个传播源 s_1 和 s_2 之间的最短路径上的节点，$u \in \rho(s_1, s_2)$ 是这些节点的邻居，$T_u(s_1, s_2)$ 表示以 u 作为根节点、远离 s_1 和 s_2 方向的子树，则当 $T_u(s_1, s_2)$ 上与根节点 u 距离为 r 的节点个数 $n(u, r)$ 满足 $br^{\alpha} \le n(u, r) \le cr^{\alpha}$ 时，图 G 被称为双源头几何树。

上述网络类型之间的关系如图 1 所示。

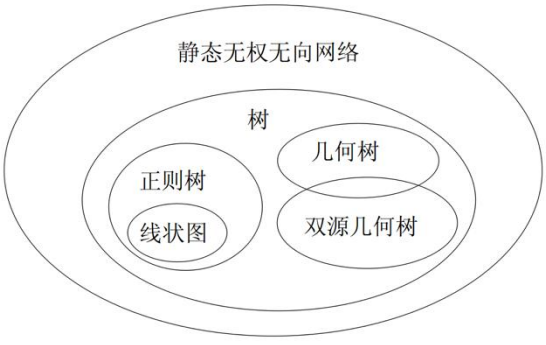

图 1 常用人际接触网络分类

2.2 疾病传播模型

现有模型通常采用自动机的形式来刻画个体在疾病传播和患病过程中的状态变化，即在任一时刻，每个个体可以处于易感态 S、潜伏态 E、感染态 I，以及恢复态 R 四种状态之一，并以一定概率在状态之间切换。

常用的疾病传播模型有 SI、SIS、SIR、SEIR、SIRS 等模型[7, 26]（见图 2）。然而在溯源问题中，现有研究集中在 SI、SIS 和 SIR 三个模型，对于 SEIR 和 SIRS 模型则研究较少。上述三种模型简要介绍如下：

（1）SI 模型：在 SI 模型中，每个节点只有 S 和 I 两种可能状态；在传播初始时刻，网络中只有一个节点或者少数节点处于 I 状态，而其他节点都处于 S 状态；在传播过程的每个时间步，任一处于 I 状态的个体都以相同的概率感染它的每个处于 S 状态的邻居[27-30]。

疾病传播溯源问题经常会对上述 SI 基本模型做一些扩展，比如假设 I 状态的节点以概率 1 感染其所有 S 状态的邻居节点，但是在每条边上的传播时间服从独立相同的指数分布[4, 5, 24, 25, 31-34] 或者高斯分布[35, 36] 等。

（2）SIS 模型：与 SI 模型相同，SIS 模型中每个节点也是只能取 S 和 I 两种可能状态；而与 SI 模型不同的是，任一 I 状态节点在每个时间步都以相同的概率变回健康状态（S 状态），而且变回 S 状态的节点依然有可能被其处于 I 状态的邻居节点再次感染[37]。

（3）SIR 模型：在 SIR 模型中，节点可能的状态集合中增加了一个 R 状态。与 SIS 模型类似，所有 I 状态节点在每个时间步都以相同的概率变回健康状态；但是不同的是，SIR 模型中被感染节点变回的健康状态不是 S 状态而是 R 状态，变成 R 状态的节点将一直处于 R 状态，不再被感染。SI 模型可以看成 SIR 模型在节点恢复健康概率为 0 时的特例[38-42]。

除上述常用模型之外，文献[43] 还定义了三种传播模型，即雪崩模型（snowball）、随机游走模型（random walk）和接触过程模型（contact process）。这些模型类似 SI，但是传播参数上有所不同。简要地说，在雪崩模型中，病原体在每个时间步以概率 1 从被感染节点传播到它的邻居节

点，因此传播路径形成了以传播源作为根节点的宽度优先搜索树。在随机游走模型中，假设初始时刻在源头节点存在有限数量的病原体，它们以每个时间步一跳的速度相互独立地在网络中随机游走，同时将疾病传播给接触到的所有节点。当病原体数量达到无限大时，随机游走模型退化为雪崩模型。接触过程模型与 SI 传播模型完全相同，当在每条连边上的传播概率为 1 时，接触过程模型同样退化为雪崩模型。

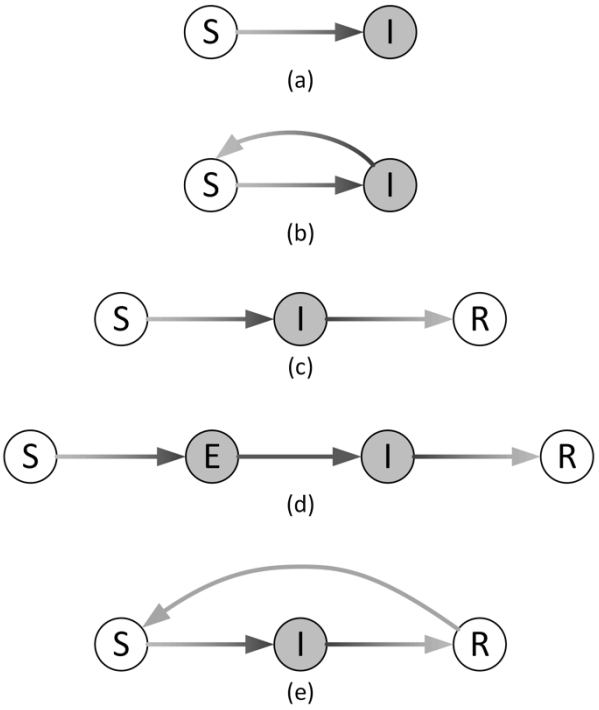

图 2 五种疾病传播模型：(a) SI (b) SIS (c) SIR (d) SEIR (e) SIRS

2.3 疾病传播溯源问题

2.3.1 问题描述

假设在人际接触网络 $G(V, E)$中，疾病按照 SI 模型从传播源 $s \in V$ 在网络中进行传播；经过时间 t，我们观察到网络中所有节点所处的状态 $\boldsymbol{O}(t)$（也称为“快照”）。溯源问题的目标就是根据网络 G、时刻 t 时所有节点的状态信息 $\boldsymbol{O}(t)$，推断传播源 s 是哪个节点（见图 3）。

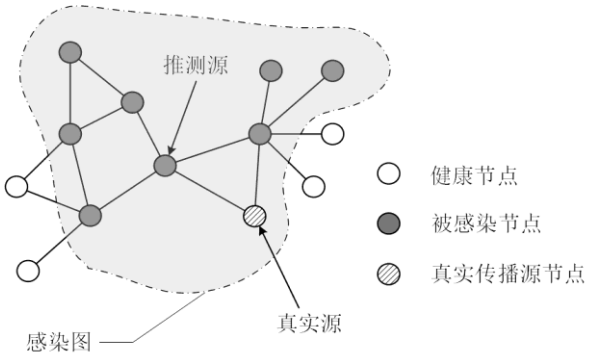

图 3 溯源问题示意图。

2.3.2 溯源问题的四个扩展

按照已知信息的不同，溯源问题可以从上述基本定义出发在四个层面上进行扩展，描述如下：

(1) 传播时间 t：在基本定义中，快照中每个节点的状态都是疾病传播过程经过 t 时间以后的结果，这里传播时间 t 一般为已知量[35, 38-40, 44-46]。然而在实际情况中，传播时间 t 还可以是未知量[4, 5, 23-25, 29, 30, 33, 37, 41, 42, 47, 48]。

(2) 快照覆盖程度：在基本定义中，快照中的状态信息包含了所有的网络节点[4, 5, 27, 28, 30, 31, 33, 34, 39, 43-45, 47-49]；然而更符合实际情况的是快照只覆盖了部分节点[29, 32, 35, 36, 40]。

(3) 传播源数量：在基本定义中，传播源的数量只有 1 个[4, 5, 33, 44, 46, 48, 50]，但是在实际情况中，可能存在多个传播源，而且传播源数量可能是已知的[29, 38, 49]，也有可能是未知的[25, 27, 28, 41, 45, 47]。

(4) 传播源的候选节点：在基本定义中，传播源可能是接触网络中任意一个节点；而在实际应用中，可能预先已知传播源的候选节点[31]。

在本综述中，如果不加特别说明，溯源问题满足以下条件：(1) 传播时间相同且未知；(2) 快照包含所有节点的状态信息；(3) 传播源只有 1 个；(4) 无传播源的候选节点。

2.3.3 算法性能评价指标

对单个传染源的溯源算法来说，性能评价有两个指标：

（1）准确率（又称检测率，detection rate），指使用该算法能正确找到源头的几率；

（2）错误距离（error distance），指算法推断出的传播源与真实传播源的测地距离。例如图 3 中推测源的错误距离为 1。

针对多个传染源的溯源问题，Luo 等人提出平均错误距离和最小感染区重合度两个性能评价指

标[25]，在此不赘述。

2.4 溯源算法脉络

基于贝叶斯理论，可以将溯源问题写成如下的极大后验估计或极大似然估计的形式[4, 5]：

$$\hat{s} = \arg\max_{s\in V} P(s \mid \boldsymbol{O}) = \arg\max_{s\in V} P(\boldsymbol{O} \mid s) \quad (1)$$

其中 $P(\boldsymbol{O}|s)$是假设 s 为传播源的情况下，疾病经接触网络扩散后得到各个节点的状态观察 $\boldsymbol{O}$ 的概率。

由于直接计算使得 $P(\boldsymbol{O}|s)$或 $P(s|\boldsymbol{O})$最大的传播源已经被证明是#P-hard 的[5, 51, 52]，因此现有的溯源算法常常采用如下三种思路：

（1）不直接计算 $P(\boldsymbol{O}|s)$或 $P(s|\boldsymbol{O})$，而是根据传染源的某些统计性质进行推断，比如各种中心性算法等；

（2）近似计算 $P(\boldsymbol{O}|s)$或 $P(s|\boldsymbol{O})$，比如置信传播算法等；

（3）采用模拟仿真策略估计 $P(\boldsymbol{O}|s)$或 $P(s|\boldsymbol{O})$，比如蒙特卡洛方法等；

（4）不直接估计 $P(\boldsymbol{O}|s)$或 $P(s|\boldsymbol{O})$，而是估计它们的上界或下界，比如 NetSleuth 算法。

基于上述分析，我们将现有溯源算法按照基本思想、典型算法、实验结果的形式分类总结如下。

3 基于传染源中心性的溯源算法

在人际接触网络中，所有已感染节点诱导出的一个子图，称为感染图[4]，其中包括所有的被感染节点，以及它们之间的连边。图 3 给出了感染图的一个实例。

直观上看，感染图是疾病从传播源出发按一定的规律向外扩散形成的，因此传播源是感染图的中心节点，具有某种中心性。基于中心性的算法的基本思想就是直接求取具有中心性的节点，作为传染源的估计。

在溯源问题提出之前，研究人员已提出多种网络中心性指标，例如度中心性（degree centrality）、距离中心性（distance centrality）、紧密度中心性（closeness centrality）、中介中心性（betweenness centrality）、特征向量中心性（eigenvector centrality）等[22, 43]。然而 Comin 等人的实验表明上述通用的中心性定义并不适用于溯源问题[43]。针对溯源问题，研究人员分别提出了传播中心性（rumor centrality）、Jordan 中心性（Jordan centrality）、动态年龄（dynamical age）和无偏中介中心性（unbiased betweenness centrality），详细介绍如下。

3.1 传播中心性

3.1.1 基本思想

Shah 等人针对正则树上的 SI 传播模型（每条边上的传播时间服从参数为 1 的指数分布）提出一个新的节点中心性指标——传播中心性（rumor centrality），并据此提出了第一个解决溯源问题的算法[4, 5, 33]。

该算法认为：感染图中某个节点是传播源的概率，与疾病从该节点出发感染其他节点的所有可能顺序的计数成正比。因此，可以使用下式计算传播中心性最大的节点，作为传染源的估计[4]：

$$\begin{aligned}\hat{s} &= \arg\max P(\boldsymbol{O} \mid s) \\ &= \arg\max_{s\in G_N} P(G_N \mid s) \\ &= \arg\max_{s\in G_N} R(s, G_N)\end{aligned} \quad (2)$$

其中 G_N 表示感染图（N 是感染图中的节点数量），$R(s,G_N)$ 是疾病从节点进行传播时，感染图 G_N 中所有节点的所有可能被感染顺序的计数，定义为传播中心性。网络中传播中心性最大的节点定义为传播中心（rumor center）。

Shah 等人指出，在正则树上，感染图的传播中心是传播源的极大似然估计；即使在一般树上，传播中心也与极大似然估计没有显著差别[4]。进一步地，Shah 等人给出了树型感染图 G_N 上任意节点 v 的传播中心性的计算公式[4]

$$R(v, G_N) = N! \prod_{u\in G_N} \frac{1}{T_u^v} \quad (3)$$

其中 N 是感染图 G_N 里的节点总数，T_u^v 是 G_N 中以节点 u 作为根，向远离 v 的方向展开的子树。为简化符号，我们在本文中也用 T_u^v 表示该子树中

的节点个数。图 4 是T_u^v的一个示例。

3.1.2 典型算法

根据树形图上T_u^v的性质$T_u^v = N - T_v^u$，Shah 等人给出了在 $O(N)$时间内使用 $O(N)$空间计算出树形感染图G_N上所有节点传播中心性的消息传递算法（Rumor Centrality Message-Passing Algorithm，RCMP）[4]。当前传播结果所对应的源头就是传播中心性最大的节点，也就是传播中心。因为此算法只能用于定位单个源头，所以也被 Luo 等人称为单源头估计（Single Source Estimation, SSE）算法[23]。

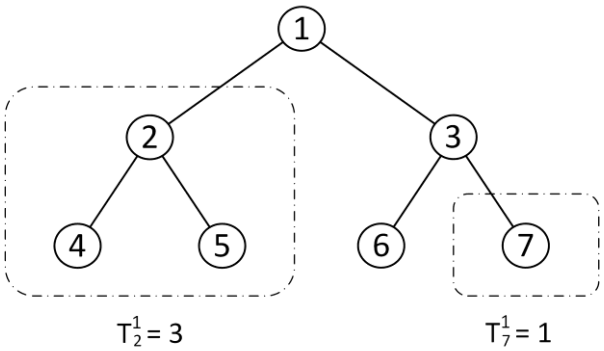

图 4 T_2^1和T_7^1的一个示例[4]

3.1.3 实验结果

Shah 等人[4, 5, 33]给出了此算法的性能：在线状图上，算法准确率接近于 0；在度 d>2 的正则网络上，当传播时间$t \to \infty$时使用此算法推断传播源的准确率α_d从 1/4 起随 d 递增，且在 d>20 后达到稳定值 1-ln2；在几何树上，当疾病在接触网络中每条连边上的传播时长分布满足一定条件，且传播时间$t \to \infty$时，算法准确率趋近于 1。

Shah 等人发现，在树形网络上传播中心性与距离中心性等价，而当网络非树形结构时，传播中心性比距离中心性更优。

3.1.4 推广和扩展

（1）推广到一般网络

当感染图G_N不是树形结构时，无法直接使用原有的算法来计算传播中心性，因此无法定位源点。

为解决此问题， Shah 等人假设疾病传播沿着最短路径进行，形成以传播源为根节点的宽度优先搜索树（见图 5），并估计传染源为[4, 5]：

$$\hat{s} = \arg\max_{s \in G_N} R(s, T_{bfs}(s)) \qquad (4)$$

其中$T_{bfs}(s)$就是在感染图G_N中以节点 s 作为根节点形成的宽度优先搜索树。Luo 等人将此算法称为 SSE-BFS 算法[25]。

由于此时需要对每个可能的传播源生成一棵宽度优先搜索树，然后在树上计算该传播源的传播中心性，所以 SSE-BFS 算法的时间复杂度为 $O(N^2)$，空间复杂度为 $O(N)$。

实验结果表明：在小世界网络[53]上，传播中心算法以 16%的准确率找到传播源；在真实的美国电力网络上，准确率为 3%；然而即便推断错误，错误距离也仅有几跳而已。

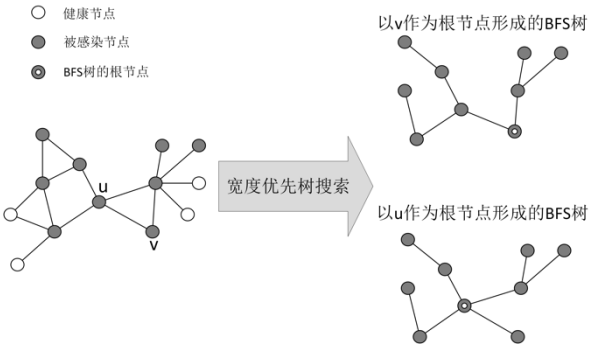

图 5 宽度优先搜索树的示意图。图中给出了从两个不同的节点 u 和 v 开始分别构造出的宽度优先搜索树。

（2）推广到多源问题

Luo 等人将单源头溯源问题中的传播中心性推广到了多源头的情况，其中疾病传播模型保持不变[23-25]。

具体地，Luo 等人借鉴 Shah 等人在文献[4, 5]中的思路，当传播网络是正则树时，从给定的传播源得到给定感染图的概率与从给定传播源开始传播疾病得到的所有可能传播顺序计数成正比的结论（即$P(G_N \mid S) \propto C(S \mid G_N)$，$G_N$表示包含 N 个节点的感染图），进而推导出[23-25]

$$\hat{S} = \arg\max_{|S|\le k_{\max}} P(G_N \mid S) = \arg\max_{|S|\le k_{\max}} C(S \mid G_N) \quad (5)$$

其中 $k_{\max}$ 是源头数量的最大可能值。

当源头只有 2 个且感染图为树形结构时，Luo 等人给出了 $C(S \mid G_N)$ 的精确解，以及在 $O(N^2d^2)$ 时间内计算出使得 $C(S \mid G_N)$ 值最大的集合 S 的双源头估计（Two Source Estimation，TSE）算法，此处 d 指感染图中最大的节点度。

Luo 等人证明，当传播网络是单源头几何树且传播时间 $t \to \infty$ 时，TSE 算法找到的两个源头以概率 1 相邻且其中一个是真正的源头；当传播网络是双源头几何树[25] 且传播时间 $t \to \infty$ 时，TSE 算法以概率 1 准确定位两个传播源。

对于源头数多于两个的情形，Luo 等人给出了基于不同源点的感染图划分算法 IP（Infection Partitioning），将感染图分成几块互不相交的节点集合，称为感染区（infection region），各自对应一个源头；然后利用 IP 算法给出了时间复杂度为 $O(N^2k_{\max}^3)$ 的多源估计划分（Multiple sources Estimation and Partitioning, MSEP）算法，这里 $k_{\max}$ 是传播源数量的上界，在执行算法前预先估计出来或者给定。

接下来 Luo 等人基于疾病按照最短路径传播的假设，使用类似 Shah 等人提出的方法[5]，继续将多源头的源点估计算法 MSEP 推广至一般的传播网络，在算法过程中不直接使用感染区，而是在感染区内从相应的源头开始做宽度优先搜索获得 BFS 树，再进行接下来的处理。作者将这种使用 BFS 的算法称为 MSEP-BFS，时间复杂度依然为 $O(N^2k_{\max}^3)$。

实验表明，MSEP 算法在正则树上计算出传播源数量的正确率达到 71%，这一数值在几何树上达到 93%。在小世界网络上，由于需要将原始网络近似成 BFS 树，MSEP-BFS 算法推断传播源数量的正确率只有 69.2%。在真实的美国电力网络上的传播实验表明，MSEP 算法能够达到 59%以上的感染区重合度。

（3）推广到快照只覆盖部分节点信息的情形

Karamchandani 等人[32] 将正则树或几何树上单源头问题的传播中心性算法推广到了快照只覆盖部分信息的情形——每个节点在快照中以概率 p 报告其状态。

该算法先依据已被感染的节点，诱导出一个子图 G_N^p。如果该子图不连通，那么需要向子图中尽可能少地增加未知状态的节点，使得从已感染节点和额外增加的节点能够诱导出连通子图（称为报告传播子图, reported rumor subgraph）。报告传播子图通常会包含所有的已知被感染节点和一部分未知状态的节点。最后在此报告传播子图上运行 SSE 算法找到的源节点即为推断出的传播源。

Karamchandani 等人证明：在几何树上，p>0 时的算法准确率与 p=1 时相当；在正则树上，当 p 大于某个阈值时，算法准确率与 p=1 时相差不大。

（4）推广到传播源有候选节点的情况

Dong 等人[31] 将正则树上单源头溯源问题推广到存在候选节点的情况，并提出相应的算法：直接选择被感染的候选节点中传播中心性最高的节点，作为传染源的估计。

当感染图不是树形结构时，同样使用从源点出发的宽度优先搜索树，所以算法的时间复杂度在树上是 $O(n)$，在一般网络上是 $O(n^3)$，空间复杂度在树和一般网络上都是 $O(n)$，此处 n 是感染图的节点数量。

当传播网络是正则树时，Dong 等人分析了算法准确率在不同候选节点类型（所有节点、连通子图、数量确定）下的精确解，并发现算法准确率随着正则树的度变大而变大，随着被感染节点数量的增多而变小。

（5）推广到有多个独立快照的情况

Wang 等人[34] 将单源头的溯源问题推广到包含多个传播快照的情形，即从同一个传播源开始有多次独立的传播过程，在每次传播过程中分别选择一个时刻，给出所有节点的状态信息。

Wang 等人发现，如果这些快照不独立（例如是同一次传播不同时刻的快照），则多个快照与最开始的快照相比并不能提高算法准确率。只有当这些快照相互独立，也就是属于多次不同的传播实验时，才有意义。Wang 等人将传播中心性推广到多次独立传播的情况，称为联合传播中心性

（union rumor centrality），定义为[34]

$$R(u,G_{n_1},\cdots,G_{n_k})=R(u,G_{n_1})\cdots R(u,G_{n_k}) \quad (6)$$

其中 G_{n_i} 为第 i 次独立传播的感染图，$R(u,G_{n_i})$ 是原始的传播中心性。与传播中心类似，联合传播中心（union rumor center）是联合传播中心性最大的节点。Wang 等人给出了计算树型网络上多个独立传播快照的联合传播中心性的算法 URCC（Union Rumor Centrality Calculation），并把它推广到了一般图的情形。该算法的空间复杂度为 $O(nk)$，在树型网络上的时间复杂度为 $O(nk)$，在一般网络上的时间复杂度为 $O(n^3k)$，这里 n 为传播网络中的节点数量。

Wang 等人通过模拟实验发现：在小世界网络和无标度网络上，从多个独立快照计算联合传播中心性比从单个快照计算传播中心性更能够准确找到传播源；即便不能正确找到，错误距离也较小。

3.2 Jordan中心性（Jordan centrality）

3.2.1 基本思想

由于公式(1)中的似然 $P(\boldsymbol{O}|s)$通常难以计算，Zhu 等人[42] 在单源头的溯源问题中提出，用最可能的疾病传播路径对应的源点来近似后验概率最大的源点。这种近似方法虽然不够准确，但是方便计算。

Zhu 考虑这样一种情形：疾病传播使用 SIR 模型，但在快照中将 S 状态和 R 状态都识别为“健康”状态，而不能将它们区分开。为解决这类溯源问题，Zhu 等人借鉴文献[54] 中离心率（eccentricity）的思想，定义感染离心率（infection eccentricity）为到所有被感染节点距离的最大值，并定义了类似 Jordan 中心（Jordan center）的 Jordan 感染中心（Jordan infection center）为给定被感染节点集合时感染离心率最小的节点。

Zhu 等人证明，在无限大的树形网络上，当疾病从单个源头开始按 SIR 模型传播一段时间后，若快照中仅记录了根据所有处于 I 状态的节点，那么从最可能传播路径推断出的源点是这些 I 状态节点所对应的 Jordan 感染中心。

3.2.2 典型算法

为尽快找到网络中的 Jordan 感染中心，Zhu 等人提出了反向传播算法（Reverse Infection Algorithm）[42]。

算法的基本思想是：首先，每个处于 I 状态的节点向网络中的邻居节点发送带有自己身份标签的消息，不同的节点发送的消息各不相同；然后在接下来的每个时间步，每个携带消息的节点都向邻居节点发送它所携带的所有消息，并接收所有由邻居节点发送来的消息，与自己本身携带的消息进行整合，以保证每个身份标签的消息最多只携带一份；当在某个时间步有节点获得所有 I 状态节点发送的消息时，算法结束，此时所有携带所有消息的节点都是 Jordan 感染中心。

本算法被 Luo 等人称为单 Jordan 中心估计（Single Jordan Center estimation, SJC）算法[29]。该算法的时间复杂度为 $O(n_{\mathrm{I}}n^2)$，空间复杂度为 $O(n_{\mathrm{I}}n)$，这里 n 是传播图中节点数量，n_{I} 是被感染节点的数量。

3.2.3 实验结果

Zhu 等人[42] 证明，在正则树上使用反向传播算法得到的Jordan感染中心与真实的源点的距离以很高的概率小于一个常数。实验证实此算法在正则树上可以达到 60%的准确率。

3.2.4 推广和扩展

（1）推广到 SIS 传播模型

Luo 等人将 Jordan 感染中心的定位方法推广到 SIS 传播模型[37]。SIS 传播模型与 SIR 传播模型类似，原本被疾病感染的节点有可能恢复到健康状态，而且此传播模型天然无法确定一个健康的节点是否曾经被疾病感染过。

Luo 等人证明，在 SIS 传播模型下，无限大树图上的Jordan感染中心同样是最可能传播路径对应的源头。

实验结果显示在度超过 4 的正则图上，SJC 算法的准确率可以达到 55%；即使没有找到正确的传播源，错误距离也在 4 跳以内。

（2）推广到一个特殊的 SI 传播模型

Luo 等人还将此溯源方法推广到了一个特殊的 SI 模型[30]。在这个 SI 模型中，任何与被感染节点相邻的健康节点被感染的概率都相同，与被感染

的邻居数量无关。在传播一段时间以后给出的状态信息中，所有健康节点的状态都能正确显示出来，但是被感染节点的状态只有一部分能正确显示，称为显式（explicit）状态，另一部分将与健康状态混淆，称为隐式（hidden）状态。

Luo 等人给出了在 $O(n)$时间内、使用 $O(n)$的空间、通过Jordan感染中心解决树形网络上单源头源点估计问题的 Jordan 中心估计（Jordan Center Estimation, JCE）算法[30] 。此算法与 Hedetniemi 等人提出的算法[55] 类似，先假设信息从叶子节点向根传播，然后假设信息从根向叶子传播，在传播过程中计算出 Jordan 中心。实验表明，JCE 算法在正则树上的平均错误距离只有 0.36。

（3）在特殊 SI 传播模型的基础上推广到多源问题

Luo 等人继续将此溯源算法推广到多源头的情形[29] 。他们定义 k-Jordan 感染中心为感染离心率最小的 k 个节点，并证明：在无限大的树形网络上，若使用最可能的疾病传播路径来估计传播源，则单源问题的解依然是所有显式状态节点的 Jordan 感染中心，而在多源问题中则是所有显式状态节点的 k-Jordan 感染中心。

Luo 等人给出了多源问题中计算 k-Jordan 感染中心的多 Jordan 中心估计（Multiple Jordan Center estimation, MJC）算法，该算法主要包括两个步骤：第一步是根据已选的传播源将所有的显式状态节点划分成 k 个互不相交的 Voronoi 集合；第二步是在各个 Voronoi 集合对应的感染图中分别使用 SJC 算法来求解各自的传播源。在随机选择 k 个初始传播源后，算法重复迭代执行上述两个步骤，直到推断出的多个源头收敛，或者迭代次数达到上限。

（4）在特殊 SI 模型的基础上推广到一般网络

对一般网络，Luo 等人虽然仍旧寻找最可能传播路径对应的传播源，但不再直接将Jordan感染中心作为传播源，而是首先寻找与被感染节点集合一致的感染树（infection tree），然后计算可生成该感染树的最可能传播路径，从而估计传播源为该树的 Jordan 感染中心[30] 。

Luo 等提出两种方法求解感染树和传播源：第一种方法是将问题建模成二次约束混合整数二次规划问题[56] 并求解，但是复杂度高；第二种方法是使用贪心技术不断调整感染树使得接近源头的节点有较小的度，而远离源头的节点度较大，此算法为反向贪心（Reverse Greedy, RG）算法，时间复杂度为 $O(n^3)$，空间复杂度为 $O(n+m)$，n 是传播网络中的节点数量，m 是连边数量。

实验结果显示，RG 算法在电力网络、小世界网络和 Facebook 社交网络上推断传播源的平均错误距离分别为 2.59、1.69 和 1.03。

3.3 动态年龄算法（dynamical age）

3.3.1 基本思想

Fioriti 等人[49] 在多源头传播源定位问题中，提出传播源是感染图中节点年龄（node age）最大的节点，此处节点年龄指节点加入网络的时间。

Zhu 等人[57] 认为，当接触网络是经过缓慢的成长过程发展起来时（例如无标度网络的 BA 生成模型[58] ），该网络的邻接矩阵（或者拉普拉斯矩阵）的特征值谱与网络节点的年龄具有较强的相关性。但是，如果接触网络没有这样一个生长过程（例如随机网络的 ER 模型[58] ），则无法使用特征值方法来计算网络节点的年龄。

3.3.2 典型算法

Fioriti 等人[49] 根据 Restrepo 等人在文献[59] 中介绍的节点动态重要性（node dynamical importance），提出了计算感染图中每个节点的动态年龄（dynamical age）的溯源算法 DA。

该算法计算将每个节点从网络中去除以后，网络邻接矩阵最大特征值的下降量；去除某个节点导致的下降量越大，则该节点的年龄越大，该节点也就越有可能是传播源或者接近传播源。详细地说，该算法定义第 i 个节点的 DA 为[49]：

$$DA_i = \frac{\lambda_m - \lambda_m^{new}}{\lambda_m} \tag{7}$$

其中 λ_m 代表当前感染图邻接矩阵的最大特征值，λ_m^{new} 代表将节点 i 从感染图中去掉以后新邻接矩阵的最大特征值。如果已知有 k 个源，那么 DA 最大的前 k 个节点就是此算法估计出的传播源。当使用 QR 分解[60-62] 计算矩阵特征值时，DA 算法的时间复杂度为 $O(n_I^3)$，空间复杂度为 $O(n_I^2)$。

对于特殊的网络，例如使用 BA 模型构建的无

标度网络，由于网络生成过程中每个节点优先选择连边度较大的节点，因此度越大的节点动态年龄也就越大，所以在 BA 网络中可以直接寻找度数最高的前 k 个节点作为传播源。

在几个真实网络上的模拟实验表明，DA 算法能正确找到传播源或者错误距离只有几跳。

3.4 无偏中介中心性

3.4.1 基本思想

Comin 等人[43] 在单源头的溯源问题中尝试了平均测地距离、紧密度中心性、中介中心性、特征值中心性等四种节点中心性指标，但使用它们的溯源算法准确率并不高，因此提出了无偏的中介中心性（unbiased betweenness centrality）。

Comin 等人设计无偏中介中心性指标的基本思想是：感染图与原始接触网络相比，节点的度作为一个局部统计量更有可能保持不变，因此需要在原始的中心性统计量中去除度值造成的偏差。在雪崩传播模型下的实验表明无偏的中介中心性表现最好。

3.4.2 典型算法

笔者将 Comin 等人提出的算法称为无偏中介（Unbiased Betweenness，UB）算法。此算法比较简单，只要计算感染图中所有节点的无偏中介中心性即可。该中心性指标定义为[43]：

$$\hat{B}_i = \frac{B_i}{(k_i)^r} \tag{8}$$

其中 k_i 表示节点的度，r 可以选择一个合适的值，B_i 是原始定义的中介中心性，即任意节点对之间最短路径经过节点 i 的比例。

本算法的复杂度为 $O(nm)$ [63]，其中 n 是感染图的节点数，m 是感染图的连边数。

3.4.3 推广到快照只覆盖部分信息的多源问题

Zang 等人将 UB 算法推广到了 SIR 传播模型上、只观察到一部分节点状态的多源问题情形[41]。笔者将他们提出的算法称为 AUB（Advanced Unbiased Betweenness）算法。

由于观察到的状态信息不完整，如有的感染节点没有被观测到，或者不能区分处于 S 状态和 R 状态的健康节点，所以 Zang 等人首先使用基于评分的逆向传播算法，根据网络拓扑结构计算所有节点的评分，将评分高于某个设定阈值的视为已被感染（处于 I 状态或者 R 状态），这样就得到了扩展感染图。

在获得扩展感染图以后，考虑到各个传播源造成的感染子图内部连边紧密，而子图之间的连边稀疏，因此 Zang 使用模块化指标[64] 对扩展感染图进行划分，认为划分后的每个社区分别由不同的传播源传播疾病形成。

基于此，Zang 等人给出了一个启发式的算法来估算传播源的个数 k：逐步增加传播源个数 k 并使用 Newman 给出的谱计算方法[65] 计算将扩展感染图划分成 k 个社区时的模块度值，当模块度值不再显著增加时停止增加 k。这时社区结构中不同的社区即代表不同的感染子图。

确定各个感染子图后，多源问题便转化成了多个单源问题，然后分别在各个感染子图中寻找无偏中介中心性指标最高的节点作为传播源。

3.4.4 实验结果

在正则树、ER 随机网络、BA 无标度网络上，AUB 算法推断出的传播源的错误距离分别为 0.12、1.81、1.93；在三个真实的网络（Facebook、Epinions、Vote）上，相应的结果分别为 2.50、2.37、1.65。

3.5 总结

本节中的溯源算法基于传染源在感染图中的中心假设，提出 4 种中心性指标，分别是传播中心性、Jordan 中心性、动态年龄中心性、无偏中介中心性。

简要地说，传播中心性指标假设从传播源出发感染其他节点所有可能的顺序最多；Jordan 中心性指标假设传播源与病毒在感染节点上最可能的传播路径相关；动态年龄指标依据传播源是感染图中节点年龄最大的节点，即最早加入感染图的节点；无偏中介中心性指标认为需要在原始的中介中心性指标上去除节点度值造成的偏差。

传播中心性、动态年龄中心性、无偏中介中心性这三种指标通常使用于完整的传播快照（或者感染图），而 Jordan 中心性则可用于传播快照中没有包含所有被感染节点的情况。此外，传播中心性和 Jordan 中心性都是基于树形网络结构提出，并通过宽度优先树搜索推广到一般网络上，所以它们在树形网络上效果较好，但在一般网络上效果

较差；动态年龄中心性和无偏中介中心性则直接在一般网络上提出。

这四种指标也可以推广到多个传播源的情况，但对传播中心性和Jordan中心性，研究人员特别设计了一些算法，能够更好地将这两种指标应用在多源问题中。

4 基于置信传播的溯源算法

公式(1)说明在理想情况下，传播源的估计 $\hat{s}$ 应该是使得 $P(\boldsymbol{O}|s)$或 $P(s|\boldsymbol{O})$都达到最大的 s。这两种似然都难以计算，复杂度甚至达到#P-hard[5, 51, 52]，因此一部分研究人员将似然转换成其他相对容易计算的统计量，如 Shah 等人将计算给定传播源时的状态信息似然 $P(\boldsymbol{O}|s)$换成了计算所有被感染节点的合法感染顺序计数[4, 5]，而 Zhu 等人将计算后验概率 $P(s|\boldsymbol{O})$最大的传播源换成了计算最可能的传播路径，然后根据计算出的传播路径来反推传播源（忽略了其他所有可能的传播路径）[42]。上述策略虽然降低了计算难度，但只适用于简单的网络拓扑结构（通常要求网络是树形结构），在一般网络上准确性大大降低。

为了提高准确性，一部分研究人员依旧希望能直接计算 $P(\boldsymbol{O}|s)$和 $P(s|\boldsymbol{O})$。为了减轻计算量，研究人员使用置信传播（belief propagation, BP）来辅助计算。最大化似然函数中的核心步骤就是计算各种边际概率，而置信传播通过因子图（factor graph）上的消息传递来做概率图模型（如贝叶斯网络和马尔可夫随机场）上的概率推断，从而计算未观察节点上的边际概率分布[40, 66]。即便是在一般网络上，置信传播也是比较好的近似算法[67]。

4.1 动态消息传递算法（dynamic message-passing, DMP）

4.1.1 基本思想

Lokhov 等人在 SIR 疾病模型的单源溯源问题上提出了依赖于动态消息传递方程的 DMP 算法[40]。该算法首先由 SIR 模型和网络拓扑推导出任意时刻 t+1 时每个节点 i 处于各个状态的概率 $P_S^i(t+1)$、$P_R^i(t+1)$、$P_I^i(t+1)$ 与前一时刻相应的变量和其他中间变量的递推关系[40]：

$$P_S^i(t+1)=P_S^i(0)\prod_{k\in\partial i}\theta^{k\to i}(t+1) \tag{9}$$

$$P_R^i(t+1)=P_R^i(t)+\mu_i P_I^i(t) \tag{10}$$

$$P_I^i(t+1)=1-P_S^i(t+1)-P_R^i(t+1) \tag{11}$$

$$P_S^{i\to j}(t+1)=P_S^i(0)\prod_{k\in\partial i\setminus j}\theta^{k\to i}(t+1) \tag{12}$$

$$\theta^{k\to i}(t+1)-\theta^{k\to i}(t)=-\lambda_{ki}\phi^{k\to i}(t) \tag{13}$$

$$\begin{aligned}\phi^{k\to i}(t)&=(1-\lambda_{ki})(1-\mu_k)\phi^{k\to i}(t-1)\\&-[P_S^{k\to i}(t)-P_S^{k\to i}(t-1)]\end{aligned} \tag{14}$$

其中 μ_i 是节点 i 在每个时间步从 I 状态恢复到 R 状态的概率，λ_{ki} 是疾病每个时间步在节点 k 和节点 i 的连边上的传播概率，∂i 是节点 i 的邻居节点集合，$P_S^{i\to j}(t+1)$、$\theta^{k\to i}(t+1)$、$\phi^{k\to i}(t)$ 是三个中间变量。

一旦确定了传播源，便确定了上述变量在初始时刻的值，继而使用上述递推关系式就可以计算出网络中每个节点在任意时刻处于任意状态的概率。

在上述递推关系的基础上，加以状态信息中各个节点的独立性假设，便可计算状态信息似然的近似解[40]：

$$P(\boldsymbol{O}\mid s)=\prod_k P_S^k(t)\prod_l P_I^l(t)\prod_m P_R^m(t) \tag{15}$$

其中 t 是给出所有节点状态信息的时间。传播源即可推断为使得 $P(\boldsymbol{O}|s)$最大的节点 s。

4.1.2 典型算法

根据 DMP 方程和状态信息似然的计算公式，可以直接计算出每个节点作为传播源时得到所观察信息的概率，推断的传播源即为使得似然最大的源节点。

DMP 算法需要已知疾病传播模型 SIR 的参数，以及传播时间 t。当传播时间 t 未知时，可对不同的 t 计算配分函数（partition function）$Z(t)=\sum_s P(\boldsymbol{O}\mid s)$。使得此配分函数最大的 t，即为所求的传播时间。求解传播时间的时间复杂度为 $O(tMN)$，空间复杂度为 $O(tN)$，其中 t 是传播时间的上界，M 是网络中边的数量，N 是网络中节点

的数量。

4.1.3 实验结果

实验结果表明：使用 DMP 算法，对真实的传播源的似然通常高于其他节点的似然。

DMP 算法的优点在于：

（1）算法准确性比现有的传播中心和 Jordan 中心都要高(在某些疾病参数上弱于Jordan中心)；

（2）即使在快照中只覆盖部分信息，此算法依然有效。

其不足之处在于：

（1）当网络中含有环时，DMP 算法中的递推关系式不再正确；

（2）节点状态的独立性假设导致 DMP 算法给出的解并不是严格意义上的极大似然解。

4.2 基于因子图的置信传播算法（belief propagation, BP）

4.2.1 基本思想

Altarelli 等人[38] 基于置信传播提出了 SIR 传播模型上另一个传播源定位算法 BP。

设在 SIR 疾病传播模型上，所有节点从进入 I 状态到进入 R 状态之间的时间（也称为恢复时间，recovery time）为 $\boldsymbol{g}$，所有连边上的疾病传播时间（transmission delay）为 $\boldsymbol{s}$，所有节点被感染的时刻为 $\boldsymbol{t}$，于是有[38]

$$p(\boldsymbol{t},\boldsymbol{g}\mid\boldsymbol{x}^0)=\sum_{\boldsymbol{s}}p(\boldsymbol{t}\mid\boldsymbol{x}^0,\boldsymbol{g},\boldsymbol{s})p(\boldsymbol{s}\mid\boldsymbol{g})p(\boldsymbol{g}) \quad (16)$$

由于在 T 时刻所有节点的状态 $\boldsymbol{x}^T$ 是已知量，所以[38]

$$\begin{aligned}p(\boldsymbol{x}^0\mid\boldsymbol{x}^T)&\propto\sum_{\boldsymbol{t},\boldsymbol{g}}p(\boldsymbol{x}^T\mid\boldsymbol{t},\boldsymbol{g})p(\boldsymbol{t},\boldsymbol{g}\mid\boldsymbol{x}^0)p(\boldsymbol{x}^0)\\&=\sum_{\boldsymbol{t},\boldsymbol{g},\boldsymbol{s}}\prod_{i,j}\omega_{ij}\prod_i\phi_iG_i\gamma_i\zeta_i\end{aligned} \quad (17)$$

其中 G_i 是节点 i 的恢复时间 g_i 的概率分布，γ_i 是节点 i 在初始时刻处于 I 状态的概率，ζ_i 是节点 i 在 T 时刻的状态是否与它的被感染时刻和恢复时间相符合的指示变量（若符合则为 1，否则为 0），ω_{ij} 为节点 i 和节点 j 之间的疾病传播时间 s_{ij} 的概率分布，ϕ_i 为指示变量，若满足以下两个条件之一则为 1，否则为 0：

条件一：节点 i 在初始时刻处于 I 状态且其被感染的时刻 t_i=0；

条件二：节点 i 在初始时刻处于 S 状态且 $t_i=\min_{j\in\partial_i}\{t_j+s_{ji}\}$。

由于很难从此公式计算出边际概率分布 $p(\boldsymbol{x}^0\mid\boldsymbol{x}^T)$，Altarelli 等人借用置信传播方法，用因子图(factor graph)表示出各个变量的相关关系，引入几个新的变量，得到如下公式[38]

$$p(\boldsymbol{x}^0\mid\boldsymbol{x}^T)\propto\sum_{\boldsymbol{t},\boldsymbol{t}',\boldsymbol{g}}Q(\boldsymbol{g},\boldsymbol{t},\boldsymbol{t}',\boldsymbol{x}^0) \quad (18)$$

其中 $Q=\frac{1}{Z}\prod_{i<j}\phi_{ij}\prod_i\varphi_iG_i\gamma_i\zeta_i$，

$\phi_{ij}=\omega_{ij}\omega_{ji}$，$\varphi_i$ 为节点 i 的被感染时刻和恢复时间是否与其邻居节点相符合的指示变量（符合为 1，不符合为 0），Z 为归一化因子。

由于因子图的拓扑结构反映出原始传播网络的拓扑结构，因此这一方法在无环网络（也就是树）上可以得到后验边际分布的精确解。注意到此公式中 $\boldsymbol{x}^0$ 与传播源 s 等价，$\boldsymbol{x}^T$ 与状态信息 $\boldsymbol{O}$ 等价，因此 $p(\boldsymbol{x}^0\mid\boldsymbol{x}^T)$ 与传播源的后验概率 $p(s|\boldsymbol{O})$ 等价。

4.2.2 典型算法

BP 算法的基本思想是：基于因子图所得的传播源后验概率，真实传播源应能在所有节点中取得最大值。计算后验概率时每次 BP 迭代过程的时间复杂度为 $O(TG^2M)$，其中 G 是最大允许的恢复时间，M 是网络中连边的数量。

此算法还可以应用于快照只覆盖部分节点的情形、传播时间未知的情形、多源头的情形、传播网络动态变化的情形。

4.2.3 实验结果

模拟实验显示，BP 算法在随机网络上的准确

率可达60%。即便快照中缺失了40%的节点信息，BP 算法在 RRG 网络上也可达到 80%的准确率。

4.3 总结

本小节中溯源算法都是基于贝叶斯理论中的概率计算提出，其中 DMP 的计算基于给定快照信息时传播源的后验概率 $P(s|\boldsymbol{O})$最大，而 BP 的计算基于疾病从传播源开始传播得到给定快照信息的似然 $P(\boldsymbol{O}|s)$最大。事实上，由贝叶斯理论中的全概率公式可知，极大化 $P(s|\boldsymbol{O})$和极大化 $P(\boldsymbol{O}|s)$是等价的。

在计算后验或似然的过程中，核心步骤是如何计算各种边际概率。为减轻计算量，提出 DMP 和 BP 两种算法的研究人员都使用置信传播，通过因子图上的消息传递来计算目标后验或目标似然。

这两种算法都与网络结构无关，理论上可以推广到动态网络上。另外，由于这两种算法都是按照疾病传播过程一步一步计算每个节点的概率，因此它们在传播时间已知的情形下可以达到更好的效果；当传播时间未知时，只能计算不同传播时间下的后验或似然。另外，它们也都对快照信息的完整性没有要求，可以在快照信息只包含部分节点状态的情况下达到溯源的目的。

5 基于蒙特卡洛（Monte-Carlo）模拟的溯源算法

除了上述对似然函数做近似计算以及直接寻找传染源的统计量两种策略之外，第三种策略直接采用仿真模拟，从而避免了计算似然函数的困难。

5.1 基本思想和典型算法

Nino Antulov-Fantulin 等人[39] 在单源头、SIR 传播模型、已知传播过程中 T 时刻所有节点的状态的溯源问题中提出了基于蒙特卡洛模拟的定位算法。

这种算法原理非常简单，即通过直接模拟疾病传播的方法来估计 $P(\boldsymbol{O}|s)$：对每个可能的传播源 θ（也就是在观察到的状态中处于 I 或者 R 的那些节点），重复做 n 次独立的、传播时间为 T 的蒙特卡洛模拟传播，然后计算 T 时刻所有节点状态与观察到的状态一致的模拟次数 n_θ，使得 n_θ 值最大的 θ 就是得到的源点定位结果。

为了加快速度，可以使用剪枝方法减少不必要的模拟过程，即一旦发现某个在观察时处于 S 状态的节点在模拟过程中进入 I 状态，那么本次模拟结果必然与观察到的状态信息不符，所以可以提前终止本次模拟。本方法显而易见需要进行大量模拟，所以耗时很长，并不实用。

为了解决这一问题，Nino Antulov-Fantulin 等人[39] 提出了基于蒙特卡洛模拟的软边界估计方法（Soft-Margin estimator）。该方法同样需要对每个可能的源头 θ 重复做 n 次独立的、传播时间为 T 的蒙特卡洛模拟传播，但是与之前的算法不同，这里每次模拟过程必须达到 T 时间才可停止。模拟结束后，按如下公式计算节点 θ 的软边界估计[39]：

$$P(\boldsymbol{R}=\boldsymbol{r}_*\mid\Theta=\theta)=\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{-[\varphi(\boldsymbol{r}_*,\boldsymbol{r}_{\theta,i})-1]^2}{a^2}\right) \tag{19}$$

其中 $\boldsymbol{r}_*$为观察信息中所有节点的状态（I 和 R 状态的节点在 $\boldsymbol{r}_*$里的相应分量为 1，S 状态的节点为 0），$\boldsymbol{r}_{\theta,i}$ 是将节点 θ 作为传染源做第 i 次传播模拟后所有节点的状态，函数 $\varphi:(R^N\times R^N)\to[0,1]$ 计算两个向量 $\boldsymbol{r}_1$ 和 $\boldsymbol{r}_2$ 的 Jaccard 相似度，即两个向量中都是 1 的元素占都不是 0 的元素数量的比重，a 是软边界的参数，具体取值可以参考文献[39] 。

笔者在本文中将此算法称为 MCSM（Monte-Carlo with Soft-Margin）算法，其时间复杂度为 $O(|\boldsymbol{r}_*|\times n\times\overline{RT}_M)$，空间复杂度为 $O(n+l|\boldsymbol{r}_*|)$，其中 $|\boldsymbol{r}_*|$ 为候选源点的数量，也就是观察信息中处于 I 和 R 状态节点的总数，l 为当传播源固定时，模拟疾病传播的次数，$\overline{RT}_M$ 为单次模拟传播的时间复杂度，与具体模拟算法有关（如可以使用文献[68, 69] 提出的疾病传播模拟快速算法）。

5.2 实验结果

在网格状网络上的模拟结果显示该算法比 Jordan 估计器（Jordan estimator）[42] 和 DMP 估计器（dynamic message-passing estimator）[40] 有更高的准确率。在巴西真实的性关系网络[70] 上做的传播模拟结果显示该算法能以 60%的准确率找到真实的传播源或者传播源的邻居。

6 基于最小描述长度的溯源算法

6.1 基本思想

Prakash 等人在多源头、SI 传播模型的溯源问题中，提出了基于最小描述长度的 NetSleuth 算法[27, 28]。

该算法的工作原理是，给定疾病传播模型、疾病传播网络和一段时间后的传播结果，真正的源头应该使得描述源头和传播结果的长度最短。根据文献[71]，描述长度可以分成两部分，一部分是模型的描述长度，另一部分是基于模型所得数据的描述长度，即[27, 28]

$$L(G_I, S, \boldsymbol{R}) = L(S) + L(\boldsymbol{R} \mid S) \qquad (20)$$

其中S是传播源，$L(S)$是模型的描述长度；$\boldsymbol{R}|S$是已知的传播结果，$L(\boldsymbol{R}|S)$是数据的描述长度。Prakash 等人使用整数的 Universal 编码长度[72] 和基于概率似然的香农熵[73] 计算这两种描述长度，从而将溯源问题转化成了求解最短描述长度的最优化问题。

6.2 典型算法

Prakash 等人基于以上分析提出了无需源头个数先验知识的传播源定位算法 NetSleuth[27, 28]。该算法首先假设传播源只有 1 个，按照单源头问题求得该传播源（必然处于 I 状态），计算此时的描述长度，然后将刚刚求得的传播源改为 S 状态，重新按照单源头问题求解第 2 个传播源（必然处于 I 状态），然后对找到的 2 个传播源计算新的描述长度，依此不断迭代直到描述长度不再减小。

在解决单源头子问题时，Prakash 等人试图寻找使得 $P(\boldsymbol{O}|s)$最大的传播源。由于此概率难以计算，Prakash 等人将 t 时刻每个节点被感染概率 $\boldsymbol{P}(t)$ 的上界作为 $\boldsymbol{P}(t)$ 的近似，得到 $\boldsymbol{P}_I(t+1) \approx \sigma\lambda_1^t \boldsymbol{u}_1 \boldsymbol{u}_1^T \boldsymbol{P}_I(0)$，这里 $\boldsymbol{P}_I(t)$ 是 t 时刻每个被感染节点处于被感染状态的概率，λ_1 和 $\boldsymbol{u}_1$ 分别是感染图的拉普拉斯矩阵 $\boldsymbol{L}_A$ 的最小特征值和相应的特征向量。因此，单源头子问题的解为 $s^* = \arg\max_s \boldsymbol{u}_1(s)$。

NetSleuth 算法不需要事先给出源点的数量，所以它是自适应的源点定位算法。由于可以使用 Lanczos 方法去计算拉普拉斯矩阵的最小特征值，所以对稀疏图计算拉普拉斯矩阵特征向量的步骤只需要 $O(E_I)$的时间复杂度，NetSleuth 算法的总时间复杂度为 $O(k(E_I+E_F+V_I))$，这里 k 是传播源个数的上界，E_I 是感染图里的连边数量，E_F 是感染图边界的连边数量，V_I是感染图里的节点数量。

6.3 实验结果

模拟实验表明，NetSleuth 算法可以在真实的路由器网络[74]上准确发现传播源的数量，并且在其找到的传播源上进行多次模拟传播得到的感染图与从真实传播源进行多次传播得到的感染图相当。

7 总结与展望

7.1 算法比较

在提出各种溯源算法的原始文献中，各自使用了不同的传播模型、传播参数、传播网络、源头数量等条件，从而导致无法直接评估各种算法性能。

为了了解这些算法在不同评价标准和适用场景下的优劣，我们重新实现了典型算法，并在相同的实验平台和多类数据集上，比较了算法性能。测试的典型算法包括基于传播中心性的 SSE、SSE-BFS，基于 Jordan 中心性的 SJC、JCE、RG，基于动态年龄的 DA，基于无偏中介中心性的 UB，基于置信传播的 DMP，基于蒙特卡洛的 MCSM，以及基于最小描述长度的 NetSleuth 算法。

算法全部使用 C++11 实现，并在 CPU 为 16 核

1.4GHz、内存为 64GB、操作系统为 Ubuntu 14.04.5 的服务器上运行。

在实验中，为公平起见，我们使用完全相同的传播网络、传播模型和问题场景。具体的说，我们首先从传播网络中随机选择一个节点作为初始感染节点，然后在该网络上使用 SIR 模型传播疾病，将 5 个时间步后的传播结果作为完整快照信息，依次使用每个溯源算法来推断疾病的单个源头。

在推断传播源时，我们均假设传播网络、疾病模型、传播参数为已知条件，但传播时间未知。在所测试的算法中，仅有 DMP 和 MCSM 两种算法与传播时间相关。为方便起见，我们在实验中限制这两种算法最多只考虑 50 个时间步的传播时间（实际传播时间的 10 倍）。为减少随机性造成的偏差，我们做了 100 次实验，每次实验都随机选择初始感染点作为真实的传播源。当快照中被感染节点数不超过 2 个时，各种溯源算法的比较已无明显意义，因此我们在实验时只使用不少于 10 个节点被感染的传播结果作为测试溯源算法时使用的快照信息。

不同的溯源算法在使用不同的传播网络和不同的传播参数时，性能可能有所不同。为了比较传播网络和传播参数对溯源算法性能的影响，我们在不同的传播网络和不同的传播参数上测试了这些溯源算法，描述如下：

（1）在传播网络方面，我们使用两类网络，第一类是使用理想模型构造的网络，分别为从随机网络构造的随机树、从 BA 模型[58] 构造的无标度网络、从 WS 模型[58] 构造的小世界网络，以及从 ER 模型[58] 构造的随机网络；第二类为从真实网络构造的网络，分别为 Stehle 等人于 2009 年 10 月 1 日在法国一所小学收集的接触网络[75] 和 Huang 等人于 2011 年 10 月 31 日在中国的华南农业大学收集的接触网络[76]（为方便起见，我们丢弃接触网络中时长不到 10 分钟的接触、消除边权信息、再取最大连通分量）。表 1 列出了这些网络的规模和来源。

（2）在传播参数方面，我们将 SIR 疾病模型中 I 状态节点恢复健康变成 R 状态的概率设定为 0.4，但使用 0.25~0.55 范围内不同的传播概率（S 状态节点被邻居感染变成 I 状态的概率）。

表 1 实验网络的规模和来源

网络类型	网络拓扑	节点数	连边数	来源
理想网络	随机树	200	199	随机网络生成树
	无标度网络	200	591	BA 模型
	小世界网络	200	600	WS 模型
	随机网络	200	600	ER 模型
真实网络	法国小学网络	207	503	文献[75]
	中国大学网络	145	780	文献[76]

表 2 不同网络拓扑结构上溯源算法的准确率（传播概率 0.3）

网络拓扑	SSE	SSE-BFS	SJC	JCE	RG	DA	UB	DMP	MCSM	NetSleuth
随机树	0.17	0.22	0.08	0.20	0.10	0.22	0.19	**0.34**	0.15	0.04
随机网络	0.07	0.01	-	-	0.16	0.03	0.09	**0.34**	0.08	0.01
无标度网络	0.02	0.01	-	-	0.01	0.03	0.03	**0.13**	0.02	0.03
小世界网络	0.08	0.02	-	-	0.11	0.08	0.06	**0.21**	0.13	0.14
法国小学网络	0.09	0.04	-	-	0.12	0.08	0.07	**0.22**	0.06	0.08
中国大学网络	0.02	0.02	-	-	0.04	0.00	0.03	**0.12**	0.00	0.05

表 3 不同网络拓扑结构上溯源算法的平均错误距离（传播概率 0.3）

网络拓扑	SSE	SSE-BFS	SJC	JCE	RG	DA	UB	DMP	MCSM	NetSleuth
随机树	1.37	1.09	1.9	1.08	1.91	1.08	1.20	**0.84**	1.53	2.63
随机网络	2.55	2.85	-	-	1.78	1.87	1.78	**1.01**	2.22	3.08
无标度网络	2.79	2.83	-	-	2.39	1.69	**1.66**	1.69	2.36	2.64
小世界网络	1.62	1.90	-	-	1.56	1.03	1.68	**0.86**	1.26	1.64

法国小学网络	1.85	2.16	-	-	1.80	1.50	1.55	**1.08**	1.77	2.70
中国大学网络	2.15	2.60	-	-	2.31	1.86	2.05	**1.32**	2.25	2.98

表 4 不同网络拓扑结构上溯源算法的运行时间（传播概率 0. 3，单位：毫秒）

网络拓扑	SSE	SSE-BFS	SJC	JCE	RG	DA	UB	DMP	MCSM	NetSleuth
随机树	0.5853	1.5846	0.84947	0.51555	203.331	1.29019	**0.50086**	7060.38	2171.05	23.5627
随机网络	30.2171	24.0489	-	-	545.628	290.938	**1.46928**	19097.7	21904	27.9201
无标度网络	116.575	88.3619	-	-	973.394	2752.8	**3.66113**	21853.1	33591.1	43.4706
小世界网络	4.69234	3.75286	-	-	329.316	5.66075	**0.70913**	19283	13585.1	24.4425
法国小学网络	7.4318	5.85183	-	-	345.169	16.1839	**0.80133**	16263.3	14997.4	25.6591
中国大学网络	121.564	82.6616	-	-	691.109	1370.52	**4.11573**	22993	31463.2	27.6628

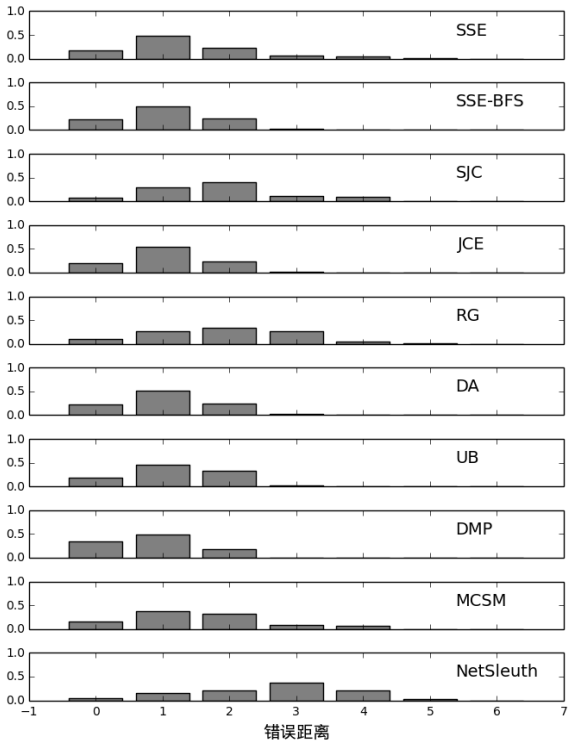

图 6 随机树上溯源算法的错误距离分布（传播概率 0.3）

7.1.1 不同溯源算法在不同传播网络上的性能

我们在上述 6 种网络拓扑结构上测试了各种溯源算法，将测试结果（包括算法的准确率、平均错误距离和平均运行时间）分别整理在表 2、表 3 和表 4 中，并在表中将最优的性能值加粗表示。由于 SJC 和 JCE 两种算法仅能在树型网络上运行，因此表格中略去它们在随机树以外网络拓扑结构上的实验结果。为方便起见，我们仅给出传播概率为 0.3 的结果，其他传播概率下的测试结果与之类似。各算法性能总结如下：

（1）准确率

从表 2 可看出，溯源算法 DMP 在实验使用的所有网络拓扑结构上都具有最大的准确率，且具体数值随网络拓扑在 0.12 与 0.34 间不等。其他溯源算法在不同的网络拓扑上也有不同的表现，如 DA 在随机树上的准确率高达 0.22，在所有算法中排名第二，但在中国大学网络上则低至 0.00。

（2）错误距离

表 3 的平均错误距离与表 2 的准确率基本一致，准确率较高的算法具有较低的平均错误距离，但也不完全一致。以 DMP 算法为例，它在所有溯源算法中具有最高的准确率，因此也基本上具有最低的平均错误距离；但是，当网络拓扑结构采用无标度网络时，DMP 算法的平均错误距离没能达到最小值，而以 0.03 的差距位于 UB 算法之下。与准确率类似，表 3 中数据同样表明，每种溯源算法的平均错误距离都随着网络拓扑结构的变化而变化，但是变化程度不如准确率剧烈。

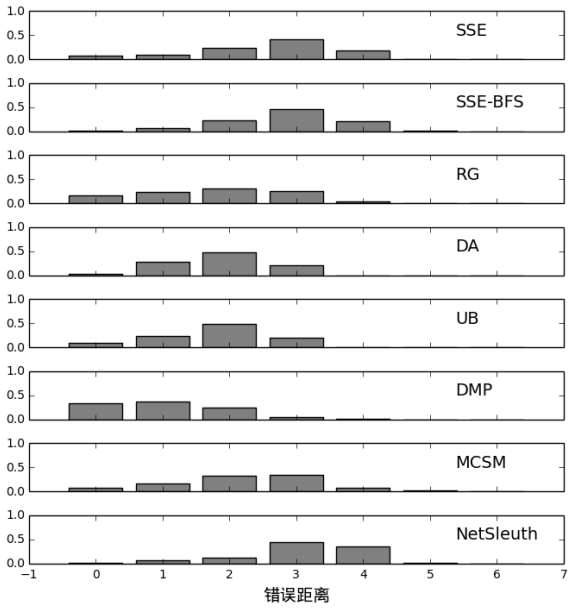

图 7 随机网络上溯源算法的错误距离分布（传播概率 0.3）

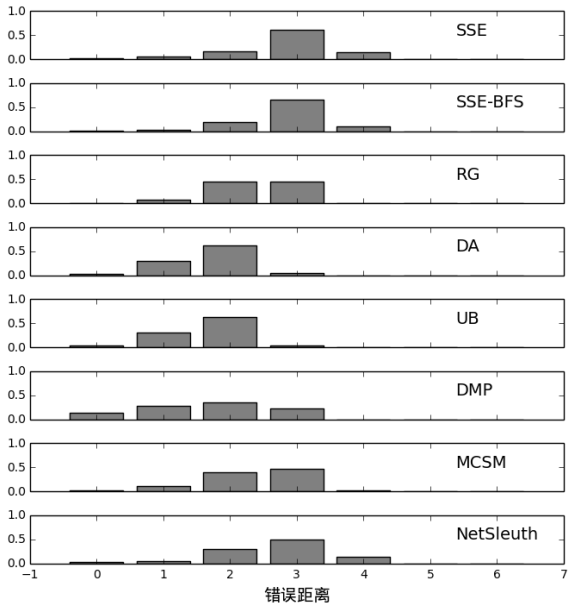

图 8 无标度网络上溯源算法的错误距离分布(传播概率 0.3）

图 6~11 列出了在每种网络拓扑结构上所有溯源算法错误距离的分布情况。图 6 显示，在随机树上，大部分溯源算法的错误距离相差不大，大多集中在 1 附近，仅有 SJC 和 RG 集中在 2 附近、NetSleuth 集中在 3 附近。从错误距离分布中也可看出，DMP 算法有超过 30%的几率达到零错误距离，与表 2 中 0.34 的准确率一致。小世界网络上的溯源结果（图 9）也显示了相似的错误距离：UB 算法集中在 2，其他算法集中在 1。然而，当网络拓扑结构为随机网络（图 7）时，各种溯源算法的错误距离分布有了明显的不同：表现最优的 DMP 算法集中在 0 和 1，RG、DA、UB 集中在 2，SSE、SSE-BFS、MCSM、NetSleuth 集中在 3。这些溯源算法在无标度网络（图 8）、法国小学网络（图 10）和中国大学网络（图 11）上也表现出了错误距离的差异性。

（3）运行速度

对每个算法来说，运行时间至关重要。为此，我们记录了实验中每个算法的运行时间，并将它们的平均值记入表 4。由表 4 可见，在所有网络上，UB 的运行时间都达到最短，即便在运行时间最长的中国大学网络上也不到 5 毫秒。与此相对，准确率和平均错误距离达到最优的 DMP 算法，单次运行时间则长达 7 秒以上，尤其是在中国大学网络上更是长达 23 秒。另外，溯源算法的运行时间对网络拓扑结构极为敏感。所有算法在随机树上的运行时间都极短，但在无标度网络和中国大学网络上的运行时间极长。以 DA 算法为例，在随机树上只需要平均 1.3 毫秒的时间，但在无标度网络上运行时间则长达 2.8 秒，约为前者的 2000 倍。

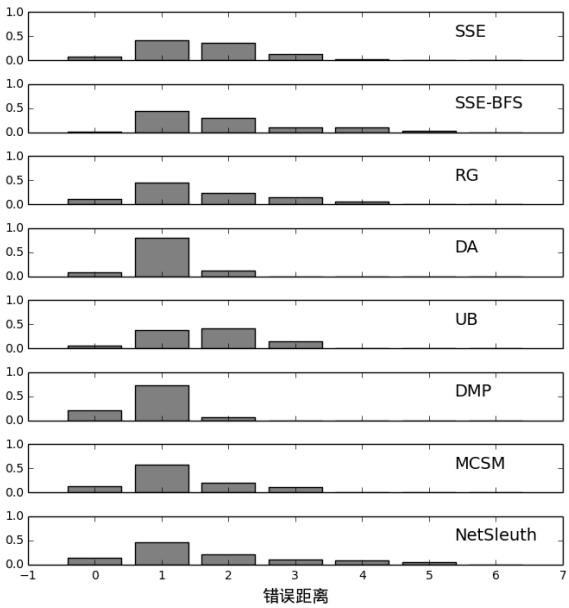

图 9 小世界网络上溯源算法的错误距离分布(传播概率 0.3）

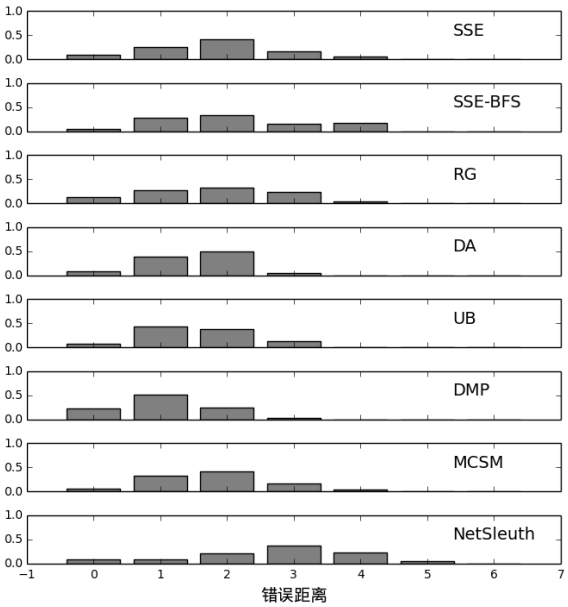

图 10 法国小学网络上溯源算法的错误距离分布（传播概率 0.3）

7.1.2 不同溯源算法在不同传播概率下的性能

为比较不同传播概率下这些溯源算法的性能，我们列出溯源算法在随机树（表 5~7）和中国大学网络（表 8~10）这两种典型网络上的实验结果。为方便起见，所有算法中的最优性能值均加粗表示。

从表 5~7 中可以看出：

（1）虽然传播概率不同，但所有溯源算法在相同的网络拓扑结构上都有着相似的性能排名。

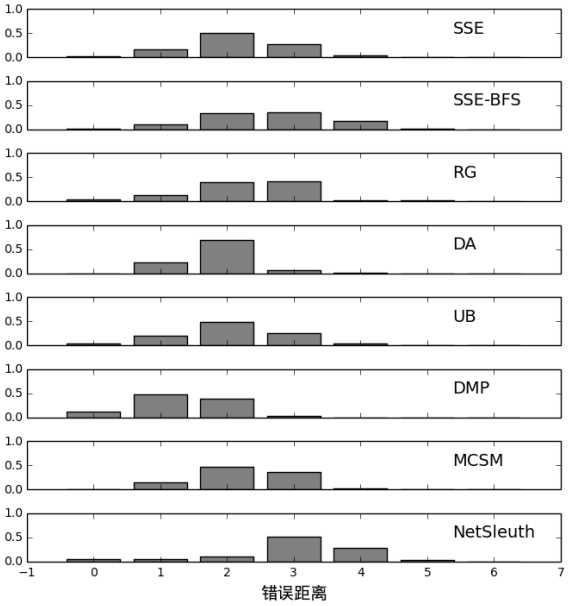

图 11 中国大学网络上溯源算法的错误距离分布（传播概率 0.3）

（2）不论在何种传播概率生成的快照场景下，DMP 都是准确率最高、平均错误距离最短、平均运行时间最长的算法。与此相反，NetSleuth 的准确率最低，平均错误距离也最大。

（3）需要特别提出的是，JCE 的准确率与平均错误距离虽然不是最优，但也接近最优的 DMP；然而，JCE 的运行时间仅有 0.5 毫秒左右，远远低于 DMP 算法，且比运行时间最短的 UB 算法高出不到 5%。我们可以认为，在树型拓扑结构的网络上，JCE 算法是综合性能最高的算法。

与表 5~7 类似，表 8~10 同样列出了各种算法在中国大学网络上的实验结果。由于中国大学网络不是树型结构，此时无法使用 JCE 算法和 SJC 算法，因此表格中去除了 JCE 和 SJC 这两列。

由表 8 和表 9 可见，在中国大学网络上，DMP 算法的准确率和平均错误距离远远优于其他 7 种算法；与随机树上的情形不同，此时 DMP 算法不再是运行时间最长的算法，但也长达 20 秒以上。相反，UB 算法虽然准确率不高，但运行时间只有 4 毫秒，考虑到 UB 算法的时间主要消耗在计算所有节点的中介中心性，优化空间较大，因此在中国大学网络上 UB 算法具有较大的优势。

事实上，其他网络拓扑结构上的实验结果同样表明，各种溯源算法之间的性能高低与产生快照的传播概率关联不强。因此，我们可以针对不同的网络结构使用不同的溯源算法。我们在 7.1.3 节中给出了分别适用于实验中 6 种网络拓扑结构的溯源算法。

表 5 随机树上不同传播概率下的准确率

传播概率	SSE	SSE-BFS	SJC	JCE	RG	DA	UB	DMP	MCSM	NetSleuth
0.25	0.15	0.17	0.15	0.20	0.14	0.17	0.17	**0.26**	0.13	0.08
0.30	0.17	0.22	0.08	0.20	0.10	0.22	0.19	**0.34**	0.15	0.04
0.35	0.11	0.11	0.03	**0.16**	0.11	0.11	0.10	**0.16**	0.08	0.03
0.40	0.08	0.08	0.10	0.11	0.13	0.11	0.15	**0.23**	0.14	0.06
0.45	0.06	0.06	0.03	0.12	0.14	0.10	0.11	**0.20**	0.17	0.07
0.50	0.02	0.02	0.07	0.13	0.13	0.04	0.14	**0.22**	0.15	0.05
0.55	0.07	0.08	0.03	**0.14**	0.13	0.09	0.13	**0.14**	0.07	0.05

表 6 随机树上不同传播概率下的平均错误距离

传播概率	SSE	SSE-BFS	SJC	JCE	RG	DA	UB	DMP	MCSM	NetSleuth
0.25	1.21	1.11	1.74	1.08	1.94	1.11	1.16	**0.96**	1.57	2.60
0.30	1.37	1.09	1.90	1.08	1.91	1.08	1.20	**0.84**	1.53	2.63
0.35	1.56	1.31	1.99	1.21	1.95	1.30	1.35	**1.13**	1.86	2.81
0.40	1.70	1.42	2.07	1.31	2.00	1.33	1.34	**1.06**	1.74	2.67
0.45	2.02	1.63	2.41	1.31	2.07	1.39	1.52	**1.11**	1.70	2.95
0.50	2.19	1.67	2.26	1.33	2.00	1.60	1.49	**0.97**	1.86	3.08
0.55	2.31	1.99	2.53	1.45	2.05	1.71	1.70	**1.24**	1.99	3.28

表 7 随机树上不同传播概率下的平均运行时间（单位：毫秒）

传播概率	SSE	SSE-BFS	SJC	JCE	RG	DA	UB	DMP	MCSM	NetSleuth
0.25	0.54269	1.24111	0.70851	0.47069	188.197	0.91081	**0.46633**	6616.69	1584.03	22.5307
0.3	0.5853	1.5846	0.84947	0.51555	203.331	1.29019	**0.50086**	7060.38	2171.05	23.5627
0.35	0.60544	1.78069	0.8779	0.51925	209.257	1.39174	**0.51899**	7113.53	2502.86	24.5587
0.4	0.55548	1.66553	0.8212	0.4862	192.776	1.6417	**0.47523**	6590.63	2603.34	22.3042
0.45	0.55339	1.85489	0.87781	0.50484	198.845	1.85753	**0.48945**	6643.33	2967.96	22.8145
0.5	0.62	2.24021	1.02296	0.56012	215.492	2.37753	**0.5316**	7124.65	3601.01	25.5622
0.55	0.60158	2.78659	1.06996	0.55673	205.638	4.20526	**0.52492**	6724.04	4059.09	22.6951

表 8 中国大学网络上不同传播概率下的准确率

传播概率	SSE	SSE-BFS	RG	DA	UB	DMP	MCSM	NetSleuth
0.25	0.01	0.02	0.06	0.01	0.02	**0.17**	0.04	0.02
0.30	0.02	0.02	0.04	0.00	0.03	**0.12**	0.00	0.05
0.35	0.03	0.00	0.01	0.01	0.02	**0.20**	0.02	0.03
0.40	0.01	0.01	0.01	0.00	0.05	**0.18**	0.07	0.01
0.45	0.02	0.01	0.03	0.02	0.00	**0.12**	0.01	0.01
0.50	0.02	0.01	0.00	0.00	0.02	**0.18**	0.00	0.04
0.55	0.02	0.00	0.02	0.01	0.01	**0.12**	0.01	0.00

表 9 中国大学网络上不同传播概率下的平均错误距离

传播概率	SSE	SSE-BFS	RG	DA	UB	DMP	MCSM	NetSleuth

0.25	2.19	2.49	2.09	1.79	2.04	**1.23**	2.21	2.90
0.30	2.15	2.60	2.31	1.86	2.05	**1.32**	2.25	2.98
0.35	2.13	2.70	2.67	1.77	2.34	**1.22**	2.20	3.50
0.40	2.33	2.93	2.59	1.93	2.37	**1.25**	2.16	3.47
0.45	2.36	3.10	2.69	1.97	2.48	**1.31**	2.17	3.69
0.50	2.43	3.18	2.85	2.02	2.60	**1.42**	2.42	3.72
0.55	2.29	3.34	2.86	1.86	2.69	**1.35**	2.29	3.74

表 10 中国大学网络上不同传播概率下的平均运行时间（单位：毫秒）

传播概率	SSE	SSE-BFS	RG	DA	UB	DMP	MCSM	NetSleuth
0.25	91.2463	62.263	573.199	905.914	**3.18784**	20731.6	27746.6	21.7595
0.30	121.564	82.6616	691.109	1370.52	**4.11573**	22993	31463.2	27.6628
0.35	129.865	88.7353	694.905	1669.43	**4.26211**	21704	29458.1	27.5541
0.40	141.424	95.4117	741.391	2363.58	**4.562**	23415.1	30136.7	33.6045
0.45	140.907	95.7367	717.625	2555.27	**4.47823**	21651.8	28207.2	33.8361
0.50	155.227	104.358	771.061	2861.6	**4.79012**	23001.6	29181.5	37.5713
0.55	142.142	97.3679	712.072	3250.29	**4.47527**	20720.5	26223.3	38.457

7.1.3 不同传播网络上的溯源算法选择

综合考虑不同传播网络上各个溯源算法的准确率（表 2）、平均错误距离（表 3）和平均运行时间（表 4），我们按照性能优先、兼顾性能和速度、速度优先三种组别，在各种网络拓扑结构上推荐了相应的溯源算法。

具体的说，性能优先意味着选择准确率和错误距离最优的溯源算法；兼顾性能和速度意味着在保留一些性能的基础上，选择运行较快的算法；速度优先表示优先选择耗时最短的算法，其次才在考虑耗时相当的算法中选择性能最优的算法。

在表 11 中我们列出了这三种组别下各自适用的溯源算法，同时也给出了这些算法在实验中的准确率、错误距离和运行耗时，以供参考。在随机树、无标度网络、法国小学网络和中国大学网络上，速度优先组别下的算法已达到兼顾性能和速度的效果，因此没有提供相应的算法。

实验结果表明：

（1）在参与实验的所有网络上，性能优先组的算法都是 DMP，它的准确率远远优于速度优先组的算法，错误距离也只在无标度网络上弱于速度优先组，但是 DMP 算法的运行时间比速度优先组高出 3 个数量级，因此 DMP 算法仅适用于追求准确率而非运行时间的场合。

表 11 在各种网络拓扑结构上推荐使用的溯源算法（时间单位：性能优先组为秒、其他两组为毫秒）

网络拓扑	性能优先				兼顾性能和速度				速度优先			
	算法	准确率	错误距离	耗时	算法	准确率	错误距离	耗时	算法	准确率	错误距离	耗时
随机树	DMP	0.34	0.84	7.06					JCE	0.20	1.08	0.516
随机网络	DMP	0.34	1.01	19.1	RG	0.16	1.78	546	UB	0.09	1.78	1.47
无标度网络	DMP	0.13	1.69	21.9					UB	0.03	1.66	3.66
小世界网络	DMP	0.21	0.86	19.3	DA	0.08	1.03	5.66	UB	0.06	1.68	0.709
法国小学网络	DMP	0.22	1.08	16.3					UB	0.07	1.55	0.801
中国大学网络	DMP	0.12	1.32	23.0					UB	0.03	2.05	4.12

表 12 溯源问题算法总结表

	算法		网络拓扑	传播模型	传播时间	源头数量	快照信息	时间复杂度	空间复杂度	参考文献
传染源中心性	传播中心性	SSE	树	SI	未知	1	完整	$O(n_I)$	$O(n_I)$	[4]
			树	SI	未知	1	部分			[32]
			正则树	SI	未知	1（有候选）	完整	$O(n_I)$	$O(n_I)$	[31]
			一般图	SI	未知	1（有候选）	完整	$O(n_I^3)$	$O(n_I)$	[31]
		SSE-BFS	一般图	SI	未知	1	完整	$O(n_I^2)$	$O(n_I)$	[4]
		TSE	树	SI	未知	2	完整	$O(n^2d^2)$		[25]
		MSEP	树	SI	未知	多个且未知	完整	$O(n^2k^3)$		[25]
		MSEP-BFS	一般图	SI	未知	多个且未知	完整	$O(n^2k^3)$		[25]
	联合传播中心性	URCC	树和一般图	SI	未知	1	完整	树 $O(nl)$，一般图 $O(n^3l)$	$O(nl)$	[34]
	Jordan 中心性	SJC	树	SIR	未知	1	仅 I 节点	$O(n_In^2)$	$O(n_In)$	[42]
			树	SIS	未知	1	仅 I 节点	$O(n_In^2)$	$O(n_In)$	[37]
		JCE	树	SI	未知	1	部分 I 节点	$O(n)$	$O(n)$	[30]
		MJC	树	SI	未知	多个且未知	部分 I 节点			[29]
		RG	一般图	SI	未知	1	部分 I 节点	$O(n^3)$	$O(n+m)$	[30]
	动态年龄	DA	一般图	SI	未知	1 或多	完整		$O(n_I^2)$	[49]
	无偏中介中心性	UB	一般图	雪崩	未知	1	完整	$O(m_In_I)$	$O(n_I)$	[43]
		AUB	一般图	SIR	未知	多个且未知	部分			[41]
置信传播	动态消息传递	DMP	一般图	SIR	已知或未知	1	完整或部分	$O(tmn)$	$O(tn)$	[40]
	因子图	BP	一般图	SIR	已知或未知	1 或多	完整或部分	每次迭代 $O(tg^2m)$		[38]
蒙特卡洛	软边界	MCSM	一般图	SIR	已知或未知	1	完整	$O(n_Ilr)$	$O(n_Il+n)$	[39]
最小描述长度		NetSleuth	一般图	SI	未知	多个且未知	完整	$O(k(n_I+m_I+E_F)$	$O(n_I^2)$	[27]

（2）与之相反，速度优先组的算法基本上都是 UB 算法（在随机树上 JCE 算法更优），虽然准确率显著低于DMP，但是最大错误距离只有2，而且耗时极短，只需数毫秒的时间即可做完。

（3）在随机网络和无标度网络上，兼顾性能和速度的算法分别是 RG 和 DA。与速度优先组的 UB 算法相比，RG 的准确率更高，DA 的错误距离更小，但它们的耗时更长：DA 耗时是 UB 的 7 倍，DMP 的 1/4000；RG 耗时是 UB 的 400 倍，DMP 的 1/40。

7.2 算法总结

7.2.1 算法原理总结与分析

自 2010 年 Shah 和 Zaman 提出网络传播的溯源问题[4] 以来，研究人员基于传染源中心性、置信传播、蒙特卡洛、最小描述长度等，相继提出了各种溯源算法。这些算法的核心思想都可以认为是试图基于贝叶斯概率理论求解传播源的极大似然 $P(\boldsymbol{O}|s)$或者极大后验 $P(s|\boldsymbol{O})$。然而，直接计算使得 $P(\boldsymbol{O}|s)$或者 $P(s|\boldsymbol{O})$最大的传播源是#P-hard 的，因此现有的溯源算法都做了各种假设或者近似，包括：

（1）假设从传播源开始的疾病传播可产生合法的感染顺序最多（“合法”指传播结果与快照信息符合，下同），例如基于传播中心性的算法 SSE、SSE-BFS、TSE、MSEP、MSEP-BFS、URCC 等；

（2）假设传播源与最可能的合法传播路径相关，例如基于 Jordan 中心性的算法 SJC、JCE、MJC、RG 等；

（3）假设传播源是感染图的中心节点，因此根据网络中各个节点的统计性质进行推断，例如基于动态年龄的算法 DA 和基于无偏中介中心性的

算法 UB、AUB 等；

（4）通过因子图上的消息传递来近似计算边际概率，从而计算目标似然或后验，例如基于置信传播的算法 DMP 和 BP 等；

（5）采用模拟仿真估计似然或后验，例如基于蒙特卡洛的算法 MCSM；

（6）通过估计似然或后验的上下界来估计各个节点作为传播源的概率大小关系，例如基于最小描述长度的 NetSleuth。

7.2.2 算法的适用场景和复杂度分析

如表 12 所示，我们从 5 个方面分析本文中溯源算法的不同场景，分别为网络拓扑、传播模型、传播时间、源头数量和快照信息。

（1）网络拓扑：基于传播中心性和 Jordan 中心性的算法都是基于树形拓扑结构提出，然后通过宽度优先搜索树推广到一般网络上；其他算法都是直接在一般网络的拓扑结构上提出。

（2）传播模型：基于传播中心性的算法、DA、NetSleuth 只关心感染图的结构，因此基于 SI 传播模型提出。但是，基于 Jordan 中心性和无偏中介中心性的算法也只关心感染图中的结构，所以虽然表 12 中列举了不同的传播模型，它们实际上对此并不敏感。

（3）传播时间：目前几乎所有的溯源算法都与传播时间无关。但是，基于置信传播的 DMP 和 BP 与基于蒙特卡洛的 MCSM，由于计算过程中需要随时间步迭代计算，因此在已知传播时间的前提下可以达到更快的速度和更优的性能。

（4）源头数量：大部分溯源算法都是解决单源头溯源问题的，但也有不少研究人员将原有的溯源算法推广到多源问题（如 MSEP、MSEP-BFS、MJC、AUB 等），或者直接提出可以解决多源问题的溯源算法（如 DA、BP、NetSleuth 等）。

（5）快照信息：基于传播中心性的算法都要求快照中包含所有节点的状态信息，而基于 Jordan 中心性的则只需部分被感染节点的信息。此外，DA、AUB、MCSM 和 NetSleuth 都需要完整的快照信息，而基于置信传播的溯源算法则不需要。

表 12 还列举了本文中所有溯源算法的时间复杂度和空间复杂度。在算法复杂度中，n 代表网络节点数量，m 代表网络中连边数量，d 代表网络中最大度值，k 代表传播源数量的上界，t 代表传播时间的上界，g 代表节点的最长恢复时间，r 代表单次模拟疾病传播的时间复杂度，l 代表模拟疾病传播的次数，n_{I} 和 m_{I} 分别代表感染图中节点数量和连边数量，E_F 是感染图边界的连边数量。可以看出，SSE 算法和 JCE 算法的时间复杂度最低，与网络规模成线性关系。

由于各个算法复杂度计算公式中的因变量不同，因此很难仅从公式上比较所有算法在时间和空间上的优劣关系。

然而通过实验，我们发现 DMP 在性能上遥遥领先其他所有算法，但耗时很长；与之相对，UB 算法耗时极短，可在需要兼顾性能和速度时使用。

7.2.3 溯源算法的实际应用

在实际应用方面，现有的溯源算法还不够成熟，主要限制在于疾病传播所依赖的复杂网络。一方面，无论何种溯源算法，均要求传播网络完全已知；另一方面，目前传播溯源方面的研究都是基于静态的传播网络。而在实际生活中，人与人之间的接触网络不仅在不断发展变化，而且很难被记录和预测，因此无法直接使用现有的溯源算法。此外，未知的疾病发展过程、缺失的患病信息，都给溯源带来了困难。

上述困难导致了在实际应用中成功案例较少，其中一个典型案例是 Pinto 等人根据 2000 年南非 KwaZulu-Natal 的霍乱疫情所做的算法验证实验[35]。由于在该案例中霍乱通过污染水传播，Pinto 等人将当地主要河流的流域形式化成传播网络，网络中的连边是不同的河道，节点是河道的交汇点。在实验中，Pinto 等人假设只知道 20%节点的感染时刻，推断出的传播源与真实案例中第一个被感染的节点只有平均不到 4 跳的距离。

7.3 免疫策略分析

溯源问题的后续工作通常都是免疫问题，即在给定传播源、传播网络和疾病传播模型的前提下，通过有计划地接种免疫一部分人群，使得传染病在人群中的大规模爆发得到有效控制。在传染病爆发初期，疫苗数量较少，因此如何选择有限数量的人群进行接种免疫使得疾病爆发规模达到最小成为下一个需要解决的重要问题。

目前常用的免疫策略都是针对网络中各个节点不同的特征进行的，按照免疫覆盖率选择特征

较高的人群。这些特征包括度（degree，该节点的邻居节点数量）、强度（strength，该节点与其他节点连边上的权重之和）、聚集系数（clustering coefficient，本文中简称 CC，指示该节点的邻居节点相互连通的程度）、特征向量中心性（邻接矩阵最大特征值对应的特征向量中该节点对应的分量，本文中简称 EC）、中介中心性（本文中简称 BC）。对 CC 值较高的节点来说，邻居节点之间的连通性较好，导致自身在传播过程中的重要性较弱，因此基于 CC 制定的免疫策略选择 CC 值较低（而不是较高）的人群。

为比较这些免疫策略，我们使用 SIR 传播模型，分别在随机网络、无标度网络、小世界网络、法国小学网络、中国大学网络上做了疾病传播和接种免疫的仿真。此处我们使用的法国小学网络[75] 和中国大学网络[76] 都是原始的动态变化网络，而非前文提到的无权静态网络；我们使用的随机网络、无标度网络和小世界网络也都如文献[76] 中那样采用与中国大学网络相同的规模，而非前文中使用的理想网络。

我们计算了免疫覆盖率为 25%时，在不同的网络上使用不同的免疫策略后能够减少的感染人数比重（计算公式：1-免疫后感染人数/原感染人数），并将结果记录在表 13 中。

考虑到理想网络都是无权网络，度和强度的值完全一致，因此在三种理想网络（随机网络、无标度网络和小世界网络）上，我们将度和强度两种免疫策略合并表示。为了方便比较，我们将每种网络拓扑上最优的免疫结果加粗表示。

表 13 不同免疫策略在不同网络上的效果

网络拓扑	度	强度	CC	EC	BC
随机网络	0.239		0.238	0.239	**0.243**
无标度网络	**0.294**		0.240	0.265	0.279
小世界网络	0.239		0.238	**0.256**	0.238
法国小学网络	**0.902**	0.895	0.789	0.752	0.863
中国大学网络	**0.661**	0.610	0.314	0.501	0.595

从表 13 中可以看出，三种理想网络中最优免疫措施各不相同，分别是中介中心性（随机网络）、度（无标度网络）、特征向量中心性（小世界网络）；但是在两种真实网络（法国小学网络和中国大学网络）上，最优免疫措施都是计算方法最简单的度。免疫策略最终都要应用到真实的人际接触网络上真实的疾病传播过程，因此可考虑直接使用接触网络中的度值作为选取接种免疫人群的指标。

7.4 研究趋势分析

经过多年的发展，虽然研究者们已经提出了多种溯源算法，但是依然存在着准确度和时间复杂度方面的不足。

在以上分析的基础上，笔者认为溯源问题有如下研究趋势：

趋势一、提高算法的准确度

理想情况下，传染源可以直接通过最大化快照信息的似然函数 $P(\boldsymbol{O}|s)$来推断。由于 $P(\boldsymbol{O}|s)$难以计算，大部分研究人员使用其他统计量（如传播路径的似然或者传播顺序的计数）就直接推断传染源，或者在计算时使用各种近似技术（如在计算似然时认为每个节点的状态相互独立）。这些近似造成了不可忽视的推断误差，因此更加准确地直接计算似然的方法，是值得研究的问题。

此外，目前大部分溯源算法的输入信息都是被感染节点和传播网络，并没有很好地利用其他信息，例如疾病传播模型、各个节点被感染的时间、网络中传播关系的先验知识等。如果能够很好地整合这些信息，就可以在计算似然函数时排除很多无需考虑的情形，进而提高传播源推断的准确性，同时也可降低算法的耗时。

趋势二、降低算法复杂度

现有的一部分溯源算法已经达到较高的性能，但是耗时过长。如果能够加快计算过程，例如在不影响算法准确度的前提下精简迭代过程、规避一些重复计算、开发并行计算方法，则能够优化现有溯源算法的复杂度，这将对 DMP 这样的高性能长耗时算法是极大的改善。同时，如果利用传播过程中得到的其他信息，例如各个节点被感染的时间、网络中传播关系的先验知识等，也可进一步精简计算过程，达到降低算法耗时的效果。

趋势三、提高现有算法的实用性

现有的溯源算法，如基于传播中心性和 Jordan 中心性的算法，已经被推广到了多种溯源问题。但是，这些溯源算法并没有覆盖所有层次上的扩

展问题。

如表 12 所示，基于传播中心性和 Jordan 中心性的算法都还没有被推广到在一般图上、有多个传播源、快照信息只覆盖部分节点的场景。更重要的是，目前的溯源算法大部分也只是针对少量的疾病传播模型，如SI和SIS模型等。然而对于更实用的疾病模型，如存在潜伏期的 SEIR 疾病模型、或者 SIRS 这种经过一段免疫期后又变回易感状态的疾病模型，现有算法无法处理，因此需要进一步的研究。

此外，目前的溯源算法基本上都是基于疾病传播网络稳定不变的情形提出，但是在现实中，由于人类的活动，疾病传播网络应该处于时刻变动之中。目前研究人员提出的算法中只有极少部分可以应用于动态变化的网络拓扑，如 DMP 算法和 BP 算法，还有很多基于动态变化网络的溯源算法有待开发。

目前的溯源问题都只是为了更好地了解传染病的发展过程而去追溯疾病的源头，但笔者认为，溯源问题还可以与其他问题结合，形成全新的问题。例如，当快照中只覆盖部分信息时，研究人员可以先进行溯源，找到疾病的源头，然后再根据源头来推测网络中其他节点被感染的概率，从而有针对性地对健康患者进行检查，进而更高效地控制疫情。在此例中，研究人员不仅可以分别解决溯源问题和预测问题，还可以将它们结合起来处理，以期降低近似计算造成的误差。

疾病传播溯源问题是复杂网络上疾病传播领域的重要组成部分。溯源问题的研究，为控制传染病的传播提供了重要的理论支撑；此外，疾病传播溯源问题上的算法还可以考虑推广到谣言传播和计算机病毒传播的研究中，从而促进复杂网络等信息科学的发展。

参 考 文 献

[1] Williams B G, Granich R, Chauhan L S, et al. The impact of HIV/AIDS on the control of tuberculosis in India. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(27): 9619-9624.

[2] Smith R D. Responding to global infectious disease outbreaks: lessons from SARS on the role of risk perception, communication and management. Social Science & Medicine, 2006, 63(12): 3113-3123.

[3] Organization W H. Global tuberculosis report 2013. Geneva, Switzerland: World Health Organization, 2013.

[4] Shah D, Zaman T. Detecting sources of computer viruses in networks: theory and experiment. Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems. New York, USA, 2010: 203-214.

[5] Shah D, Zaman T. Rumors in a network: who's the culprit? IEEE Transactions on Information Theory, 2011, 57(8): 5163-5181.

[6] Pastor-Satorras R, Castellano C, Van Mieghem P, et al. Epidemic processes in complex networks. Reviews of Modern Physics, 2015, 87(3): 925.

[7] Hethcote H W. The mathematics of infectious diseases. SIAM Review, 2000, 42(4): 599-653.

[8] Newman M E. Spread of epidemic disease on networks. Physical Review E, 2002, 66(1): 16128.

[9] Duan W, Fan Z, Zhang P, et al. Mathematical and computational approaches to epidemic modeling: a comprehensive review. Frontiers of Computer Science, 2015, 9(5): 806-826.

[10] Eames K T. Modelling disease spread through random and regular contacts in clustered populations. Theoretical Population Biology, 2008, 73(1): 104-111.

[11] Fefferman N H, Ng K L. How disease models in static networks can fail to approximate disease in dynamic networks. Physical Review E, 2007, 76(3): 31919.

[12] Smieszek T, Fiebig L, Scholz R W. Models of epidemics: when contact repetition and clustering should be included. Theoretical Biology and Medical Modelling, 2009, 6(1): 11.

[13] Lappas T, Terzi E, Gunopulos D, et al. Finding effectors in social networks. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, USA, 2010: 1059-1068.

[14] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington , USA, 2003: 137-146.

[15] Seo E, Mohapatra P, Abdelzaher T. Identifying rumors and their sources in social networks. Proceedings of the SPIE, Maryland, USA, 2012, 8389:1-13.

[16] Farajtabar M, Gomez-Rodriguez M, Du N, et al. Back to the past: source identification in diffusion networks from partially observed cascades. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, San Diego, California, USA, 2015: 232-240.

[17] Nguyen D T, Nguyen N P, Thai M T. Sources of misinformation in online social networks: who to suspect? IEEE Military Communications Conference, Orlando, USA, 2012: 1-6.

[18] Zhu K, Ying L. Source localization in networks: trees and beyond.

arXiv preprint arXiv:1510.01814. 2015.

[19] Louni A, Santhanakrishnan A, Subbalakshmi K P. Identification of source of rumors in social networks with incomplete information. arXiv preprint arXiv:1509.00557. 2015.

[20] Zhang Z, Xu W, Wu W, et al. A novel approach for detecting multiple rumor sources in networks with partial observations. Journal of Combinatorial Optimization, 2017, 33(1): 132-146.

[21] Zang W, Wang X, Yao Q, et al. A fast climbing approach for diffusion source inference in large social networks. Proceedings of the Second International Conference on Data Science, Sydney, Australia, 2015: 50-57.

[22] Jiang J, Wen S, Yu S, et al. Identifying propagation sources in networks: state-of-the-art and comparative studies. IEEE Communications Surveys and Tutorials, 2016, PP(99): 1

[23] Luo W, Tay W P. Identifying infection sources in large tree networks. 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), Seoul, Korea, 2012: 281-289.

[24] Luo W, Tay W P. Identifying multiple infection sources in a network. Proceedings of the 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove,USA, 2012: 1483-1489.

[25] Luo W, Tay W P, Leng M. Identifying infection sources and regions in large networks. IEEE Transactions on Signal Processing, 2013, 61(11): 2850-2865.

[26] Newman M. Networks: an introduction. Oxford, UK: Oxford University Press, 2010.

[27] Prakash B A, Vreeken J, Faloutsos C. Efficiently spotting the starting points of an epidemic in a large graph. Knowledge and Information Systems, 2014, 38(1): 35-59.

[28] Prakash B A, Vreeken J, Faloutsos C. Spotting culprits in epidemics: how many and which ones? 2012 IEEE 12th International Conference on Data Mining (ICDM), Brussels, Belgium, 2012: 11-20.

[29] Luo W, Tay W P. Estimating infection sources in a network with incomplete observations. 2013 IEEE Global Conference on Signal and Information Processing, Austin, Texas, USA, 2013: 301-304.

[30] Luo W, Tay W P, Leng M. How to identify an infection source with limited observations. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(4): 586-597.

[31] Dong W, Zhang W, Tan C W. Rooting out the rumor culprit from suspects. 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, 2013: 2671-2675.

[32] Karamchandani N, Franceschetti M. Rumor source detection under probabilistic sampling. 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, 2013: 2184-2188.

[33] Shah D, Zaman T. Rumor centrality: a universal source detector. Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems, London, UK, 2012: 199-210.

[34] Wang Z, Dong W, Zhang W, et al. Rumor source detection with multiple observations: fundamental limits and algorithms. Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems, Austin, USA , 2014: 1-13.

[35] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks. Physical Review Letters. 2012, 109(6): 68702.

[36] Louni A, Subbalakshmi K P. A two-stage algorithm to estimate the source of information diffusion in social media networks. Proceedings of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto,Canada, 2014: 329-333.

[37] Luo W, Tay W P. Finding an infection source under the SIS model. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, British Columbia, Canada, 2013: 2930-2934.

[38] Altarelli F, Braunstein A, Dall Asta L, et al. Bayesian inference of epidemics on networks via belief propagation. Physical Review Letters, 2014, 112(11): 118701.

[39] Antulov-Fantulin N, Lančić A, Amuc T, et al. Identification of patient zero in static and temporal networks: robustness and limitations. Physical Review Letters, 2015, 114(24): 248701.

[40] Lokhov A Y, Mézard M, Ohta H, et al. Inferring the origin of an epidemic with a dynamic message-passing algorithm. Physical Review E. 2014, 90(1): 12801.

[41] Zang W, Zhang P, Zhou C, et al. Locating multiple sources in social networks under the SIR model: a divide-and-conquer approach. Journal of Computational Science, 2015, 10: 278-287.

[42] Zhu K, Ying L. Information source detection in the SIR model: a sample path based approach. IEEE/ACM Transactions on Networking, 2016, 24(1): 408-421.

[43] Comin C H, Da Fontoura Costa L. Identifying the starting point of a spreading process in complex networks. Physical Review E. 2011, 84(5): 56105.

[44] Zhang Y B, Zhang X Z, Xu C, Zhang B. Fast source localization method for social network. Journal of Northeastern University (Natural Science), 2016, 37(4): 467-471 (in Chinese)
(张聿博, 张锡哲, 徐超, 张斌. 社交网络信息源快速定位方法. 东北大学学报（自然科学版）, 2016, 37(4): 467-471)

[45] 付世海. 基于社团结构的网络多传播源定位算法研究[硕士学位论文]. 沈阳: 东北大学, 2013. Fu S. The multi-source localization research of network propagation based on community structure[M.S. thesis]. Shenyang:

Northeastern University. 2013.

[46] 晏迪. 面向网络扩散源点定位的观察点部署策略研究及定位算法优化[硕士学位论文]. 沈阳: 东北大学, 2013. Yan D. The research of observers placement about source localization for network diffusion and optimization of localization algorithm[M.S. thesis]. Shenyang: Northeastern University. 2013.

[47] 臧文羽. 大规模在线社交网络负面信息传播分析与引导[博士学位论文]. 北京: 中国科学院计算技术研究所, 2016. Zang W. Propagation analysis and guidance of negative information in large-scale social networks[Ph.D. thesis]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences. 2016.

[48] 董文祥. 网络中信息传播：信息源选择与检测的若干关键问题研究[博士学位论文]. 合肥: 中国科学技术大学, 2014. Dong W. Research on selection and detection of information sources for networked diffusion[Ph.D. thesis]. Hefei: University of Science and Technology of China. 2014.

[49] Fioriti V, Chinnici M. Predicting the sources of an outbreak with a spectral technique. arXiv preprint arXiv:1211.2333. 2012.

[50] 徐超. 基于部分传播路径的社交网络传播源点定位方法研究[硕士学位论文]. 沈阳: 东北大学, 2014. Xu C. The research of social network source location based on partial propagation paths[M.S. thesis]. Shenyang: Northeastern University. 2014.

[51] Brightwell G, Winkler P. Counting linear extensions. Order, 1991, 8(3): 225-242.

[52] Valiant L G. The complexity of enumeration and reliability problems. SIAM Journal on Computing, 1979, 8(3): 410-421.

[53] Watts D J, Strogatz S H. Collective dynamics of “small-world” networks. Nature, 1998, 393(6684): 440-442.

[54] Harary F. Graph theory. Boston, USA: Addison Wesley Longman Publishing Co., 1972.

[55] Hedetniemi S M, Cockayne E J, Hedetniemi S T. Linear algorithms for finding the Jordan center and path center of a tree. Transportation Science, 1981, 15(2): 98-114.

[56] Achterberg T. SCIP: solving constraint integer programs. Mathematical Programming Computation, 2009, 1(1): 1-41.

[57] Zhu G, Yang H J, Yang R, et al. Uncovering evolutionary ages of nodes in complex networks. The European Physical Journal B: Condensed Matter and Complex Systems, 2012, 85(3): 1-6.

[58] Albert R, Barabási A. Statistical mechanics of complex networks. Reviews of Modern Physics, 2002, 74(1): 47.

[59] Restrepo J G, Ott E, Hunt B R. Characterizing the dynamical importance of network nodes and links. Physical Review Letters, 2006, 97(9): 94102.

[60] Francis J G F. The QR transformation, I. The Computer Journal, 1961, 4(3): 265-271.

[61] Francis J G F. The QR transformation, II. The Computer Journal, 1962, 4(4): 332-345.

[62] Kublanovskaya V N. On some algorithms for the solution of the complete eigenvalue problem. USSR Computational Mathematics and Mathematical Physics, 1961, 1(3): 637-657.

[63] Brandes U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 2001, 25(2): 163-177.

[64] Newman M E, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 26113.

[65] Newman M E. Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 2006, 74(3): 36104.

[66] Mezard M, Montanari A. Information, physics, and computation. Oxford, UK: Oxford University Press, 2009.

[67] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Burlington, Massachusetts, USA: Morgan Kaufmann, 2014.

[68] Antulov-Fantulin N, Lančić A, Atefančić H, et al. Fast SIR algorithm: a fast algorithm for the simulation of the epidemic spread in large networks by using the susceptible–infected–recovered compartment model. Information Sciences, 2013, 239: 226-240.

[69] Vestergaard C L, Génois M. Temporal Gillespie algorithm: fast simulation of contagion processes on time-varying networks. PLoS Computational Biology, 2015, 11(10): e1004579.

[70] Rocha L E, Liljeros F, Holme P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. PLoS Computational Biology, 2011, 7(3): e1001109.

[71] Grünwald P D. The minimum description length principle. Massachusetts, USA: The MIT Press, 2007.

[72] Rissanen J. Modeling by shortest data description. Automatica, 1978, 14(5): 465-471.

[73] Cover T M, Thomas J A. Elements of information theory. New Jersey, USA: Wiley-Interscience, 2006.

[74] Bikhchandani S, Hirshleifer D, Welch I. A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of Political Economy, 1992, 100(5): 992-1026.

[75] Stehlé J, Voirin N, Barrat A, et al. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. PLoS ONE, 2011, 6(8): e23176.

[76] Huang C, Liu X, Sun S, et al. Insights into the transmission of respiratory infectious diseases through empirical human contact networks. Scientific Reports, 2016, 6: srep31484.

HUANG Chun-Lin, born in 1987, Ph. D. candidate. His research interest includes epidemic spreading on complex networks.

LIU Xing-Wu, born in 1976, associate professor, master supervisor. His research interest includes graph algorithms and distributed algorithms.

DENG Ming-Hua, born in 1969, professor, Ph. D. supervisor. His current research interest includes bioinformatics and system biology.

ZHOU Yang, born in 1980, Ph. D. candidate. Her main research interest includes molecular epidemiology of tuberculosis.

BU Dong-Bo, born in 1973, professor, Ph. D. supervisor. His current research interest includes algorithm design and analysis, and bioinformatics.

Background

Epidemic source identification is to estimate infection sources based on the knowledge of health states of individuals and the structure of the underlying propagation network. It is an important task for preventing infectious epidemic outbreaks beforehand, and helping medical scientists design effective drugs for newly occurred plagues. However, due to complex propagation inference on networks, it remains a great challenge in the field of epidemic spreading on complex networks.

After the problem first raised by Shah and Zaman, different algorithms have been proposed, although none of them guarantees exact solution. The algorithms can be categorized into four families: 1) likelihood approaches based on network structure, including calculate rumor centralities, Jordan centralities, dynamical ages and unbiased betweenness centralities of all nodes in the network; 2) Bayesian approaches based on belief propagation, including deriving dynamic message-passing equations or using belief propagation technique, such that likelihoods of the snapshot observation are inferred with the pre assumed epidemic source; 3) Monte-Carlo approaches based on Soft-Margin, employing direct numerical simulations with the disease spreading model; 4) the approach based on minimum description length principle. All existing approaches employ approximate technique in the algorithm designing process due to high computation complexity on likelihoods, and only the approach based on belief propagation gives exact likelihood on tree networks. Most of the algorithms performs well only on certain network structures. For example, SleuthNet only performs well on grid networks, and rumor centrality only performs well on geometric trees.

In this survey, we summarized the existing approaches to epidemic source identification. Comparative evaluation of identification algorithms shows the following observations: (1) All algorithms are robust to different epidemic spreading parameters. (2) DMP algorithm, though time-consuming, makes the best source identification algorithm through rather high accuracy and low error distance. (3) UB algorithm runs rather fast on almost all networks with comparative accuracy and error distance with other algorithms.

The study is funded by the National Science and Technology Major Project under Grant 2008ZX10003009-005, the National Basic Research Program of China (973 Program) under Grant 2012CB316502, the National Nature Science Foundation of China under Grants 11175224, 11121403, 31270834 and 61272318, and the Open Project Program of State Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, China (No.Y4KF171CJ1).