一种基于通联数据的信息扩散路径推测算法

项英倬 魏 强 游 凌
(盲信号处理国家重点实验室 成都 610041)

摘 要 信息的传播和扩散对于研究市场营销、病毒木马的传播等具有重要意义。但是,在许多场景下仅能获取网络中用户的通联数据,难以获取用户间通信的内容。针对该问题,文中提出了一个基于概率的信息传播模型来对网络中的通联数据进行建模,以此估计网络中用户通信内容的相关性,进而推测网络中信息的扩散路径。文中证明了求解该模型的复杂度为 NP-hard,并提出了 PathMine 算法来获取模型的一个近似最优解。实验表明,所提 PathMine 算法能够高效地挖掘网络中信息的传播模式,优于已知的其他方法。

关键词 信息扩散,网络分析,信息流,子模函数

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/jsjkx.180901759

Information Diffusion Path Inferring Algorithm Based on Communication Data

XIANG Ying-zhuo WEI Qiang YOU Ling
(National Key Laboratory of Science and Technology on Blind Signal Processing, Chengdu 610041, China)

Abstract Information diffusion and propagation play an important role in viral marketing and virus diffusion. However, in many occasions only the connected data of users in the network can be obtained, which makes it difficult to obtain the content of communication between users. To deal with such challenges, this paper proposed an information diffusion model based on probability in order to predict the relativity of communications between users, and then infer the diffusion path of information in the network. In addition, this paper proved that the complexity of solving the model is NP-hard, and proposed PathMine algorithm to get a near optimal solution. Experiment results show that the proposed PathMine algorithm outperforms other state-of-art algorithms.

Keywords Information diffusion, Network analysis, Information flow, Submodule function

1 引言

网络中信息的传播、影响力的分析、木马和病毒的传播等是社交网络和信息网络中重要的研究领域[1-3],挖掘网络中的信息传播模式将有助于这些问题的分析。信息的传播模式指目标信息(木马、病毒、某个观点等)在网络中传播的模式,该模式难以通过观察得到。挖掘网络中信息的传播模式面临两个挑战:1)传播的信息内容是未知且无法获取的;2)背景通信流量占比远高于目标信息流量占比。在一般场景下,能够利用的数据只有通联双方和通信发生的时间。比如,在僵尸网络中,比较容易获取节点间的通信行为以及通信时间,但是,难以对节点间通信的负载进行分析。目前,国内外关于信息传播的研究大多是基于通信内容的方法[4-5]、通信内容与通信属性相结合的方法[6-8],或者基于复杂网络的一些方法[9-10]。这些方法在通信内容未知且通信属性只有时间时,难以发挥很好的效果。

本文所研究的场景比较特殊,即在仅已知网络中用户通联双方以及通联时间的情况下,推断用户间传递内容的相关性,进而分析用户间信息的传播路径与模式。通常情况下,容易观测到某用户与其他用户通信的时间序列,而通信内容未知,那么,如何通过这些通信数据分析网络中信息的传播方式,以及网络中什么样的信息传递模式能够解释这些观测的通联数据呢?

为了解决这些问题,本文采用生成模型对用户间通联数据的相关性和信息传播模式进行建模,并试图找到一个隐含的网络子图,使得通联数据的似然函数能够最大化。本文认为,网络中的指令、僵尸网络中的指控数据等信息的传播存在一个固定的模式,如果网络中的通联数据能够构成一个有向多边网络(Multi-edge Digraph),那么信息的传播模式将是该网络的一个有向子图。相关概念利用图 1 说明,图中所有的边构成一个有向图,边代表节点间发生过通信;而由粗边构成的子图代表网络中信息的传播模式。比如,在僵尸网络中,信息的传递模式可以理解为节点间的指控关系,而整个网络的边代表所观测到的通信行为。

到稿日期:2018-09-17 返修日期:2018-12-12 本文受国家自然科学基金(61174124)资助。

项英倬(1990-),男,博士生,主要研究方向为人工智能,E-mail:xiangyzh@foxmail.com;**魏 强**(1987-),男,工程师,主要研究方向为信息处理,E-mail:weiqianglg@163.com(通信作者);**游 凌**(1971-),男,研究员,主要研究方向为网络态势感知。

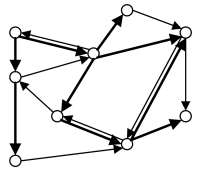


图 1　信息传递模式的说明

Fig. 1　Illustration of information diffusion model

本文第 2 节给出信息扩散路径推测问题的定义,并对问题进行建模、分析和求解;第 3 节通过人造数据对算法的性能进行对比评估,进而将算法应用于 Enron 邮件数据集中,分析 Enron 公司员工间信息传播的模式。

2　PathMin 算法

2.1　问题定义

一般来说,网络中用户间信息的传播遵循一定的模式,而这个模式通常可以表示用户间信息传递的关系。

定义 1　网络中用户间信息的传播模式指在通信网络 $G_o(V,E)$中代表用户间信息传播的一种固有模式。本节中用网络 $G(V,E)$代表用户间信息传播关系的网络,那么 $G(V,E)$是一个有向图,且是 $G_o(V,E)$的一个子图,代表用户间的信息更加倾向于沿着网络 $G(V,E)$中的边传播。网络 $G(V,E)$中的每条边称为信息传递边,而网络 $G_o(V,E)$的非信息传递边称为普通边。

定义 2　网络信息扩散路径的推测问题指给定节点的集合 V,以及节点之间的通联数据 $A=\{(v_i,v_j,t_i)\mid v_i,v_j\in V, i\neq j\}$,推测节点间传递信息的相关性,并得到 $G(V,E)$。

我们容易得到通信网络 $G_o(V,E)$的拓扑,即只需要将每条通联数据当作一条边并将所有的边组合起来,但是用户间信息的传递模式 $G(V,E)$是不容易得到的。由于通信内容未知,为了得到网络中信息的传播模式,本文首先需要根据节点的通信行为推测通信内容的相关性,然后分析信息的传播模式。文献[11-13]中指出信息转发的时间间隔 Δt 满足指数分布或幂律分布,即 $p(\Delta t)\sim e^{-\alpha\Delta t}$ 或者 $p(\Delta t)\sim(\Delta t)^{-\alpha}$。如果用户在某次通信后短时间内又与其他用户通信,那么可以推测用户间传递的信息是相关的,这里使用 $f(t_{i,j}\mid t_{k,i})$来衡量传递信息的相关性。

上述场景中由于不包含通信内容,目前国内外解决该问题的主流方法是采用复杂网络的相关手段。比如,Altenburger 等[9]通过节点间的通信数量、节点的度等网络结构信息来衡量节点的相似度;Yang 等[10]指出社交网络中边的中介中心性以及边上的数据量对信息的传播至关重要。然而,这些方法和模型并没有考虑通联数据的时间属性,本文的贡献在于提出了信息传播模型,融合了节点间通联的时间属性,对网络中的信息传播模式进行了挖掘。

2.2　信息传播模型

为了方便后续数据处理,首先对通联数据 A 进行整理。

定义 3　用户的行为记录指用户接收信息与发送信息的行为所构成的记录,有 $R_{\text{records}}=\{R_i\mid i\in V\}$,其中:

$$R_i=\{\cdots r_x r_y\cdots s_p\cdots \mid r_x:=\text{recievetime of node } x\to i,\ s_p:=\text{sendtime of node } i\to p\} \tag{1}$$

然后定义一个最大观测时间窗口 T,并将每个用户的行为记录切成许多时间切片 ts。由于需要挖掘用户传递信息的行为,因此将用户收到其他用户发送给他信息的开始时间作为时间切片的起点,并只考虑用户的发送行为。这里认为用户在时间窗口外的发送行为与窗口内收到信息的时间间隔过长而不会存在信息传递的可能。图 2 给出了用户 i 的行为记录以及时间切片。在时间切片 r_x 与 r_x+T 之间只有两个发送行为,因此该时间切片可以表示为 $ts_{r_x,i}=\{r_x,s_u,s_v\}$,每个时间切片由两个下标来确定,第一个下标代表该时间切片开始的时间,第二个下标代表用户 i 的时间切片,并且每个时间切片的第一个数据为该用户接收信息的行为以及时刻,后续的数据全部为该用户的发送行为以及时刻。切割所有用户的行为记录,则将通联数据 A 转化为时间切片的集合 TS,容易证明,通联数据 A 与时间切片 TS 是可以互相转化的,时间切片只是将数据变换为另一种表现形式。

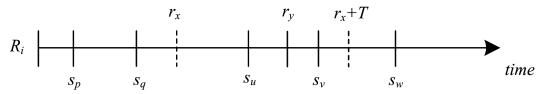


图 2　节点行为记录及时间切片示意图

Fig. 2　Illustration of nodes action records and time slices

在图 2 中,由于并不知道用户 i 收到来自用户 x 的消息后是否将消息转发给了用户 u 或者 v,因此需要根据不同情况进行讨论。如果由用户 x 指向用户 u 之间的边属于 $G(V,E)$,那么用户有较大的概率进行信息传播,这里使用 $f(t_{i,u}\mid t_{x,i})$表示这一概率;如果由用户 x 指向用户 v 之间的边不属于 $G(V,E)$,那么使用一个非常小的数值 ε 来表示这种情况发生的概率,一般来说该参数设置为 10^{-6} 即可。由于用户可能会将消息发送给 $G(V,E)$中的一个或多个子节点,因此用 β 表示用户未发送给其子节点的概率,该参数设置为 0.1 左右即可。对于不同的场景,ε 和 β 这两个参数可以根据实际情况自行设定。这样,可以得到一个时间切片的似然函数:

$$\begin{aligned} f(ts_{r_x,i}\mid G,G_o)&=\prod_{k\in Vout(i)}\varepsilon^m(1-\varepsilon)^n\beta^q f(t_{i,s_k}\mid t_{r_x,i})\\ &=\prod_{k\in Vout(i)} f'(t_{i,s_k}\mid t_{r_x,i},\varepsilon,\beta) \end{aligned} \tag{2}$$

其中,s_k 指用户 i 发送信息给用户 k 的时间;n 指时间切片中用户 i 发送信息给该用户子节点的通信数量;m 为用户 i 发送信息给非子节点的通信数量,并且 $m+n$ 等于用户 i 在 $G_o(V,E)$中的出度与 $G(V,E)$中出度的差;q 为用户未发送给该用户子节点的数量。图 3 给出了式(2)的一个说明。图中虚线表示 $G_o(V,E)$中的非信息传递边,实线表示 $G(V,E)$中的信息传递边,一般来说用户间信息传递模式 $G(V,E)$是未知的,可以发现,第一个潜在传递模式更容易产生右边的时间切片,因此由式(2)表征的似然函数要大一些。

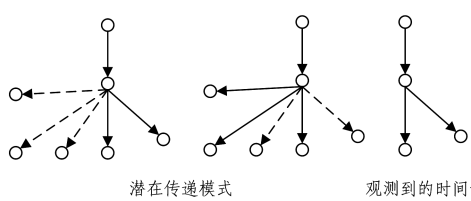


图 3　时间切片与潜在传播模式

Fig. 3　Time slices and potential diffusion models

考虑所有的时间切片集合 TS，可以得到全部观测数据的似然函数：

$$f(TS|G,G_o)=\prod_{i\in V,r_x\in R_i} f(ts_{r_x,i}|G,G_o) \tag{3}$$

这样便将通信网络中用户间信息传递模式挖掘问题转化为了如何找到 G 以使得式(3)最大化的问题。下面对式(3)的性质进行分析：首先，该公式是非负的，因为式(2)是非负的；其次，该公式是单调的，也就是说 $f(TS|G,G_o)<f(TS|G',G_o)$ 对于 $G(V,E)$ 以及 $G'(V,E')$，$E\subset E'$ 成立，因为根据式(2)，当把一条普通边转变为信息传递边后，似然概率中的一个 ε 变为了更大的概率值，所以向 $G(V,E)$ 中增加边后并不会降低解的性能，$G(V,E)=G_o(V,E)$ 应该是最佳的解。然而，一般来说，挖掘出的信息传播模式 $G(V,E)$ 较为稀疏，这样更能够表征原网络 $G_o(V,E)$ 中的一些特性，尤其是得到信息传播的骨干路径图，因此本文将网络 $G(V,E)$ 中边的数量限制为 k。重新改写通信网络中用户间信息传递模式挖掘问题为：

$$G'=\arg\max_{|G|\leqslant k} f(TS|G,G_o) \tag{4}$$

一般来说，采用暴力搜索的方法求解式(4)需要花费的时间随问题规模呈指数增长，下面证明求解式(4)是 NP-hard。

定理 1 由式(4)定义的通信网络中，用户间信息传递模式挖掘问题是 NP-hard。

证明：可以将该问题规约到 MAX-k-COVER 问题[14]中来。在 MAX-k-COVER 问题中，给定一个有限集合 W，$|W|=n$ 以及子集 $S_1,\cdots,S_m\subset W$。目标函数：

$$F_{MC}(A)=|\bigcup_{i\in A} S_i| \tag{5}$$

表示由 A 索引的子集覆盖 W 中元素的个数。MAX-k-COVER 的目标是选取 k 个子集使得 F_{MC} 最大化。假设有一个势为 n 的时间切片集合 TS 以及求解的目标网络 G，使得 $\max_{|G|\leqslant k} F_{TS}(G)=\max_{|A|\leqslant k} F_{MC}(A)$。网络 G 的节点为 $V=\{1,\cdots,m\}\cup\{r\}$，可以做一个双射 $(i,r)\leftrightarrow S_i$，如果将一个普通边转化为信息传递模式的边，$f(TS|G,G_o)$ 将会增加，因此可以做映射 $f(TS|G\cup(i,r))\leftrightarrow S_i$。如果在 G 中增加一条边 $(i,r)\in G$，则选择 S_i 进入 A，那么一个含有 k 条边的网络 G 等价于 MAX-k-COVER 问题中的一个解 A。因此，MAX-k-COVER 的每个解 A 都可以等效为由式(4)定义的问题的一个解 G。证毕。

定理 1 证明了通信网络中用户间信息传递模式挖掘问题为 NP-hard，因此难以在多项式时间内找到该问题的求解算法。针对这类问题，一般采用启发式搜索算法或者是求解一个近似最优解。下面研究该问题的求解算法。

2.3 PathMine 算法

为了求解式(4)，首先对其两边取对数，将似然函数 $f(TS|G,G_o)$ 转化为对数似然函数 $\log f(TS|G,G_o)$。更进一步，对等式右边做一个等价转化，将其改为优化增量最大化。首先假设一个空网络 K，其边全部是普通边。

$$G'=\arg\max_G(\log f(TS|G,G_o)-\log f(TS|K,G_o)) \tag{6}$$

将式(2)、式(3)代入式(6)得到：

$$G'=\arg\max_{i\in V,r_k\in R_i}\sum_{j\in Vout(i)}\log w(i,j) \tag{7}$$

其中，$w(i,j)=\varepsilon^{-1}f'(t_{i,s_j}|t_{r_x,i},\varepsilon,\beta)$。为了方便，定义：

$$F(TS|G)=\sum_{i\in V,r_k\in R_i}\sum_{j\in Vout(i)}\log w(i,j) \tag{8}$$

下面研究式(8)的一些性质。

定理 2 假设 V 为节点的集合，TS 为时间切片的集合，那么 $F(TS|G)$ 是子模函数(Submodular Function) $F:2^W\rightarrow\mathbb{R}$，其中 $W\subseteq V\times V$ 为有向边的集合。

证明：首先考虑单个时间切片 ts、网络 $G\subset G'$ 以及一条边 $e=(r,s)\notin G'$，假设 $w_{i,j}$ 是网络 G 中边 (i,j) 的权重，$w'_{i,j}$ 是 G' 中边 (i,j) 的权重。因为 $G\subset G'$，显然对于所有的边，有 $w'_{i,j}\geqslant w_{i,j}\geqslant 0$。如果 (i,j) 是网络 G 和 G' 的共同边，那么 $w_{i,j}=w'_{i,j}$。设 $M_{A,e}=\sum_{i\in A\setminus\{r\}} w(i,s)$，容易得到 $M_{G,e}\leqslant M_{G',e}$，因此有：

$$\begin{aligned}&F(ts|G\cup\{e\})-F(ts|G)\\&=\log(\frac{M_{G,e}+w(r,s)}{M_{G,e}})\\&\geqslant\log(\frac{M_{G',e}+w(r,s)}{M_{G',e}})\\&=F(ts|G'\cup\{e\})-F(ts|G')\end{aligned} \tag{9}$$

由于子模函数的非负线性组合仍然是子模函数，因此 $F(TS|G)=\sum F(ts|G)$ 也是子模函数。证毕。

一般来说，子模函数的最大化问题都是 NP-hard[14]，但是本文通过贪婪算法求得了一个近似最优解，从一个空图 K 开始，每次循环向图中添加一个可以使式(8)的增量最大的边，直到 K 里有 k 条边截止。根据上述思想，本文给出 PathMine 算法的伪代码，如算法 1 所示。

算法 1 PathMine Algorithm

```
Require: TS, k, G_o
G←K
for all ts∈TS
  S_ts←all possible subordinates
While |G|<k
  For all (i,j)∈G_o\G
    δ_{i,j}=0, M_{i,j}←∅
  For all ts:(i,j)∈ts
      If w(i,j)≥w(Parent_ts(j),j) then
          δ_{j,i}=δ_{j,i}+w(i,j)−w(Parent_ts(j),j)
      M_{i,j}←M_{i,j}∪{ts}
  (i*,j*)←arg max_(i,j) δ_{i,j}
  G←G∪{(i*,j*)}
  For all ts∈M_{i*,j*}
    Parent_ts(j*)←i*
  Return G
```

在代码的实现中，本文采用了局部更新以及延迟计算[15]的方法来加快迭代的速度。通常，每个时间切片并不是特别大，仅涉及网络中的几个点，如果一次循环中 PathMine 算法选择的边不涉及该时间切片中的任何一个节点，那么该时间切片对下一个循环的增量贡献为零，因此在计算下一次的循环过程中，仅需要计算有影响的时间切片即可，这样可以大幅减少每次循环中的计算量。而延迟更新的思想则是在每次循环过程中，优先选择较大增量的边进行比较。假设 G_1，$G_2,\cdots,G_k$ 是贪心算法每次循环后依次得到的网络，令 $\Delta_e(G_i)=F(TS|G_i\cup\{e\})-F(TS|G_i)$ 代表第 i 次循环中将边 e 添加到

网络后得到的增量。根据 $F(\cdot|G)$ 的子模性，当 $i\leqslant j$ 时，有 $\Delta_e(G_i)\geqslant\Delta_e(G_j)$，因此在每次循环中，边 e 的增量只能是单调递减的，这意味着在一次循环中增量较小的边不可能在下一次的循环中增量突然变大。这样，每次循环都维护一个增量的序列，每次循环时，算法可以优先从最大增量的边开始；由于在新的循环中该增量可能会降低，因此只要重新计算其增量即可，并将改变重新插入到增量序列中；如果其增量不变，那么就选取这条边作为本次循环的最优边。

PathMine 算法比较容易进行并行化处理，因此进一步提升了计算速度以及对大规模网络的适应性。比如，在初始化计算每个时间切片的似然函数以及更新边的增量时，可以进行大规模的并行处理，从而加快算法的速度。

3　实验分析

3.1　模拟数据仿真

本节使用模拟数据对算法的性能进行分析，并将其与基础方法、基于静态图的算法进行对比。为了生成模拟数据，首先确定一个节点的集合 V，然后使用 KroneckerGraph[16-17] 来生成真实的有向网络结构，用户之间的信息将在网络的有向边上传播。该网络结构通常代表用户间信息的传递模式。本文考虑了随机图[18]（Kronecker 参数矩阵为[0.5，0.5；0.5，0.5]）、层次社区结构[19]（[0.962，0.107；0.107，0.962]）以及随机幂律随机树[20] 3 种不同的网络结构。实验中随机选择网络中的一个节点作为信息传递的起始节点，并为每条边设置一个转发概率来控制信息传播的广度，然后设置一个转发时间间隔参数来控制用户转发信息的速度。这样可以模拟生成一系列用户传递信息的数据，并保证网络中每条边尽可能参与一次信息的传播。接着，本实验生成大量的随机通联数据来模拟网络中用户的背景通信数据。为了使模拟数据更加符合真实数据中低信噪比的情况，这里用户间信息传递数据与背景通信数据的比值一般小于 0.01，本文使用信噪比（SN）来表示这一个指标。例如，本文实验模拟生成了一个具有 64 个节点以及 75 条有向边的层次型社交网络，然后模拟用户的信息传递行为，生成了 180 条信息传递记录以及 12000 条背景通信数据，使得层次社交网络中全部的边均至少参与一次信息的传递。

实验中用于比较的算法包括基于静态图的方法[9-10] 和基本方法。基于静态图方法的基本思想是根据通联数据中连接出现的频次、网络通联性等得分最高的 k 个边作为挖掘结果；基础方法的思想是采用 $p(u,v)=\sum_{ts\in TS}P_{ts}(u,v)$ 作为边的权重，即信息由边 u 传给边 v 的似然程度，并选取权重最大的 k 个边作为挖掘结果。

为了评估算法的性能，本文通过对比挖掘出的信息传递模式图 G 与模拟生成的网络 G' 来衡量结果的好坏。本文定义查全率（Recall）为挖掘出的网络中的边在真实网络 G' 中的比例；准确率（Precision）为挖掘出的边是正确的比例。本文希望的结果是在高准确率的基础上尽可能提高查全率，也就是说，如果挖掘出一小部分边，希望这些边尽可能都在真实的网络 G' 中，而随着挖掘出的边的增多，准确率会逐渐下降，而查全率单调升高。

图 4 给出了几种算法的仿真实验结果。可以看出，当信噪比为 0.03 时，3 种算法的性能均优于信噪比为 0.004 的性能，说明信噪比对算法性能的影响还是比较明显的。在 3 种算法中，无论是不同的网络结构类型还是不同的信噪比，基于静态图的方法的性能都是最差的，而且当信噪比进一步降低到 0.004 之后，该算法在随机图网络中已经失效而无法挖掘出可信的结果。这说明对于具有时序信息数据的挖掘，传统的基于复杂网络指标的挖掘算法难以挖掘出用户的信息传递模式。通联频率高的两个节点之间的边未必是对信息传递影响最大的边，这与普遍认为通联频率高、数据量大的线路最重要有一定出入。这些通联频率高的边可能在某种程度上说明这条线路的两个端点之间有大量的数据要传输，然而对于网络中用户间信息的传递来说，这条边可能并不重要。这说明在不同的网络结构中，有的边可能需要承载大量的数据，然而这些数据以背景通联数据为主；而用户间信息的传递可能未必会通过该边进行传播，用户间信息的传递模式网络与观测到的通联数据网络存在很大差异。

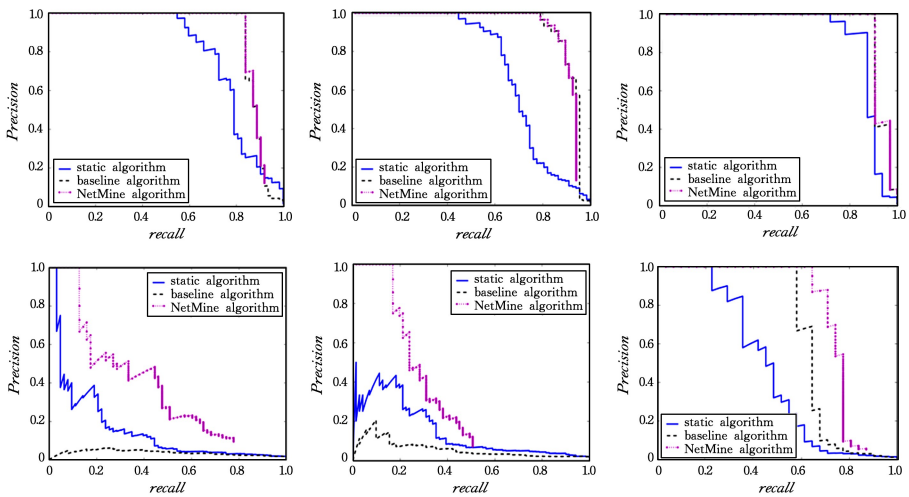


图 4　算法性能的对比

Fig. 4　Comparison of algorithms

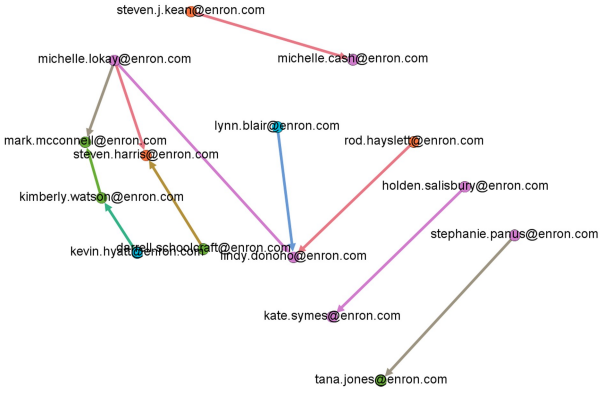
本文提出的 PathMine 算法在信噪比为 0.03 时与基础方法的性能差别不大，且均能达到不错的效果。这两种算法对于不同的网络结构均能够保证查全率为 0.9 左右时准确率达到 0.95 以上，这说明所提算法具有良好的可靠性以及稳定性。而当信噪比降到 0.004 时，本文提出的基本算法已经基本失效，只在层次社交网络结构中还有一定性能；而 PathMine 算法在这 3 种网络结构中均可以取得最好的效果。

一般来说，如果能够获取更多的用户间信息的传递数据，则能够显著提升算法的性能。而本实验仅模拟生成了 180 条用户间信息的传递数据，这说明在有效数据量较小的情况下，PathMine 算法仍可以很好地挖掘出用户间信息传递的模式。

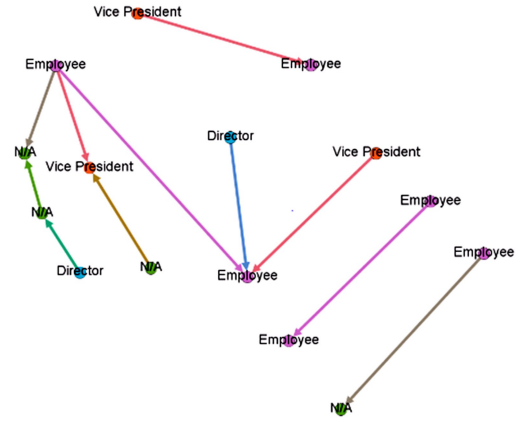
从不同的网络结构看，所提算法对层次社交网络结构的挖掘效果最好，而对随机图和随机树这两种随机网络的挖掘效果稍差。本文认为一个可能的原因在于随机网络中模拟用户传递信息的通联数据与随机数据过于类似，从而导致算法难以获取有用信息。尽管本文提出的 PathMine 算法的性能在不同的网络结构下有一定差异，但是算法性能仍然相对比较稳定，这进一步说明了 PathMine 算法的可靠性。

3.2 Enron 邮件集数据实验

安然邮件集[21]是安然公司几千名员工办公邮箱中的邮件数据集合，最初由联邦能源局公开，由卡内基梅隆大学的 WilliamCohen 收集并用于科学研究[22]。本文使用了其中一个含有 151 名标注了员工岗位职级的版本。由于算法仅需要邮件通信的双方以及时间，因此本文舍弃了邮件的内容，仅提取了邮件的发送者、接收者以及邮件发送时间，然后将这些数据存入 MySQL 数据库中。本文将有效的邮件数据输入到 PathMine 算法中，并设置算法的参数 k 为 11，即仅挖掘由 11 条边构成的信息传递网络，结果如图 5 所示。



(a)用户间信息传递模式图



(b)用户职级图

图 5 安然邮件集中用户信息传递模式图与职级图

Fig. 5 Information diffusion patterns and quarters of employees in Enron company

从图 5 中可以看出，安然公司员工间信息的传递模式大部分是由高级职位的员工传向低级职位的员工，这可以看作是一种指令类信息的传播过程，由高层向低级别员工传递信息。但还有其他情况，比如高管 StevenHarris 属于信息的汇聚方，信息由下属流向他。本文通过研究其邮件内容和标题发现，这名高管几乎所有的邮件都是在回复他人，很少主动发送邮件，也就是说，对于这名高管，通常是员工将信息汇总到他这里，然后他再进行相应处理，最后转发出去。因此，这是一个信息汇聚的传播模式。

从图 5 中看出，网络中用户间信息传递的模式中每条信息的传递路径并不是特别长，电子信息技术的发展拉近了人与人之间的距离，从高管到员工需要转发的次数并不多。

结束语 本文研究了通信网络中用户间信息传递模式的挖掘问题，并提出了 PathMine 算法对用户间信息的传递模式进行挖掘。通过仿真实验发现，PathMine 算法能够较好地挖掘出相应的信息传递模式，而且仅利用了通信发生的时间这一属性就取得了出乎意料的结果。通过与基于静态网络的挖掘算法进行对比发现，网络中流量大的边在信息的传播中所起的作用未必很大，这对于研究网络中信息的传播、市场营销等具有很强的指导意义。本文还将算法应用于安然邮件集合中，分析了安然公司中员工间信息的传递模式。

后续可以在运行速度方面对算法进行进一步优化，使其能够更快地处理海量数据。

参 考 文 献

[1] LAZARSFELD P F, KATZ E, ROPER E. Personal influence: The part played by people in the flow of mass communications [M]. New York: Routledge, 2017.

[2] ROGERS E M. Diffusion of innovations[M]. Simon and Schuster, 2010.

[3] WATTS D J, DODDS P S. Influentials, networks, and public opinion formation[J]. Journal of Consumer Research, 2007, 34(4): 441-458.

[4] MCCALLUM A, WANG X, CORRADA-EMMANUEL A. Topic and role discovery in social networks with experiments on enron and academic email[J]. Journal of Artificial Intelligence Research, 2007, 30: 249-272.

[5] ZHANG Y,WU Y,YANG Q. Community Discovery in Twitter Based on User Interests[J]. Journal of Computational Information Systems,2012,8(3):991-1000.

[6] FEI H,JIANG R,YANG Y,et al. Content based social behavior prediction:a multi-task learning approach[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM,2011:995-1000.

[7] LIN C,MEI Q,JIANG Y,et al. Inferring the diffusion and evolution of topics in social communities[C]// 11th IEEE International Conference on Data Mining(ICDM 2011). Vancouver:IEEE,2011.

[8] ZHU J,XIONG F,PIAO D,et al. Statistically modeling the effectiveness of disaster information in social media[C]// Global Humanitarian Technology Conference (GHTC). IEEE, 2011:431-436.

[9] ALTENBURGER,KRISTEN M,JOHAN U. Monophily in social networks introduces similarity among friends-of-friends[J]. Nature Human Behaviour,2018,2(4):284.

[10] YANG M,HSU W H,KALLUMADI S T. Predictive Analytics of Social Networks: A Survey of Tasks and Techniques[J]. Journal of Applied Physics,2014,112(1):014301-014301-5.

[11] ALBERT-LÁSZLÓ B. The origin of bursts and heavy tails in human dynamics[J]. Nature,2005,435(7039):207.

[12] MALMGRENR D,STOUFFER D B,MOTTER A E,et al. A Poissonian explanation for heavy tails in e-mail communication [J]. Proceedings of the National Academy of Sciences,2008,105(47):18153-18158.

[13] LESKOVEC J,MCGLOHON M,FALOUTSOS C,et al. Cascading behavior in large blog graphs: Patterns and a model[J]. arXiv:0704.2803.

[14] KHULLER S,MOSS A,NAOR J. The budgeted maximum coverage problem[J]. Information Processing Letters,1991,70(1):39-45.

[15] LESKOVEC J,KRAUSE A,GUESTRIN C,et al. Cost-effective outbreak detection in networks[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2007:420-429.

[16] GOYAL A,BONCHI F,LAKSHMANAN L V S. Learning influence probabilities in social networks[C]// Proceedings of the third ACM International Conference on Web Search and Data Mining. ACM,2010:241-250.

[17] LESKOVEC J,FALOUTSOS C. Scalable modeling of real graphs using kronecker multiplication[C]// Proceedings of the 24th International Conference on Machine Learning. ACM, 2007:497-504.

[18] ERDS P,RÉNYI A. On the evolution of random graphs [J]. Transactions of the American Mathematical Society,2011, 286(1):257-274.

[19] CLAUSET A,MOORE C,NEWMAN M E J. Hierarchical structure and the prediction of missing links in networks[J]. Nature,2008,453(7191):98.

[20] GROVER A,ZWEIG A,ERMON S. Graphite:Iterative generative modeling of graphs[J]. arXiv:1803.10459,2018.

[21] SHETTY J,ADIBI J. The Enron email dataset database schema and brief statistical report[R]. Information Sciences Institute Technical Report,University of Southern California,2004:120-128.

[22] COHEN W W. Enron Email Data Set[OL]. http://www.cs.cmu.edu/~enron/.