计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20121044
51(1):1-16,2014

SFEN-Inf：一种微博信息传播网络推理算法

郑众杰　林学练

(北京航空航天大学计算机学院 北京 100191)
(godwithzhongjie@163.com)

SFEN-Inf: An Inference Algorithm for Weibo Information Diffusion Network

Zheng Zhongjie and Lin Xuelian
(School of Computer Science and Engineering, Beihang University, Beijing 100191)

Abstract One feature of the Online Social Network is the rapid spreading and frequent interaction of information. To understand the properties of social network deeply, we usually need to know the structure of the information diffusion network. However, the network is always latent. We can observe timestamps and contents of information, but it is hard to find the information diffusion network. So, inferring the network of information diffusion according to observed data, which called information diffusion network inferrence problem, has important research meaning. We study the problem in this article, define a general probabilistic model that combines timestamp and content similarity and develop SFEN-Inf algorithm to infer the information diffusion network structure. We conduct our experiments on Sina Weibo datasets which composed of typical topics. We compare our algorithm with state of the art NetInf algorithm and conclude that SFEN-Inf has about 2 times higher accuracy and improves time efficiency by several orders of magnitude.

Key words information diffusion; network inference; SFEN-Inf; NetInf; weibo

摘 要 在线社交网络的特点之一是信息的快速传播和频繁交互。为了更好的研究社交网络的特性，我们需要知道信息传播网络的结构。然而，信息传播网络通常是潜在的。我们能观察到信息包含的时间戳和文本等内容，却难以直接观测到信息传播网络。因此，如何根据观测到的数据准确地推理出潜在的传播网络结构，即信息传播网络的推理问题，具有重要的研究意义。本文对该问题进行研究，根据信息的时间和文本内容的相似性，建立了信息传播的概率模型，提出了信息传播网络推理算法 SFEN-Inf。我们针对典型的微博事件对 SFEN-Inf 算法进行实验分析，并将 SFEN-Inf 算法与著名的 NetInf 算法进行比较。实验结果表明，SFEN-Inf 算法在传播网络推理效果上提高约 2 倍，并且算法的时间效率有较明显的提高。

关键词 信息传播；网络推理；SFEN-Inf；NetInf；微博

中图法分类号 TP393

信息传播是在线社交网络的主要特点之一[1]。常用的社交媒体[2][17]，如新浪微博和微信，每天有大量的用户发帖，这些用户生成的内容被其他用户引用和转发，继而在由虚拟用户构成的社交网络中传播，形成信息传播网络。信息传播网络是影响力最大化分析[4]、关键用户发现[1]和精准广告推荐[5]等社交网络分析和应用的基础之一[3]，因此被学术界和工业界所关注。

信息传播网络和社交网络有关，假设用户只能原创发贴，或者引用和转发其关注的用户发的帖子，那

收稿日期：2014-10-25
基金项目：国家“九七三”重点基础研究发展计划基金项目（2014CB340300）

么信息传播网络是用户、用户关注的人和用户的粉丝形成的社交网络的子集。一方面，信息传播网络的结构通常是潜在的或者难以被直接观察的。比如微博运营商出于用户隐私保护的原因，不会向外界公布完整的社交网络，这不利于信息传播网络的分析。另一方面，我们能观察到网络中节点(指微博用户)的属性，信息的文本内容，以及用户发表该信息的时间等。因此，根据观察到的数据推理潜在的信息传播路径，就成了发现传播网络的主要途径。

针对观察到的信息的时间戳和文本内容，本文提出了一种推理信息传播网络的方法。首先，我们分别分析了信息的时间和文本内容对信息传播的影响，进而利用概率模型定义信息传播网络推理问题，并把该问题转化为目标函数最优化问题，以简化求解。其次，我们提出 SFEN-Inf (Simplified Feature Enhanced Network Inference)算法来求解该问题，算法的思想是找到给定数量的边，使得目标函数的值最优。最后，本文以“北京高校食堂死猪”，“李开复癌症”，“城管打死临武瓜农”和“芦山地震”四个事件的微博数据为基础对 SFEN-Inf 算法进行实验验证。实验表明，SFEN-Inf 算法比 NetInf 算法[10]在推理效果上约有 2 倍的提高，而且算法的运行时间有较明显的减少。因此，本文的贡献在于提出的传播概率模型同时考虑了时间和文本因素，并且简化了网络推理问题，使得 SFEN-Inf 算法具有高效，可扩展的特点，能适用于大规模网络。

本文内容结构如下：第 1 节对相关工作进行简要介绍；第 2 节是问题定义和数学建模；第 3 节提出 SFEN-Inf 算法；第 4 节是实验和算法的性能评估；第 5 节是总论和工作展望。

1 相关工作

网络结构推理问题是一个已经被广泛研究的基础性问题[6,7]，但是结合社交网络信息传播和影响特性，网络推理问题又提出新的挑战。Gruhl 等[8]和 Adar 等[9]首先提出了信息传播网络推理问题。2010 年，Gomez 和 Jure 在文献[10]中对信息传播网络推理问题进行研究，提出了 NetInf 算法。他们首先分析了节点“感染”信息的时间次序，提出了一种包含时间因素的概率模型，将传播网络问题归结为最优化问题，并在理论上证明该问题是 NP-hard 问题。其次，他们转换该问题，使得转换后的最优化问题满足子模特征(Submodularity)，得以找到一种贪心算法求出近似最优解。NetInf 算法的思想很直观：如果两个用户发布信息的时间间隔总是相对较小，那么很可能在节点间存在一条边。以新浪微博为例，A 在某时刻发布帖子，B 在接下来很短的时间内引用或转发 A 的信息，并且常常这样，那么就认为 A 和 B 之间有连通关系。

NetInf 算法有如下不足之处：(1)模型仅考虑时间因素而忽略社交网络的其他特征，比如文本和节点属性等。这也导致了 NetInf 算法在合成数据集上推理效果显著，但是在真实数据集上效果不明显，通常准确率和召回率在 0.3 左右。(2)算法时间效率不高，通常推理 500 个节点，4000 条边左右的网络需要小时级别的时间，因此无法适用于大规模的社交网络。

Myers 在文献[11]提出 ConnIE 算法，Gomez 在文献[12]提出 NetRate 算法。ConnIE 算法和 NetRate 算法都是利用凸规划思想求解问题。ConnIE 算法不仅能推理出传播网络，而且能给出网络中每条边存在的概率。NetRate 用生存函数和风险函数来刻画网络推理问题，相比 NetInf 算法和 ConnIE 算法，NetRate 算法推理效果有所提高，但不明显，而且时间消耗更多。

Gomez 关于该问题的最新研究是在文献[13]中提出的 INFOPATH 算法，其贡献在于否定之前工作中关于静态网络的假设，引入动态网络的概念。

上述工作都较好的研究了网络推理问题，但都存在着在真实数据集上推理效果不显著和时间效率不高的问题。本文提出的 SFEN-Inf 算法同时考虑了时间和文本两方面因素，并且给出了更为简化的网络推理模型，使得 SFEN-Inf 算法具有比 NetInf 算法更高的准确率和时间效率，从而更好的适应大规模网络的情况。

2 问题定义

在社交网络中，我们通常能观察到节点“感染”信息的时间、信息的文本和节点属性等。但是却无法观察承载信息流动的网络结构。本文的目标是根据观测数据推理潜在的信息传播网络。

图 1 是本文针对“北京高校食堂死猪”事件的新浪微博数据所推理出的部分新浪微博传播网络。图中，节点代表新浪微博的用户，有向边代表用户之间的关注关系，尺寸较大的节点表示影响力较高的用户，较粗的边表示用户之间信息传播发生的概率较大。信息通常是从影响力较大的节点流向普通节点，但是影响力较大的节点和普通节点之间的关系并不一定最强(边可能很细)。这说明微博名人对普通用户的影响通常很大，但是普通用户跟他的朋友们关系更加密切，信息交互更加频。

图 1 “北京高校食堂死猪”事件部分网络图

2.1 基本假设

本文基于两个基本假设：(1)存在这样的网络G，使得所有的信息流动和交互都在该网络上发生。(2)潜在的网络G是静态的，不随着时间的变化而变化。

我们利用信息发布的时间和信息的内容来推理潜在的网络G，把推理出来的网络称作$G^{\wedge}$。

2.2 信息级联

信息在网络上流动会留下“足迹”，我们称之为信息级联[3]，每一条信息级联c表示信息的生命周期中，从源节点起不断“感染”其他节点的过程。比如，新浪微博中某个用户发表一条原创帖子，有若干看见该帖子的人会转发或评论该条帖子，直到该帖子不再传播，此时该条信息级联c停止。

给定潜在的网络G，我们观察到n条信息级联c在G上传播，这些信息级联的集合用$\boldsymbol{C}$表示如下：

$$\boldsymbol{C}=\{c_1,c_2,...,c_n\} \tag{1}$$

其中，每条信息级联c_i有如下形式：

$$c_i=\{(v_1,t_1,m_1)_i,(v_2,t_2,m_2)_i,...,(v_k,t_k,m_k)_i\} \tag{2}$$

其中$1\le i\le n$，v表示网络节点，t表示信息传播到该节点的时间，m表示信息的文本内容，k表示信息级联包含的节点数(信息级联的长度)。图 2 是信息级联的一个示例。图中有向边表示信息的流向，随着时间的推移，信息从一个节点传播到令一个节点，那么信息级联为：$\{(v_1,t_1,m_1),(v_2,t_2,m_2),...,(v_6,t_6,m_6)\}$，其中$t_1<t_2<...<t_6$。

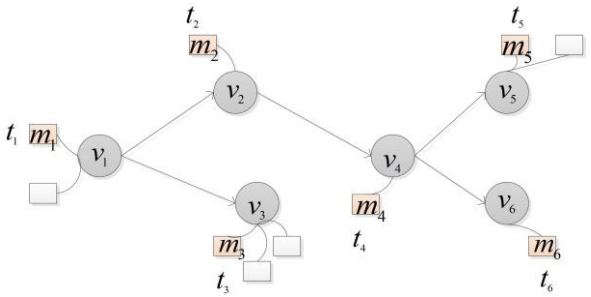

图 2 信息级联示例

2.3 传播概率

本文的传播模型基于独立级联(IC)模型[4]，该模型认为一个被“感染”的节点会以一定的概率独立地影响他的邻居节点，被“感染”的节点只能影响未被“感染”的节点，并且被“感染”的节点不会再次被“感染”。以新浪微博为例，某个时刻，A 用户发了一条微博，那么 A 用户会独立的以不同的概率影响他所有的粉丝，有些粉丝会被影响而转发或评论该条微博，有些则不会。

已知信息级联c，信息从节点v_i传播到节点v_j的概率表示为$P_c(i,j)$，传播概率不仅受时间的影响，而且受文本内容的影响，它们共同决定传播概率的大小，$P_c(i,j)$表示如下：

$$P_c(i,j)=F_c(t_i,t_j,m_i,m_j) \tag{3}$$

公式(3)中，我们将传播概率表示成时间和文本的函数F。因为信息只能沿着时间往前传播，所以对于给定的信息级联元组$(v_i,t_i,m_i)_c$和$(v_j,t_j,m_j)_c$，当$t_i\ge t_j$时，$F_c(t_i,t_j,m_i,m_j)=0$。

2.3.1 文本相似性对传播概率的影响

我们用Δ_m表示m_i和m_j之间的相似性，如下：

$$\Delta_m=\left\|m_i-m_j\right\| \tag{4}$$

其中m_i和m_j分别是文本的词的集合。为了计算Δ_m，我们引入 Jaccard 距离来衡量Δ_m的大小：

$$\Delta_m=1-\frac{\left|m_i\cap m_j\right|}{\left|m_i\cup m_j\right|} \tag{5}$$

如果Δ_m的值越小，那么m_i和m_j的相似性越高，信息从节点v_i传播到v_j的概率越大。同样以新浪微博为例，转发微博通常会引用原创微博的内容，甚至有许多用户在他们的转发微博中完全引用原创微博的内容，此时文本相似性很高。我们令$f_c(m_i,m_j)$表示文本m_i和m_j的相似性对传播概率的影响，如下：

$$f_c(m_i,m_j)=\beta e^{-\Delta_m} \tag{6}$$

其中β是待估参数。

2.3.2 时间对传播概率的影响

我们用 Δ_t 表示时间 t_i 和 t_j 的间隔，如下：

$$\Delta_t = \left\| t_i - t_j \right\| \tag{7}$$

直观上，如果 Δ_t 的值越大，那么传播概率 $P_c(i,j)$ 就越小。比如，某新浪微博用户发了一条帖子，最初的几个小时内，该帖子被转发的概率最大，经过的时间越长，该帖子再被转发的概率就越小。

我们分别用三种不同的参数模型：指数分布(**EXP**)、幂律下降(**POW**)和瑞利分布(**RAY**)来表示时间和传播概率之间的关系[14-16]：

指数分布：

$$F_c(t_i,t_j,m_i,m_j)=\begin{cases}\alpha f_c(m_i,m_j)e^{-\alpha\Delta_t} & t_i<t_j\\ 0 & t_i\geq t_j\end{cases} \tag{8}$$

幂律分布：

$$F_c(t_i,t_j,m_i,m_j)=\begin{cases}\dfrac{\alpha}{\delta} f_c(m_i,m_j)\left(\dfrac{\Delta_t}{\delta}\right)^{-1-\alpha} & t_i<t_j\\ 0 & t_i\geq t_j\end{cases} \tag{9}$$

瑞利分布：

$$F_c(t_i,t_j,m_i,m_j)=\begin{cases}\alpha f_c(m_i,m_j)\Delta_t e^{-\alpha\frac{(\Delta_t)^2}{2}} & t_i<t_j\\ 0 & t_i\geq t_j\end{cases} \tag{10}$$

其中 α 和 δ 是待估参数。

将公式(6)代入公式(8)，得到指数分布的传播概率为：

$$P_c(i,j)=F_c(t_i,t_j,m_i,m_j)=\begin{cases}\alpha\beta e^{-(\Delta_m+\alpha\Delta_t)} & t_i<t_j\\ 0 & t_i\geq t_j\end{cases} \tag{11}$$

幂律分布和瑞利分布同理。

2.4 网络推理问题

2.2 和 2.3 节分别定义了信息级联和传播概率，接下来我们给出网络推理模型。假设 $G(V,E)$ 是潜在的传播网络，因为信息级联的集合 $\boldsymbol{C}$ 已知，所以传播网络的节点集 V 已知，但传播网络的边集 E 需要推理。我们将某条信息级联 c 在网络 G 上传播的条件概率定义为 $P(c|G)$，因为本文的传播模型基于 IC 模型，$P(c|G)$ 表示如下：

$$P(c|G)=P(c|E)=\prod_{(v_i,v_j)\in E}P_c(i,j) \tag{12}$$

因此，我们需要在给定 $\boldsymbol{C}$ 的情况下找到使得 $P(c|E)$ 概率最大的边的集合 E^*：

$$E^*=\arg\max_{|E|\leq k}P(E|\boldsymbol{C}) \tag{13}$$

其中 k 是提前给定的网络中边的数量。当然，如果 G 是完全图一定会使得问题最优，但是现实中的网络通常为稀疏网络。根据贝叶斯公式，我们有：

$$P(E|\boldsymbol{C})=\frac{P(E)P(\boldsymbol{C}|E)}{P(\boldsymbol{C})}$$

因为 $\boldsymbol{C}$ 已知，假设各条信息级联相互独立，那么：

$$P(E|\boldsymbol{C})\propto P(E)P(\boldsymbol{C}|E)=P(E)\prod_{c\in\boldsymbol{C}}P(c|E)$$

对于任意的网络 $G(V,E)$，由于静态网络的假设，我们认为 $P(G)=P(E)$ 为常量，那么：

$$P(E|\boldsymbol{C})\propto\prod_{c\in\boldsymbol{C}}\prod_{(v_i,v_j)\in E}P_c(i,j)$$

将函数取对数，得到：

$$P(E|\boldsymbol{C})\propto\sum_{c\in\boldsymbol{C}}\sum_{(v_i,v_j)\in E}\log P_c(i,j)$$

即：

$$P(E|\boldsymbol{C})\propto\sum_{(v_i,v_j)\in E}\sum_{c\in\boldsymbol{C}}\log P_c(i,j)$$

令 $w_c(i,j)=\log P_c(i,j)$，那么公式(13)转化为：

$$E^*=\arg\max_{|E|\leq k}\sum_{(v_i,v_j)\in E}\sum_{c\in\boldsymbol{C}}w_c(i,j) \tag{14}$$

公式(14)刻画了网络推理问题，其中 $w_c(i,j)$ 为边 (v_i,v_j) 的权重，表明在给定的信息级联 c 中，节点 v_i 影响节点 v_j 的能力。令 $W(i,j)$ 表示边 (v_i,v_j) 在所有信息级联 $\boldsymbol{C}$ 中的权重总和：

$$W(i,j)=\sum_{c\in\boldsymbol{C}}w_c(i,j) \tag{15}$$

将等式(15)代入(14)得：

$$E^*=\arg\max_{|E|\leq k}\sum_{(v_i,v_j)\in E}W(i,j) \tag{16}$$

现在，网络推理的最优化问题简化为：我们只需要在由信息级联 $\boldsymbol{C}$ 确定的所有候选边里面找到 k 条边，使得所有边的权重之和最大。相比 Gomez 提出的 NetInf 模型和算法，我们的网络推理模型更加简单，并且该模型中传播概率的定义考虑了除时间外的其他因素，使得推理结果更为准确。

3 SFEN-Inf 算法

我们提出的算法称为 SFEN-Inf(**S**implified **F**eature **E**nhanced **N**etwork **Inf**erence)，该算法是传播概率敏感的(传播概率的刻画是否准确将直接影响算法的效果)。在 NetInf 算法中，每一条信息级联 c 对应传播树 T，并且算法的每轮迭代会更新传播树的结构，使得信息级联 c 以传播树 T 传播的概率 $P(c|T)$ 最大。SFEN-Inf 算法并未引入传播树的概念，而是直接

比较候选边的传播概率，筛选出 k 条传播概率最大的边组成传播网络的边集。SFEN-Inf 算法思想简单，尽管没有将信息级联 c 表达为传播树 T，但是 SFEN-Inf 算法的传播概率模型结合时间和文本相似性，使得传播概率的刻画更加准确，因此算法迭代过程选出的边很可能就是传播网络中的边，SFEN-Inf 算法更好的适用于推理那些 $w_c(i,j)$ 能准确刻画的网络。

算法初始状态，边集 E 为空(第 1 行)，对于每一条信息级联 c，计算所有候选边的权重 $w_c(i,j)$ (第 4 行)，如果边集 E 包含候选边 (v_i,v_j)，则更新边 (v_i,v_j) 在所有的信息级联 $\boldsymbol{C}$ 中的权重总和 $W(i,j)$，否则将该候选边 (v_i,v_j) 添加到边集 E 中(第 5-9 行)，直到遍历信息级联的集合 $\boldsymbol{C}$，返回边集 E 中 $W(i,j)$ 最大的 k 条边，算法运行结束。举例说明：信息级联 $\{(v_1,1.0,m_1),(v_2,1.2,m_2),(v_3,1.3,m_3)\}$ 的候选边集为：$\{(v_1,v_2),(v_1,v_3),(v_2,v_3)\}$，分别计算 $w_c(1,2)$, $w_c(1,3)$ 和 $w_c(2,3)$，并根据候选边是否存在于 E 中更新 $W(1,2)$, $W(1,3)$ 和 $W(2,3)$。

算法 1. SFEN-Inf 算法.

输入：信息级联的集合 $\boldsymbol{C}$ 和边的数量 k

输出：信息传播网络 $G^{\wedge}$

1: $E \leftarrow \Phi$
2: FOREACH c IN $\boldsymbol{C}$ DO
3: FOREACH (v_i,v_j) IN c AND $t_i < t_j$
4: Compute $w_c(i,j)$
5: IF (v_i,v_j) NOT IN E
6: $W(i,j) = w_c(i,j)$
7: $E \leftarrow (v_i,v_j)$
8: ELSE
9: $W(i,j)\ +=\ w_c(i,j)$
10: END IF
11: END FOR
12: END FOR
13: RETURN Top k $W(i,j)$ IN E As k Edges

算法时间复杂度分析：对于给定的信息级联 c，我们只考虑 $t_i < t_j$ 的情况，假设 c 由 n_c 个节点组成，那么对于给定的 c，算法的候选边有：

$$1+2+...+(n_c-1)=O(\frac{n_c^2}{2})$$

因此，除去计算文本相似性的时间外，SFEN-Inf 算法的时间复杂度为：

$$O(|\boldsymbol{C}|\times\bar{n}^2)$$

其中 $\bar{n}$ 是所有信息级联 $\boldsymbol{C}$ 的平均节点个数。

4 实验

4.1 实验概述

我们在 Mac(Intel Core i7,8 GB Mem)笔记本上进行实验，数据集是新浪微博四个不同的典型事件，实验评估指标是准确率(**Precision**)、召回率(**Recall**)和 **F1-Score**，实验跟 NetInf 算法对比。为简单起见，我们的模型参数设置为：$\alpha=\beta=\delta=1$。

NetInf 算法在合成数据集上 F1-Score 达到 0.9，但在真实数据集上 F1-Score 通常低于 0.3[10]。本文只在真实数据集上进行实验，因为提出的传播概率模型考虑了文本相似性，不容易在合成网络上模拟节点产生的文本。

通过实验，我们发现 NetInf 算法的 F1-Score 低于 0.2，但是 SFEN-Inf 算法的 F1-Score 超过 0.5，推理效果提高约 2 倍。而且 SFEN-Inf 算法运行时间小于 NetInf 算法(表 2)。

4.2 数据集说明

我们根据关键词爬取四个典型事件的微博，关键词分别是“北京、高校、食堂、死猪”,“李开复、癌”,“临武、瓜农”,“芦山、地震”，并过滤无关微博，得到如下数据：北京高校食堂死猪事件共 18244 条微博(2013.8-2013.9.22)，李开复癌事件共 143910 条微博(2013.8-2013.9.22)，城管打死临武瓜农事件共 990905 条微博(2013.7.17-2013.9.22)，芦山地震事件共 1516245 条微博(2013.8-2013.9.22)，四大事件数据总量约 22G(表 1)。

4.3 信息级联和 GroundTruth

4.3.1 信息级联

信息级联 c 是指用户发布一条原创微博后，其粉丝(包括粉丝的粉丝)对该微博进行转发或评论。信息级联 c 的每个元组 (v,t,m) 按时间递增排列。这样，一条原创微博对应一条信息级联，每个典型事件中大量的原创微博形成信息级联的集合 $\boldsymbol{C}$ (表 1)，作为 SFEN-Inf 算法的输入。图 3 为“芦山地震”事件中名为“中国国家地理”的用户的某条原创微博被转发形成信息级联的情况。

图 3 单条原创微博被转发形成的信息级联

表 1 新浪微博数据集说明

主题	原创微博	微博总量	信息级联
高校 死猪	978	18244	**239**(7155)
李开复 癌	26604	143910	**4333**(61941)
临武 瓜农	80542	990905	**12666**(384968)
芦山 地震	512165	5951404	**121721**(1801799)

表 1 中信息级联的数量(加黑)比原创微博少的原因是我们过滤掉了那些没有被任何用户转发的原创微博。

4.3.2 GroundTruth

真实的信息传播网络是潜在的，为了评价算法优劣，我们需要生成 GroundTruth。引言中提出，假设用户只能转发其关注用户发的帖子时，那么信息传播网络就是社交网络的子集。但是新浪微博运营商不对外公布用户完整的关注列表和粉丝列表，也就无法得到社交网络。我们通过爬虫分析微博转发结构生成 GroundTruth，具体做法如下：(1)选出某典型事件所有原创微博。(2)爬取每条原创微博的转发列表，形成转发树(如图 3)。(3)将所有原创微博的转发树合成该事件的信息传播网络。(文献[18]通过分析微博内容的“//@”标记来获取原创微博的转发网络并不准确，因为用户能自由添加或删除“//@”标记)。我们发现微博网络是稀疏网络，为简单起见，GroundTruth 的边数作为 SFEN-Inf 算法的输入 k 。

4.4 算法性能分析

我们将从网络推理效果和时间性能评价 SFEN-Inf 算法的优劣。

4.4.1 推理效果

图 4 和图 5 分别是“高校死猪”和“李开复癌症”事件的 F1-Score 图，图 6 和图 7 分别是“临武瓜农”和“芦山地震”事件的 Precision-Recall 图。因为文章篇幅的缘故，我们并未给出每个典型事件的两种图。

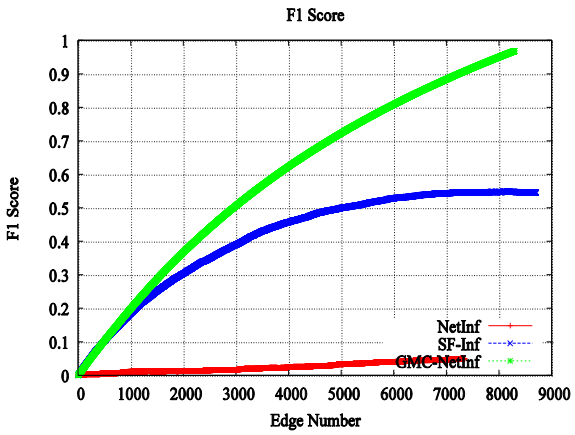

图 4 “北京高校食堂死猪”事件

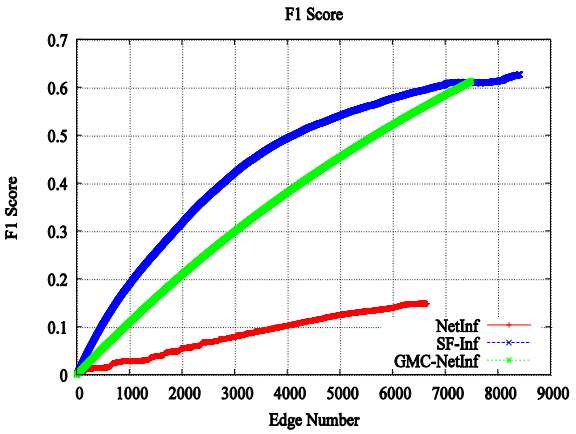

图 5 “李开复癌症”事件

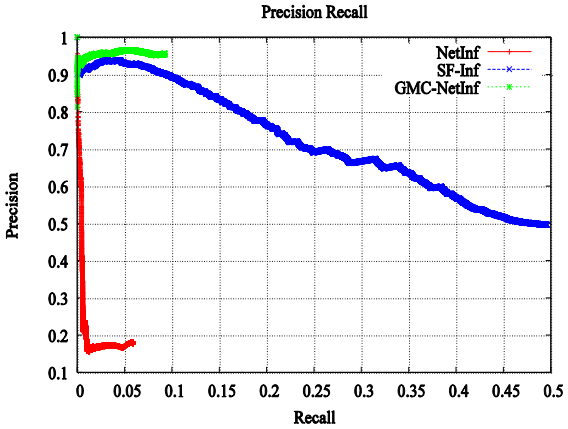

图 6 “城管打死临武瓜农”事件

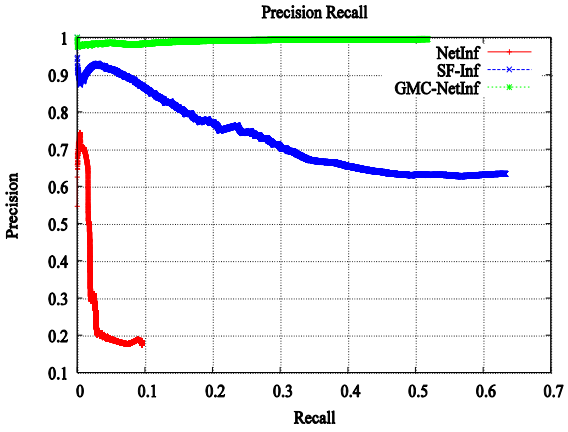

图 7 “芦山地震”事件

由实验结果可知，NetInf 算法的 F1-Score 低于 0.2，而 SFEN-Inf 算法的 F1-Score 超过 0.5，在算法性能上 SFEN-Inf 胜过 NetInf。

NetInf 算法在真实数据集上的准确率偏低，这是因为要使 NefInf 算法达到很好的效果，通常信息级联的数量要超过边的数量，现实中很难满足该条件。一条信息级联 c 对应一个传播树(图 2)，我们把从根到叶子的每条路径作为一条新的信息级联，那么从一条信息级联就衍生出若干条新的信息级联。例如，图 2 中的信息级联能衍生出 3 条新的信息级联，这样信息级联的数目增加，超过网络边的数量(表 1 括号中的

内容)。此时 NetInf 算法效果很好，F1-Score 超过 0.95，如同在合成网络上的性能(见图 4-图 7 中 GMC-NetInf：**G**enerate **M**ore **C**ascades NetInf)。但是，这种方法需要知道信息级联的准确结构，事实上本文的 GroundTruth 就是根据所有准确的信息级联的树结构合成而来。因此这部分的实验在于验证合成的 GroundTruth 的有效性。

针对“北京高校食堂死猪”事件，图 8 给出了 2.3 节提到的三种不同的时间参数模型对 F1-Score 的影响。实验表明，SFEN-Inf 算法的推理效果受时间参数模型的影响，但不明显。

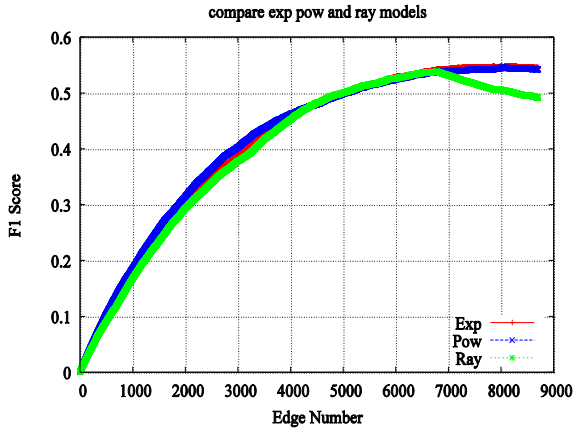

图 8 三种不同时间参数模型对比

4.4.2 时间性能

SFEN-Inf 算法因为网络推理模型简单，所以时间性能优于 NetInf 算法(表 2)，这使得 SFEN-Inf 算法能适应大规模网络的情况，应用前景广泛。

表 2 SFEN-Inf 和 NetInf 算法时间性能对比

主题	NetInf	SFEN-Inf
高校 死猪	4.38s	2.83s
李开复 癌	11m42s	21.72s
临武 瓜农	1h15m	5m22s
芦山 地震	48h20m	42m32s

4.5 网络可视化和结果分析

图 1、图 9 和图 10 分别是针对“高校死猪”事件、“李开复癌”事件和“临武瓜农”事件推理出的新浪微博传播网络。

信息通常从影响力较大的节点(新闻媒体，网络名人等)流向普通节点，但是“名人”和“平民百姓”之间联系未必最强。我们推理出的传播网络不仅能解释观察到的数据，重构信息传播路径，而且能发现影响力较大的节点和有频繁信息交互的人群。实际上，我们爬取的四个事件是当时的突发事件，通过本文的网络推理分析，我们能够更好的研究诸如“突发事件预警”，“网络安全”等重要课题。

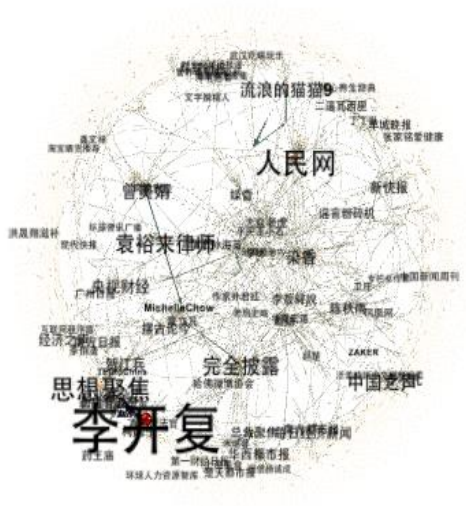

图 9 “李开复癌症”事件部分网络图

图 10 “城管打死临武瓜农”事件部分网络图

5 总结和展望

本文提出 SFEN-Inf 算法推理潜在的信息传播网络。通过实验，我们发现 SFEN-Inf 算法比著名的 NetInf 算法在推理效果和时间性能上都有较明显的提高。SFEN-Inf 算法能更好的适应大数据背景下的计算复杂性。通过将传播网络可视化，我们能直观的发现网络中的关键用户和牢固的用户关系。

在线社交网络有诸多特征，比如用户兴趣、线上线下互动、信用机制等，如何根据这些特征，建立符合社交特性本身的传播网络模型是我们的后续研究目标。

致谢 感谢黄海飞对此文的帮助！

参 考 文 献

[1] D.J. Watts and P.S. Dodds. Influentials, Networks,and Public Opinion[J]. Formation. Journal of Consumer Research, 2007, 34(4): 441-458

[2] Louis Lei Yu, Sitaram Asur, Bernardo A. Huberman. What Trends in Chinese Social Media[C] //Proceedings of the 5th SNA-KDD Workshop on Social Network Mining and Analysis(SNA-KDD'11). 2011

[3] Adrien Guille, Hakim Hacid, Cécile Favre, et al. Zighed. Information Diffusion in Online Social Networks: a Survey[C] //SIGMOD Record (SIGMOD), 2013, 42(2): 17-28

[4] David Kempe, Jon M. Kleinberg, Éva Tardos. Maximizing the Spread of Influence Through a Social Network[C] //Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'03).2003: 137-146

[5] Matthew Richardson, Pedro Domingos.Mining Knowledge-sharing Sites for Viral Marketing[C] //Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.2002: 61-70

[6] Jure Leskovec, Mary McGlohon, Christos Faloutsos, et al. Patterns of Cascading Behavior in Large Blog Graphs[C] //Proceedings of the Seventh SIAM International Conference on Data Mining(SDM'07). 2007: 551-556

[7] Jure Leskovec, Ajit Singh, Jon M. Kleinberg. Patterns of Influence in a Recommendation Network[C] //Proceedings of the 10th Pacific-Asia Conference Advances in Knowledge Discovery and Data Mining.2006: 380-389

[8] Daniel Gruhl, David Liben-Nowell, Ramanathan V. Guha, ek al. Information Diffusion Through Blogspace[C] //Proceedings of the 13th international conference on World Wide Web.2004:491-501

[9] Eytan Adar, Lada A.Adamic. Tracking Information Epidemics in Blogspace[C] //International Conference on Web Intelligence(WI'05). 2005: 207-214

[10] M.Gomez-Rodriguez, J.Leskovec, and A. Krause. Inferring Networks of Diusion and Influence[C] //Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'10). 2010: 1019-1028

[11] Seth A. Myers, Jure Leskovec. On the Convexity of Latent Social Network Inference[C] //Proceedings of the 24th Annual Conference on Neural Information Processing Systems(NIPS'10). 2010: 1741-1749

[12] Manuel Gomez-Rodriguez, David Balduzzi, Bernhard Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks[C] //Proceedings of the 28th International Conference on Machine Learning(ICML'11). 2011: 561-568

[13] Manuel Gomez-Rodriguez, Jure Leskovec, Bernhard Schölkopf. Structure and Dynamics of Information Pathways in Online Media[C] //Proceedings of the Sixth ACM International Conference on Web Search and Data Mining(WSDM'13). 2013: 23-32

[14] A.-L.Barabási. The Origin of Bursts and Heavy Tails in Human Dynamics[J]. Nature, 2005, 435: 207-211

[15] J. Leskovec, M. McGlohon, C. Faloutsos, et al. Cascading Behavior in Large Blog Graphs[C] //Proceedings of the Seventh SIAM International Conference on Data Mining(SDM'07). 2007: 26-28

[16] R. D. Malmgren, D. B. Stouffer, A. E. Motter,et al. A Poissonian Explanation for Heavy Tails in E-mail Communication[C] //Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(47): 18153–18158

[17] 王晶，朱珂，汪斌强. 基于信息数据分析的微博研究综述[J]. 计算机应用，2012,32(7)：2027-2029,2037

[18] Yaqiong Wang, Hongfu Liu, Hao Lin, et al. SEA: a System for Event Analysis on Chinese Tweets[C] //Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'13). 2013:1498-1501

郑众杰，男，1990 年生，研究生，研究领域：数据挖掘、社交网络

林学练，男，1978 年生，博士，讲师，研究领域：并行数据处理、中间件技术