件概率独立的定义.

定义 **1**. 设变量集 $U=\{a,b,\ldots\}$的概率分布为 D,且集合 $X,Y,Z\subset U$.X 和 Y 在分布 D 上关于 Z 条件概率独立记作 $I\langle X,Z,Y\rangle$,其中,$I\langle X,Z,Y\rangle$是指对于 X,Y,Z 的任意状态取值 x,y,z,都有 $p(x,y|z)=p(y|z)p(x|z)$.特别地,当 $Z=\varnothing$时,$I\langle X,Z,Y\rangle$退化为概率独立 $I\langle X,Y\rangle$.

定义 1 描述变量间的相对独立性.相对独立性的物理含义反映了排除条件变量(Z)的中介作用后,测试变量(X,Y)的相关性,即反映了测试变量的直接相关性.所以,我们可以通过 $I\langle X,Z,Y\rangle$的真值来判断例 1 中 X 和 Y 是否直接相关来获得其间结构化关联:$I\langle X,Z,Y\rangle$为真,即说明 $X\sim Y$ 为真.由此可见,条件概率独立相对于概率独立具有更丰富的语义,可描述团体影响关联蕴含结构的信息.

根据信息论理论,条件互信息熵常用来定量描述变量的条件独立程度.

定义 **2**. X 和 Y 关于 Z 的条件互信息熵记为 $Inf((X,Y)|Z)$,其中,

$$Inf((X,Y)\mid Z)=\sum_{X,Y,Z\text{的所有状态取值}}p(x,y,z)\log_2\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)} \tag{3}$$

$Inf((X,Y)|Z)$表示 Z 确定之后,Y 与 X 的互信息熵.$Inf((X,Y)|Z)=0$ 表示 $I\langle X,Z,Y\rangle$,且其值越大,说明 X 和 Y 关于 Z 的直接关联性越强.特别指出:当 $Z=\varnothing$时,$Inf((X,Y)|Z)$退化为互信息熵 $Inf(X,Y)$.

根据以上论述,团体直接影响关联的形式化定义如下.

定义 **3**. 设 M 为社会网络 $G(V,E)$上的团体集,团体 $m_i,m_j\in M,m_i$ 对应的点集记为 $m_i(x)$,且有 $V=\bigcup_i^{|M|}m_i(x)$,$m_i(x)\cap m_j(x)=\varnothing$(为简化问题,仅考虑团体不重叠的情况).$m_i$ 和 m_j 存在直接影响关联,当且仅当 $ind(m_i,m_j)>0$.其中,$ind(m_i,m_j)=\begin{cases}w_{ij}, & w_{ij}>\varepsilon\\ 0, & w_{ij}\le\varepsilon\end{cases}$,$w_{ij}=Inf((m_i(x),m_j(x))|(V-(m_i(x),m_j(x))))$,$\varepsilon$为给定阈值.

我们可以根据定义 3 构建 IG.然而,这种直观的构建方法不能运用于较大规模的网络.其原因是:定义 3 中,w_{ij} 的计算需考虑集合 $m_i(x),m_j(x),V-\{m_i(x),m_j(x)\}$的所有状态取值;集合的状态取值数是该集合大小的指数倍,例如,当网络点数量$|V|=n,|m_i(x)|+|m_j(x)|=a$ 时,$V-\{m_i(x),m_j(x)\}$的状态取值就有 2^{n-a} 种,由此可见,w_{ij} 的计算复杂度为 $O(2^n)$.为此,我们将利用同质性假设条件,通过建立团体关联和单点关联的关系,给出一种 w_{ij} 的快速计算方法.在介绍我们的方法之前,我们首先引入条件概率独立相关的数学性质来说明方法的思路.

引理 **1**[19]. 设 X,Y,Z,Q 是分布 D 上两两不相交的变量集,则 $I\langle X,Z,Y\rangle$满足:

(1) 对称律:$I\langle X,Z,Y\rangle\Rightarrow I\langle Y,ZQ,X\rangle$.

(2) 分解律:$I\langle X,Z,Y\rangle\wedge I\langle X,Z,Q\rangle\Leftrightarrow I\langle X,Z,Y\cup Q\rangle$.

在同质性假设下,团体和点的直接影响关联存在如下关系.

定理 **1**. $M=\{m_1,\ldots,m_{|M|}\}$为团体集,$X=\{x_1,\ldots,x_{|M|}\}$为点集,且 $x_i\in m_i(x),i\in 1,\ldots,|M|$.在同质性假设下,在历史观测数据 H 上有 $ind(x_i,x_j)=0$,当且仅当 $ind(m_i,m_j)=0$.

证明:设 $m_i(x)=x_1\cup\ldots\cup x_i$, $m_j(x)=x'_1\cup\ldots\cup x'_j$, $\Upsilon=V-\{m_i(x),m_j(x)\}$.由定义 2 和定义 3 可知,$ind(x_i,x_j)=0$ 当且仅当 $I\langle x_i,\Upsilon,x_j\rangle$,所以原命题即证 $I\langle x_i,\Upsilon,x_j\rangle\Leftrightarrow I\langle m_i(x),\Upsilon,m_j(x)\rangle$成立.显然,在同质性假设下,有 $I\langle x_i,\Upsilon,x_j\rangle\Leftrightarrow I\langle\forall x_i,\Upsilon,\forall x'_j\rangle$ 成立.对 $I\langle\forall x_i,\Upsilon,\forall x'_j\rangle$ 中的$\forall x'$运用分解律有 $I\langle x_i,\Upsilon,x'_1\rangle\wedge\ldots\wedge I\langle x_i,\Upsilon,x'_j\rangle\Leftrightarrow I\langle x_i,\Upsilon,m_j(x)\rangle$ 成立,即 $I\langle m_i,\Upsilon,m_j(x)\rangle$.根据对称律再次运用分解律,同理可证 $I\langle m_i(x),\Upsilon,m_j(x)\rangle$,所以 $I\langle x_i,\Upsilon,x_j\rangle$为 $I\langle m_i(x),\Upsilon,m_j(x)\rangle$的充要条件. □

定理 1 说明点集 X 和团体 M 的关联图同构,所以我们可使用如下的办法快速构建 $IG(M,I,W)$.

- 给定一个以团体为节点(为与社会网络 G 区别,此处将 IG 图中的点称为节点)的完全图 $IG^*(M,I,W)$,如果在分布 D 上有 $Inf(x_i,x_j)\le\varepsilon$(其中,$x_i\in m_i(x)$),说明 m_i,m_j 独立(不存在关联),则直接删去边 $I_{i,j}$;
- 否则,需根据条件概率独立进一步判断关联类型:如果 $ind(x_i,x_j)=0$,则说明在 m_i,m_j 不存在直接关联,删掉该边;如果 $ind(x_i,x_j)>0$,则说明在 m_i,m_j 存在直接关联,将 $I_{i,j}$ 的权值设置为 $w_{i,j}=ind(x_i,x_j)$.

当所有无关联的边被删除后,$IG^*(M,I,W)$即同构于 $IG(M,I,W)$.显然,在该过程中,我们最多需要$|M|(|M|-1)/2$ 次条件独立计算,且每次计算的复杂性为 $2^{|M|}$.所以在 M 远小于 n 的情况下,该方法可被看作线性时间.算法 1 用伪码的形式清晰地描述了上述过程.

算法 **1**. 团体关联图构建.

输入:团体集 M“感染”的历史观测数据 D.

输出:团体关联图 $IG(M,I,W)$.

1: 初始化图 $IG^*(M,I,W)$为无向完全图;$i\leftarrow 0$

2: **while** $\forall x_i\in m_i$ **and** $m_i.visited$=**false**

3: $m_i.visited\leftarrow$**true**

4: **for each** m_i **and** m_j.visited=**false** //根据条件独立关系删边

If $Inf(x_i,x_j)\geq\varepsilon$

5: **if** $w\leftarrow ind((x_i,x_j)|X-(x_i,x_j))>0$ //$X=\{x_1,\ldots,x_{|M|}\},x_j\in m_j$

6: $w_{ij}\leftarrow w$

7: **else**

8: $I\leftarrow I-e_{ij}$

else

$I\leftarrow I-e_{ij}$

9: i++

10: **return** $IG^*(M,I,W)$

2.4 GIC传递规则

关联程度越高的团体,其间点发生影响传递的可能性越高.根据此特征,本节将基于 IG 给出一种粗粒度的影响传递规则.

给定团体关联图 $IG(M,I,W)$,记团体 $m\in M$ 的邻居集为 $N(m)$.$N(m)$将对 m 产生影响,且如果团体 $u,u'\in N(m)$有 $w_{u,m}>w_{u',m}$,那么点 $x_i\in u(x)$比 $x_j\in u'(x)$有更大可能激活 $x_m\in m$.自然地,我们可以用 $p_{u,m}=\lambda\times w_{u,m}\big/\sum_{k\in N(m)}w_{k,m}$ 来表示 x_i 对 x_m 的影响概率,其中,$\lambda\in[0,1]$为设定激活因子,用以调节影响大小.又由 IC 模型可知,“未感染”的点只受“感染”点影响,且已被“感染”点状态不会发生改变.设 u 的“感染”比例为 η_u,显然,当 $p_{u,m}$ 不变,且 η_u 值越大而 $1-\eta_m$ 值越小时,u 将有更多机会影响 m.所以,$u\rightarrow m$ 的期望传递影响比例 $\eta_{u\rightarrow m}$ 可表示为

$$\eta_{u\rightarrow m}=\eta_u\times p_{u,m}\times(1-\eta_m) \tag{4}$$

特别指出,当 u 中没有点“感染”($\eta_u=0$)或 m 中点已全部被“感染”($1-\eta_m=0$)时,u 对 m 的传递影响比例为 0.

根据公式(4),GIC 模型的动态传递规则如下.

GIC 将影响传播过程划分为 $t(t=0,1,2,\ldots)$个离散时间片进行模拟.为在 G 上激活一次传播,$t=0$ 时刻,我们向种子团体集 S 以广播的方式散布“疾病”,并假设 $m_i\in S$ 以固定比例 R_i“感染”(很多现实场景下,R_i 是可以获得的,例如用户的点击率可被预测).当传播被激活后,S 在 $t>0$ 时刻将迭代的把影响传递给其邻居,乃至其邻居的邻居.IC 模型约定:t 时刻的“感染”点仅在 $t+1$ 时刻具有“传染”性.类似地,在动态模拟团体影响传播过程中,我们将记录团体 m_i 在 t 时刻“感染”的比例 $\eta_i^t(\eta_i^0=R_i)$,并根据 η_i^t 计算团体间的传递影响.设 m_i 的邻居为 $N(m_i)$,我们将分 $|N(m_i)|=1$,$|N(m_i)|>1$ 两种情况给出 η_i^{t+1} 的计算式:当 $N(m_i)=\{u_j\}$时,根据公式(4),直接有 $\eta_i^{t+1}=\eta_{j\rightarrow i}^{t+1}=\eta_j^t\times p_{j,i}\times(1-\eta_i^t)$;当 $N_{in}(m_i)=\{u_j,u_k\}$时,由以上规则可知,u_k 可能重叠激活 m_i 中已被 u_j 部分.为避免重叠激活的部分的累加,我们将 u_k 对 m_i 的影响比例表示为 $\eta_{k\rightarrow i}^{t+1}=\eta_k^t\times p_{k,i}\times(1-\eta_j^t-\eta_{j\rightarrow i}^{t+1})$.类似地,我们可以计算 $N(m_i)$中任意团体对 m_i 的期望影响比例,并对其求和得到 $t+1$ 时刻 m_i 获得的总影响比例 $\eta_i^{t+1}=\sum_{u_j\in N(m_i)}\eta_{j\rightarrow i}^{t+1}$.迭代重复这一过程,直至对任意团体 m_d 都有 $\eta_d^t\times|m_d|<1$(即没有点再被“感染”,其中,$|m_d|$表示 m_d 包含的点数)时停止.特别指出:由于团体激活过程可能成环,为简化问题,我们忽略团体间的回传影响[20],即若在 $t=\alpha$时刻 $u_j\rightarrow m_i$,那么在 $t\geq\alpha$时刻 $u_j\nleftarrow m_i$.

例 2:在图 2 中,设 $S=\{a\}$,$\lambda=1$,且 $\eta_a^0=R_a=30\%$.

- 当 $t=1$ 时,a 将影响邻居 b,c.其中,

$$\eta_b^1=\eta_{a\to b}^1=\eta_a^0\times p_{a,b}\times(1-\eta_b^0)=30\%\times\frac{3}{3+3+4}\times(1-0)=9\%,$$

$$\eta_c^1=\eta_{a\to c}^1=30\%\times\frac{1.5}{1.5+4.5}\times(1-0)=7.5\%.$$

- 在 t=2 时刻,b,c 将共同影响 d,有 $\eta_d^2=\eta_{b\to d}^2+\eta_c^1\times p_{c\to d}\times(1-\eta_d^1-\eta_{b\to d}^2)\approx 7.8\%$.

可以看出,GIC 是 IC 模型的扩展,其主要区别有两点.

1) IC 模型中点的状态为布尔取值,最多能够被 1 个邻居影响,而 GIC 模型将节点(团体)状态扩展为连续取值(比例),能够被多个邻居共同影响.
2) IC 模型中点间的影响的用概率表示,而 GIC 模型中节点间的影响用激活点数的期望比例表示.

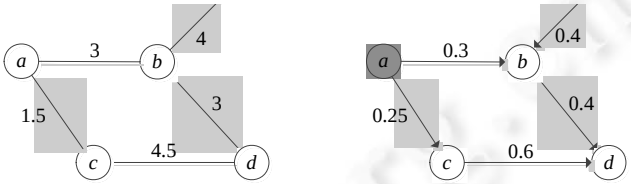


Fig.2 Example of influence diffusion

图 2 影响传播示例

3 团体影响最大化算法

基于 GIC 模型,本节从算法角度定义了团体影响最大化问题,并给出了一种贪心算法快速求解该问题.

3.1 问题定义

给定网络 G 上的团体关联图 $IG(M,I,W)$,整数 k 和团体的初始"感染"率 R_i 作为输入,根据 GIC 模型的模拟传播结果,在 IG 中搜索 k 个团体 $S\subset M$ 作为种子集,使其期望的影响范围$\sigma(S)$最大.

定理 **2**. 团体影响最大化是 NP-hard 问题.

证明:已知点影响最大化问题是 NP-hard.当网络中的每个团体仅包含 1 个点时,团体影响最大化即退化为点影响最大化问题,所以团体影响最大化是 NP-hard 问题. □

由定理 2 可知,在 $P\neq$NP 的假设下,不存在多项式时间的算法能够得到团体最大化问题的精确解.

3.2 函数子模性与贪心选择

定义 **4**(子模性)[21]. F 是定义域为集合 U 的函数,给定 S_1,S_2 为 U 上的 2 个子集.F 具有子模性,如果 F 满足:

$$F(S_1\cup\{u\})-F(S_1)\geq F(S_2\cup\{u\})-F(S_2),$$

其中,$u\subset U$,$S_1\subseteq S_2\subseteq U$.

引理 **2**[21]. 如果一个最优问题的优化目标(函数)F 满足单调性和子模性,那么使用贪心策略求解该问题所获得的结果能够保证返回(1−1/e)的最优.

引理 2 为近似求解特殊优化问题提供了理论支持.本节通过证明团体最大化问题的影响函数$\sigma(S)$满足单调、子模性,给出了一种具有保证的贪心算法.首先,我们基于 GIC 模型给出$\sigma(S)$的函数表达式.

在 GIC 模型中,$\forall m_j\in V/S$ 能够被 S 影响,当且仅当 S 到 m_j(在 IG 中)至少存在 1 条轨(如果路径 $u\to v$ 中的顶点各不相同,则该路径称为 u 到 v 的一条轨).记$\Gamma=\{\Gamma^1,\Gamma^2,\ldots\}$为 S 到 m_j 的轨集,$\Gamma^q=\langle m_i=w_1,w_2,\ldots,w_m=m_j\rangle$表示$\Gamma$中第 q 条轨,其中,$w_1\in S$ 且 $w_u\notin S$,$2\leq u\leq m$.由此,我们可基于轨给出一种特殊的树结构(简称为 Tr)来辅助分析 m_j 受 S 影响的大小,如图 3 所示.

在 Tr 中,m_j 和 S 分别作为树的根和叶子,如果$\Gamma^q\in\Gamma$中 m_j 的直接前驱为 m_k,则 m_k 为 m_j 的儿子节点,依此类推.

显然,在 Tr 中,m_j 只能被其儿子节点 $m_k\in chlid(Tr,j)$直接影响.当不考虑重复激活时,m_k 对 m_j 的期望影响为 $\eta_{k\to j}=\eta_{S\to k}\times p_{k,j}\times(1-0)$.又由于 $chlid(Tr,j)$中团体对 m_j 的影响相互独立,所以我们可以合并所有 $chlid(Tr,j)$对 m_j 的

影响,得到$\eta_{S\to j}$的递推式:

$$\eta_{S\to j}=\begin{cases}R_j, & o_j\in S\\ 1-\prod_{k\in chlid(Tr,j)}(1-\eta_{S\to k}\times p_{k,j}), & o_j\notin S\end{cases} \tag{5}$$

设$|m_j|$表示 m_j 所包含的点数,根据公式(5),影响范围函数$\sigma(S)$可表示为

$$\sigma(S)=\sum_{m_j\in M}|m_j|\times\eta_{S\to j} \tag{6}$$

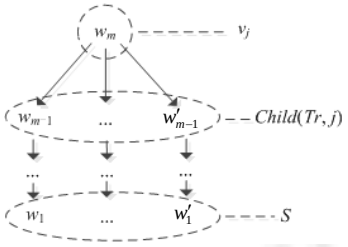


Fig.3 Tr tree

图 3 Tr 树

推论 **1**. 影响函数$\sigma(S)$具有单调性.

证明:给定 $IG(V,E,W)$,$S\subseteq W\subseteq V$.$W-S$ 到$\forall m_j\in M$ 的轨集数量$|\Gamma_{w-s}|\geq 0$,所以有$\Gamma_w\supseteq\Gamma_s$,$chlid(Tr_W,j)\supseteq chlid(Tr_S,j)$.我们首先归纳证明$\eta_{x\to y}$的单调性.当 $m_j\in S$ 时,根据公式(5)直接有$\eta_{W\to j}=\eta_{S\to j}=R_j$,所以$\eta_{W\to j}\geq\eta_{S\to j}$成立.当 $m_j\notin S$ 时,假设对于$\forall m_k\in chlid(Tr_W,j)$都有$\eta_{W\to k}\geq\eta_{S\to k}$成立.因为 $chlid(Tr_W,j)\supseteq chlid(Tr_S,j)$,对于 $m_j\notin S$ 有:

$$\eta_{W\to j}=1-\prod_{k\in chlid(Tr_W,j)}(1-\eta_{W\to k}\times p_{k,j})\geq 1-\prod_{k\in chlid(Tr_S,j)}(1-\eta_{W\to k}\times p_{k,j}).$$

又已假设$\eta_{W\to k}\geq\eta_{S\to k}$,有$\prod_{k\in chlid(Tr_S,j)}(1-\eta_{W\to k}\times p_{k,j})\geq 1-\prod_{k\in chlid(Tr_S,j)}(1-\eta_{S\to k}\times p_{k,j})=\eta_{S\to j}$.所以$\eta_{W\to j}\geq\eta_{S\to j}$成立,即,$\eta_{x\to y}$为单调递增函数.又已知单调增函数的和函数依然为单调增函数,$\sigma(S)$的单调性得证. □

推论 **2**. 影响函数$\sigma(S)$具有子模性.

为了证明$\sigma(S,T)$满足子模性,我们首先归纳证明$\eta_{x\to y}$ 满足子模性,即证:对于任意 $m_i\in M$,有$\eta_{S'\to j}-\eta_{S\to j}\geq\eta_{W'\to j}-\eta_{W\to j}$成立,其中 $S'=S\cup m_i$,$W'=W\cup m_i$,$S\subset W$.当 $m_j\in S$ 时,显然有$\eta_{S'\to j}-\eta_{S\to j}=\eta_{W'\to j}-\eta_{W\to j}=0$,子模性成立.当 $m_j\notin S$ 时,假设对于 S 的任意前驱 $m_k\in chlid(T_{W'}r,j)$都有$\eta_{S'\to k}-\eta_{S\to k}\geq\eta_{W'\to k}-\eta_{W\to k}$.根据公式(5)有:

$$\eta_{S'\to j}-\eta_{S\to j}=1-p_{k,j}\times\prod_{k\in chlid(Tr_{S'},j)}((1-\eta_{S'\to k})-(1-\eta_{S\to k}))=p_{k,j}\times\prod_{k\in chlid(Tr_{S'},j)}(\eta_{S'\to k}-\eta_{S\to k}).$$

等式两边取对数得 $\log(\eta_{S'\to j}-\eta_{S\to j})=\log p_{k,j}+\sum_{k\in chlid(Tr_{S'},j)}\log(\eta_{S'\to k}-\eta_{S\to k})$.

又$\eta_{S'\to k}-\eta_{S\to k}\geq\eta_{W'\to k}-\eta_{W\to k}$,有 $\log(\eta_{S'\to j}-\eta_{S\to j})-\log(\eta_{W'\to j}-\eta_{W\to j})=\sum_{k\in chlid(Tr_{W'},j)}\log\frac{(\eta_{S'\to k}-\eta_{S\to k})}{(\eta_{W'\to k}-\eta_{W\to k})}\geq 0$,即$\eta_{S'\to j}-\eta_{S\to j}\geq\eta_{W'\to j}-\eta_{W\to j}$成立.

如上所述,在假设团体的邻居满足子模性后,该团体同样满足子模性,根据归纳法,有 $\eta_{x\to y}$ 满足子模性成立.又已知子模函数的和函数同样是子模的,所以$\sigma(S)$具有子模性. □

3.3 CGIM算法

根据推论 1 和推论 2,我们将给出一种贪心算法来快速求解团体影响最大化问题,具体描述如算法 2 所示.

算法 **2**. CGIM.

输入:$IG(M,I,W)$, k, $R=\{R_1,\ldots,R_n\}$.

输出: seed set S.

1: $S\leftarrow\varnothing$

2: **while** $|S|<k$ **do**

3: **for each** m_i **in** $M-S$

4: $m_j \leftarrow \arg\max_{m_i \in V-S} \sigma(S \cup m_i) - \sigma(S)$

5: $S \leftarrow S \cup m_j$

6: **return** S

初始时,CGIM 算法将种子集 S 初始化为空集.每一轮迭代中,我们以 $S \cup m_i$ 作为备选种子,通过 GIC 模型估计 $S \cup m_i$ 的影响范围 $\sigma(S \cup m_i)$,并选取边际影响收益 $\sigma(S \cup m_i)-\sigma(S)$ 最大的 m_i 加入 S.重复此过程,直到 S 的大小为 k,则返回.

CGIM 算法每次迭代需计算一次所有备选种子的边际收益,总共 k 轮迭代,一共需要 $O(k \times |M|)$ 次.每次边际收益的计算需要通过一次 GIC 传播模拟过程来确定收益大小.最坏情况下,GIC 传播模拟需要遍历 $IG(M,I,W)$ 中所有点和边,时间复杂度为 $O(|M|+|I|)$.由此,算法 2 的总的时间复杂度为 $O(k \times |M| \times (|M|+|I|))$.由于团体的规模 $|M|$ 远小于社会网络中点的规模 n,所以该方法时间复杂度相对于 n 仍可看作线性时间.

4 实验分析

- 人工数据集

采用 LFR(lancichinetti fortunato radicchi)算法[15]生成的人工网络.特别指出:LFR 算法在生成网络的同时,能够自动给出社区划分基准,而社区是团体一种自然表现形式,属于本文研究对象的范畴.LFR 算法的关键参数说明如下:n 表示模拟网络的点个数,m 表示模拟网络的边数,w 表示社区数量.在本文实验中,我们将通过调节算法参数生成多种规模网络来对 CGIM 算法进行测试,见表 3.

Table 3 Experimental network

表 **3** 实验网络

	n (K)	m (K)	w (K)	*Avg-s*
net_1	10	20	0.4	25
net_2	40	50	0.4	100
net_3	100	400	2	50
dblp	8.345	343.261	0.7	119

- 真实数据集

采用作者合作网络,其中,节点表示作者,边表示两个作者之间存在合作关系.我们从 DBLP 中 2012 年计算机领域的 700 个期刊或会议中抽取了 83 450 个作者、343 261 条合作关系.为获得该合作网络中的团体划分,我们视一个期刊或者会议为一个团体,并将作者划分至投稿次数最高的团体.如,作者 A 向会议(或期刊)L_1,L_2 的投稿次数分别为 3,5,那么 A 将被划分至 L_2.

- 历史观测数据集生成

为了获得历史观测数据,我们将通过以下方式生成.

假定实验网络中点的传播概率 p 相同.在每次“疾病”传播过程中,我们从测试网络中随机的选择 1%的点作为“感染”点,并根据 IC 模型进行影响传播模拟.在传播模拟结束后,我们记录各个团体的“感染”状态作为一条记录,并生成多条记录作为本文实验的观测数据集.

为了验证 CGIM 算法的性能,我们将在点的粒度上使用多种传统方法求解团体影响最大化问题,并以此作为对比基准.实验中,算法的效果好坏可通过各算法输出种子的影响范围(即 $\sigma(S)$,见公式(6))来评价,其中,种子的影响范围越大,说明算法效果越好.特别指出:为了避免不同传播模型对影响范围估计的差异,本文将统一在 IC 模型下对各算法输出种子的影响范围进行计算.由相关工作的分析可知:贪心算法的优势在于选点质量高,而启发算法的优势在于效率,所以本文将选取 CELF 算法、TIM 算法和 degreeDis 算法作为比对算法.

在本文实验中,我们在不同 k 值下对比了各算法的影响范围和运行时间.此外,我们还测试了团体大小、数量对算法的影响.本文程序采用 Java 编写,实验环境为 Quad-Core 2.0 GHz CPU,8GB 内存的个人电脑.

4.1 实验结果

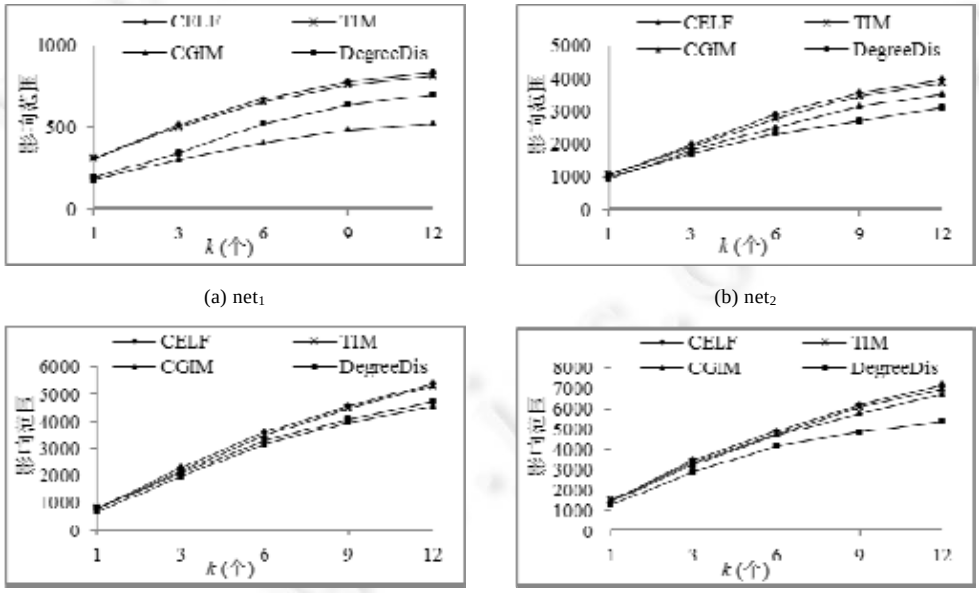
4.1.1 影响范围对比

为了验证网络特征变化对算法效果的影响,本文首先考虑团体规模的因素,并分别在团体平均大小不等的人工网络上验证了各算法输出种子的影响范围.

- 由图 4(a)所示,当团体平均大小 *Avg-s* 为 25 时,CELF,TIM 在 net_1 上的影响范围明显高于其他算法,其次是 DegreeDis,CGIM 的效果较差.
- 当 *Avg-s*=50 时(如图 4(c)所示),CELF,TIM 相对于其他算法的优势逐渐减弱,而 CGIM 与 DegreeDis 的差距被拉近.
- 当 *Avg-s*=100 时(如图 4(b)所示),CELF,TIM 的影响范围依然保持领先,但 CGIM 的表现已明显好于 DegreeDis.

由此可见,团体规模大小对算法效果的影响显著.DegreeDis 的影响范围随团体规模的增大出现了下降的原因是:在规模较大的团体中,度数较高的点之间有很大可能在多跳之后存在较严重的影响重叠,所以度启发规则的选种策略存在单个种子影响力大而总体影响力小的问题.CGIM 的影响范围随团体规模的增大明显增加的原因是:团体中点数越多,单个点的状态变化对整体的影响就越小,即在同质性假设下,点的概率值更贴近团体"感染"比例.所以,基于 GIC 更能反映团体的影响关系,从而间接提高了 CGIM 选点的质量.此外,我们还注意到:图 4(a)中,DegreeDis 的影响曲线在 k=3 和 k=6 之间存在明显跳动.该现象反映了 DegreeDis 算法存在不稳定的问题.CELF 和 CGIM 的曲线随着 k 的增大平缓增加,反映出基于传播模型的算法相对于度启发式算法更具稳定优势.

为了验证 CGIM 算法的实用性,我们在 DBLP 网络上进行对比测试,如图 4(d)所示.图中显示:CGIM 的效果随 k 变化始终较为贴近 CELF,而 DegreeDis 在 $k\geq 6$ 时效果较差.从各算法曲线的对比观察中可以发现:DegreeDis 随着 k 增大,其影响范围增长缓慢,这是该算法效果不佳的主要因素.这是由于 DBLP 网络中存在较为明显的领域性质,例如,同属于数据库领域的 Sigmod 会议和 VlDB 会议往往对其领域内的其他会议和期刊同时造成影响.因为属于同一个领域的团体(会议)间往往存在较大的影响重叠,该特点直接导致了度启发规则的效果不佳.



(a) net_1 (b) net_2

(c) net_3 (d) dblp

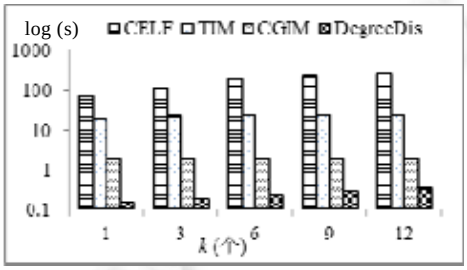
Fig.4 Comparison of influence

图 4 影响力对比

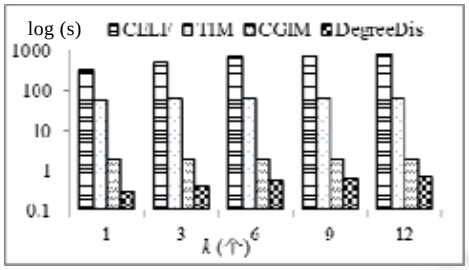
4.1.2 效率对比

图 5(a)~图 5(c)分别显示了在 net_1,net_2,net_3 上各算法的执行时间.特别指出:由于不同算法的执行时间相差较大,我们选择对数刻度描述时间轴.实验结果印证了 CELF 的低效性.如图 5(a)所示,CELF 在小规模网络(net_1)上选取 1 个团体种子的时间即为分钟级别,且随着 k 的增大,执行时间增加明显,效率远远低于 TIM,CGIM 和 DegreeDis;CGIM 的效率明显高于 CELF,TIM 而略低于 DegreeDis.这说明当团体数目远小于点的数量时(net_1: n=10k,w=0.2k),GIC 模型相对于 IC 模型在估计团体边际收益时的效率优势明显.此外,我们还注意到,当 k 增大时,CGIM 的执行时间基本保持不变.这是由于 net_1 中团体数仅为 200,以至于 CGIM 选择局部最优解所消耗的时间相对于扫描历史观测数据来说基本可以忽略引起的.

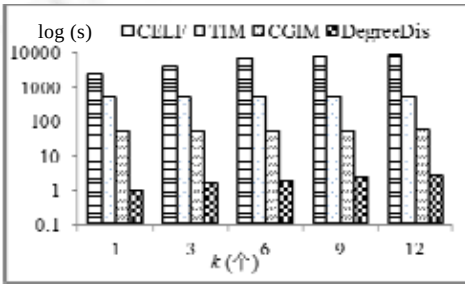
横向对比图 5(a)~图 5(c)我们发现,CGIM 执行时间对网络团体数量规模变化敏感.例如,当团体数同为 w=200 时,CGIM 的执行时间在 net_2 上(如图 5(b)所示)与在 net_1 上(如图 5(a)所示)相比基本无变化,而在 w=2000 的 net_3 上(如图 5(c)所示),CGIM 的执行时间相对于 net_1 上升了 51 倍.由此可见,团体数目对 CGIM 的效率影响较大.为此,我们保持点、边规模不变(n=20k,m=50k),在 w={1k,2k,3k,4k,5k}的网络上进一步验证团体数对 CGIM 效率的影响.由图 5(d)可知,CGIM 在 w={1k,2k,3k,4k,5k}时的执行时间分别为{21,73,179,913,2445}s.可以看出,CGIM 随着团体个数增加,执行时间成指数级上升.造成该问题的原因是 CGIM 依赖于团体关联图的构建,而当团体数量较多时,计算团体间条件概率独立需要大量的时间开销.由此可得出结论:CGIM 算法在团体数目较多的网络上,效率优势不明显,相比之下,更适合处理点数规模较大而团体规模相对较小的网络.



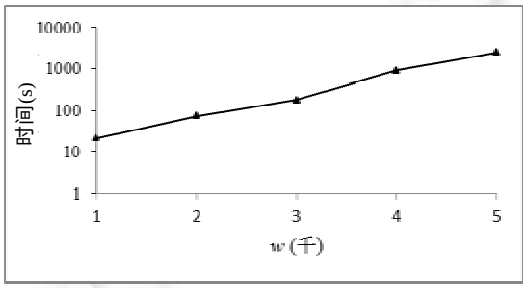
(a) net_1 (log scale)



(b) net_2 (log scale)



(c) net_3 (log scale)



(d) 团体数与 CGIM 效率

Fig.5 Comparison of execution time

图 5 执行时间对比

5 结 论

本文提出并研究了团体影响最大化问题,通过发现历史“感染”数据中团体的概率关联,建立了团体传播模型 GIC,由此给出了一种高效的团体最大化算法 CGIM.与传统的面向点的影响最大化方法不同的是,本文方法不依赖于点影响关系的获取,即可快速定位最有影响力的团体种子集.实验结果表明:当网络中团体数量远小于点数量时,CGIM 算法比 CELF,TIM 算法更高效,且比 degreeDis 算法更准确,适合于处理点数规模较大而团体规

模相对较小的网络.

未来可行的研究方向包括以下 4 个方面.

1) 针对本文团体关联图构建算法(算法 1)不适用于团体数量规模较大网络的问题,我们将研究如何利用概率性质对关联计算进行剪枝,从而提高该算法的效率.
2) 本文工作假设团体之间的点互不重叠,如何在考虑重叠情况下对团体最大化问题进行快速求解,则是我们试图研究的第 2 个问题.
3) 我们还将考虑社会网络的动态性,研究如何在点随时加入、退出团体的情况下求解团体最大化问题.
4) 最后,我们还将试图给出 CGIM 的并行版本,并基于 hadoop,spark 等平台进一步提高算法的可扩展性.

References:

[1] Guille A, Hacid H, Favre C, Zighed DA. Information diffusion in online social networks: A survey. ACM SIGMOD Record, 2013,42(2):17−28. [doi: 10.1145/2503792.2503797]

[2] Lü L, Zhang YC, Yeung CH, Zhou, T. Leaders in social networks, the delicious case. PloS One, 2011,6(6):e21202. [doi: 10.1371/journal.pone.0021202]

[3] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Lise G, Ted ES, Pedro MD, Christos F, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2003. 137−146. [doi: 10.1145/956750.956769]

[4] Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N. Cost-Effective outbreak detection in networks In: Pavel B, Rich C, Xindong W, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. San Jose: ACM Press, 2007. 420−429. [doi: 10.1145/1281192.12812 39]

[5] Wang Y, Cong G, Song G, Xie K. Community-Based greedy algorithm for mining top-*k* influential nodes in mobile social networks. In: Bharat R, Balaji K, Andrew T, Qiang Y, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2010. 1039−1048. [doi: 10.1145/1835804.1835935]

[6] Fortunato S. Community detection in graphs. Physics Reports, 2009,486(3-5):75−174. [doi: 10.1016/j.physrep.2009.11.002]

[7] Borgs C, Brautbar M, Chayes J, Lucier B. Maximizing social influence in nearly optimal time. In: Chandra C, ed. Proc. of the Symp. on Discrete Algorithms. Portland: SIAM, 2014. 946−957.

[8] Tang Y, Xiao X, Shi Y. Influence maximization: Near-Optimal time complexity meets practical efficiency. In: Curtis ED, Feifei L, Özsu MT, eds. Proc. of the Int'l Conf. on Management of Data. Snowbird: ACM Press, 2014. 75−86. [doi: 10.1145/2588555.2593670]

[9] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. In: John FEI, Françoise F, Peter AF, Mohammed JZ, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Paris: ACM Press, 2009. 199−208. [doi: 10.1145/1557019.1557047]

[10] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Bharat R, Balaji K, Andrew T, Qiang Y, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2010. 1029−1038. [doi: 10.1145/1835804.1835934]

[11] Jung K, Heo W, Chen W. Irie: Scalable and robust influence maximization in social networks. In: Zaki MJ, Siebes A, Yu JX, Goethals B, Webb GI, Wu XD, eds. Proc. of 2012 IEEE the 12th Int'l Conf. on Data Mining. Brussels: IEEE Computer Society, 2012. 918−923. [doi: 10.1109/ICDM.2012.79]

[12] Cao JX, Dong D, Xu S, Zheng X, Liu B, Luo JZ. A *K*-core based algorithm for influence maximization in social networks. Chinese Journal of Computers, 2015,38(2):238−248 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2015.00238]

[13] Myers S, Leskovec J. On the convexity of latent social network inference. In: John DL, Christopher KIW, John S, Richard SZ, Aron C, eds. Proc. of the Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2010. 1741−1749.

[14] Gomez-Rodriguez M, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks. In: Lise G, Tobias S, eds. Proc. of the Int'l Conf. on Machine Learning. Washington: Omni Press, 2011. 561−568.

[15] Gomez Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence. In: Bharat R, Balaji K, Andrew T, Qiang Y, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2010. 1019−1028. [doi: 10.1145/ 2086737.2086741]

[16] Gomez Rodriguez M, Leskovec J, Schölkopf B. Structure and dynamics of information pathways in online media. In: Stefano L, Alessandro P, Paolo F, Aristides G, eds. Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. Rome: ACM Press, 2013. 23−32. [doi: 10.1145/2433396.2433402]

[17] Mehmood Y, Barbieri N, Bonchi F, Ukkonen A. CSI: Community-Level Social Influence Analysis. Springer-Verlag, 2013. 48−63. [doi: 10.1007/978-3-642-40991-2_4]

[18] Hu Z, Yao J, Cui B, Xing E. Community level diffusion extraction. In: Timos KS, Susan BD, Zachary GI, eds. Proc. of the Int'l Conf. on Management of Data. Melbourne: ACM Press, 2015. 1555−1569. [doi: 10.1145/2723372.2723737]

[19] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Reasoning. Morgan Kaufmann Publishers, 1988. 84−86.

[20] Roughgarden T, Tardos E, Vazirani VV. Algorithmic Game Theory. Cambridge: Cambridge University Press, 2007. 648−651.

[21] Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 1978,14(1):265−294. [doi: 10.1007/BF01588971]

附中文参考文献:

[12] 曹玖新,董丹,徐顺,郑啸,刘波,罗军舟.一种基于 *k*-核的社会网络影响最大化算法.计算机学报,2015,38(2):238−248. [doi: 10.3724/SP.J.1016.2015.00238]

张平(1984－),男,湖北武汉人,博士生,主要研究领域为社会化网络,Web 数据管理.

岳昆(1980－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为不确定数据管理,不确定性知识发现与推理,数据密集型计算环境下的数据挖掘与知识发现.

王黎维(1981－),女,博士,副教授,CCF 专业会员,主要研究领域为数据质量,数据溯源,科学工作流.

黄浩(1986－),男,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘.

彭智勇(1963－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为复杂数据管理,可信数据管理,Web 数据管理.