物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

一种新的复杂网络影响力最大化发现方法

胡庆成　张勇　许信辉　邢春晓　陈池　陈信欢

A new approach for influence maximization in complex networks

Hu Qing-Cheng　Zhang Yong　Xu Xin-Hui　Xing Chun-Xiao　Chen Chi　Chen Xin-Hua

引用信息 Citation: Acta Physica Sinica, 64, 190101 (2015)　DOI: 10.7498/aps.64.190101
在线阅读 View online: http://dx.doi.org/10.7498/aps.64.190101
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2015/V64/I19

一种新的复杂网络影响力最大化发现方法*

胡庆成† 张勇 许信辉 邢春晓 陈池 陈信欢

(清华大学计算机科学与技术系, 信息技术研究院, 清华信息科学与技术国家实验室, 北京 100084)

(2014 年 12 月 8 日收到; 2015 年 6 月 10 日收到修改稿)

复杂网络中影响力最大化建模与分析是社会网络分析的关键问题之一, 其研究在理论和现实应用中都有重大的意义. 在给定 s 值的前提下, 如何寻找发现 s 个最大影响范围的节点集, 这是个组合优化问题, Kempe 等已经证明该问题是 NP-hard 问题. 目前已有的随机算法时间复杂度低, 但是结果最差; 其他贪心算法时间复杂度很高, 不能适用于大型社会网络中, 并且这些典型贪心算法必须以了解网络的全局信息为前提, 而获取整个庞大复杂且不断发展变化的社会网络结构是很难以做到的. 我们提出了一种新的影响力最大化算法模型 RMDN, 及改进的模型算法 RMDN++, 模型只需要知道随机选择的节点以及其邻居节点信息, 从而巧妙地回避了其他典型贪心算法中必须事先掌握整个网络全局信息的问题, 算法的时间复杂度仅为 $O(s\log(n))$; 然后, 我们利用 IC 模型和 LT 模型在 4 种不同的真实复杂网络数据集的实验显示, RMDN, RMDN++ 算法有着和现有典型算法相近的影响力传播效果, 且有时还略优, 同时在运行时间上则有显著的提高; 我们从理论上推导证明了方法的可行性. 本文所提出的模型算法适用性更广, 可操作性更强, 为这项具有挑战性研究提供了新的思路和方法.

关键词: 复杂网络, 影响力最大化, 信息传播, 贪心算法

PACS: 01.75.+m, 05.90.+m, 89.75.–k, 89.70.Hj **DOI:** 10.7498/aps.64.190101

1 引 言

影响力分析是复杂网络的重要研究内容, 现实世界中的诸多系统都以复杂网络 (Complex Network) [1−3] 形式存在, 比如互联网、社会系统、计算网络、生物网络和社交网络等. 在很多科学领域中, 都使用网络来表示系统中成员之间的关系, 如在社交网络中用节点代表人, 边代表人与人之间的联系, 人们的行为和思想会受到其他人的影响而发生变化. 特别是随着在线社交网络蓬勃发展, 以交友、信息分享等为目的的社交网络成为我们传播信息、表达观点、分享信息的理想平台, 同时为影响力研究提供了真实的经验数据支撑, 这种复杂的社会网络关系对信息传播和扩散起着至关重要的作用. 影响力最大化问题通过分析人们相互之间的影响模式和影响力传播方式, 既能从社会学角度加深理解人们的社会行为, 同时也能促进政治、经济和文化活动等领域的交流与传播, 其在理论和现实应用中都有重大的意义. 例如有效的控制疾病传播、流言散布、计算机病毒扩散, 还可以传播新产品、新思想、新技术以推进社会化进程. 复杂网络中影响力最大化问题一直是研究的热点和难点.

影响力最大化问题 (influence maximization) 是在给定预算的前提下, 如何选择 s 个初始传播种子节点, 最终使得它们传播范围最大化. 长期以来, 最有影响的节点发现 [4−8] 一直是研究的热点, Kitsak 等 [9] 提出了 K-shell 分解来确定最具有影响力的单源传播节点, 对多节点传播集合也只是给出与传播节点之间的距离有关系的假设. 文献 [10,11] 给出了复杂网络中最有影响力节点的相关研究进展

* 国家重点基础研究发展 (973 计划) (批准号: 02011CB3023302) 和国家高技术研究发展计划 (863 计划) (批准号: SS2015AA020102) 资助的课题.

† 通信作者. E-mail: hqc10@mails.tsinghua.edu.cn

及各种方法的分析综述. Domingos [12] 和 Richardson [13] 等首先将影响力最大化问题归纳为一个算法问题, 主要是在社交网络中找出最有影响力的成员, 提供给他们免费的样品, 希望通过他们向网络中其他成员推荐, 从而达到营销的目的, 这种通过口口相传 (word of mouth) 的影响力传播方式, 商业营销目标以最小的费用将新产品最大范围地推广到整个网络. Kempe 和 Kleinberg 等 (简称 KKT 算法) [14] 形式化表示了该问题, 首次证明了求解复杂网络上的影响力最大化问题是个 NP-hard 问题, 并给出与最优解比为 $1-1/\mathrm{e}\approx 63\%$ 的近似贪心算法 (greedy algorithm, GA), 其主要缺点是时间复杂度大, 速度慢. 同时给出了独立级联 (independent cascade) 模型和线性阈值 (linear threshold) 模型. Leskovec 等 [15] 提出了最优化的贪心算法 CELF (cost-effective lazy forward) 框架, 利用模型中子模函数的性质, 算法在取得近似最优解的同时效率比贪心算法提高了近 700 倍. Goyal 等 [16] 提出的 CELF++ 算法, Zhou 等 [17] 提出的 UBLF 算法, 都是基于 CELF 提出了时间复杂稍优的算法, 但是总体在求解大规模网络问题时仍有问题. Chen 等 [18] 提出了两种改进的贪心算法 NewGreedy 和 MixGreedy 算法, 此外还提出了一种改进的度数最大算法 DegreeDiscount 算法. NewGreedy 算法是从原始网络中去掉对传播没有影响的边, 得到一个小网络, 然后在小网络中做影响力传播, 优点是不需要每次都从整个网络上去考虑. MixGreedy 算法是先用 NewGreedy 算法计算第一步, 后面用 GELF 算法, 把两个算法结合起来使用, 实验结果表明, 这两个改进的算法都比贪心算法的时间复杂度低, 但是相比度数最大的算法, 时间复杂度还是很高. 选取度数最大的算法是求解影响力最大化问题比较好的启发式方法, DegreeDiscount 算法 [18] 是对探索式算法的一种优化策略的改进, 使得实验结果与贪心算法相近, 而运行效率有了很大提升. Kimura 等 [19] 提出一种通过分解极大强连通子图寻找影响力最大化算法, 在此基础上又出现了基于用户间最大影响路径的方法 [20], 但是通过最短路径传播的假设限制性太强, Wang 等 [21] 发现影响力的传播大多发生在社团之间, 由此提出一种贪心策略结合动态规划的算法用于初始用户的选取, 较大提升了算法的执行效率, Galstyan 等 [22] 利用收益递减策略研究市场收益最大化. Li 等 [23] 提出了积极、消极影响力最大化 PRIM 模型.

总之, 影响力最大化问题在 IC 模型和 LT 模型下运行是 NP-hard [14], 在整个网络中随机选取 s 个节点 (random heuristic) 时间复杂度最少, 但效果也是最差的; 其他贪心算法和启发式算法都以不同精度接近近似最优解, 但是必须以了解网络的全局信息为前提 (如网络拓扑结构等), 在许多情况下, 很难了解整个社会网络关系; 同时一些算法会陷入"富人俱乐部 (rich club)" [24] 这一局部最优的现象.

本文提出了一种新的影响力最大化模型称为随机节点最大度邻居方法 (RMDN), 该方法只需要知道被随机选择的节点以及与它直接相连的邻居节点信息, 从而巧妙地回避了其他算法中必须事先掌握整个网络全局信息的问题. 算法的时间复杂度仅为 $O(s\log(n))$, 运行时间随着网络的规模地扩大, 比经典最大度 (degree heuristic) 算法 ($O(m)$) 运行时间效果呈线倍数增加, 然而我们所提出的算法运行结果与已有典型算法相近, 且有时还略优.

在本文第二部分介绍了影响力最大化研究的背景知识与相关工作, 及实验传播模型; 在第三部分介绍了本文的算法模型及相应理论推导分析过程. 第四部分给出了各种算法在 4 个实际社会网络的实验和分析比较; 最后部分是总结和未来工作计划.

2 相关研究

在社会网络中, 影响力最大化问题可以帮助我们有效地控制疾病传播、流言散布、计算机病毒扩散, 还可以传播新产品、新技术、新思想以加快推进社会化进程. 由于影响力最大化问题是 NP-hard 的, 时间复杂度高, 而且在线社交网络的规模日益庞大, 所以设计新的算法模型以期获得最优解和提高算法的执行效率一直是研究的重要内容.

为了表述方便, 表 1 列出了贯穿全文的重要变量参数.

表 1 重要变量参数对照说明

Table 1. Important variables used in the paper.

变量参数	描述	变量参数	描述
n	网络节点个数	m	网络边的个数
S	初始传播种子节点集合	T	最终被传播节点集合
s	被选中的节点数	$k_{\min}$	节点最小度数
p	传播概率	$k_{\max}$	节点最大度数
R	算法迭代循环的次数	$P(k)$	节点度数是自然数 k 的概率

2.1 影响力最大化问题的定义和贪心算法

问题定义: 给定网络 $G=(V,E)$ 和常数 $s \leqslant |V|$, V 代表为社会网络中个体节点, E 代表个体之间的关系. 如果初始传播种子节点集合为 S, 传播过程结束后预期激活节点集合为 $T=\delta(S)$, 找出节点集合 $S \subseteq V$ 且 $|S|=s$, 使得传播范围 $\delta(S)$ 最大.

KKT 算法 [14] 给出了影响力最大化问题的一般贪心算法 (如算法 1 所示): 从一个空集合开始, 并且每轮迭代重复的添加一个当前最具影响力的节点; 最后, 得到大小为 s 的初始传播种子集合. 已经证明算法的解以 $(1-1/e)$ 近似逼近最优解, 算法缺点非常明显, 运行十分地耗时.

算法 1 GeneralGreedy (G,s)

```
initialize S = ∅ and R = 10000
for i = 1 to s do
  for each vertex v ∈ V\S do
    s_v = 0
    for i = 1 toR do
      s_v^+ = |δ(S ∪ {v}|
    end for
    s_v = s_v/R
  end for
  S = S ∪ {argmax_{v∈V\S} s_v}
end for
output S.
```

2.2 基于度数的节点启发式算法

中心度是分析社会网络的一个最重要的和常用的概念工具之一. 在一个社会网络中, 节点度数越高, 说明该节点在网络结构中的位置越重要或影响力越大. 在复杂网络中以度数递减的顺序选择 s 个最大度数节点的启发式选择策略, 是长期以为一个标准方法, 在社会科学中被称为"度中心性" [25]. 此方法的一个缺点就是静态选择初始节点, 没有考虑影响的扩散过程和网络中具有"富人俱乐部"现象, 影响范围会陷入局部最优, 而不能保证最终全局最优. DegreeDiscount 算法 [18] 是对度数最大节点的一种改进, 算法基本思想是当一个节点 u 的邻居节点中有一些节点已经被选作为初始的节点, 在选取下一个度数最大的节点时, 对节点 u 的度数重新计算, 最后再选出打折后度数最大的节点, 使得实验结果与贪心算法相近, 而运行效率有了很大提升. 影响力最大化相关典型算法的运行时间复杂度 (如表 2 所示).

表 2 影响力最大化典型算法的时间复杂度比较

Table 2. Time complexity of algorithms.

算法	时间复杂度
random heuristic algorithm	$O(s)$
degree heuristic algorithm	$O(m)$
degree discount algorithm	$O(s\log(n)+m)$
single discount algorithm	$O(s\log(n)+m)$
NewGreedy IC algorithm	$O(sRm)$
CELF Greedy algorithm	$O(snRm/700+)$
generalGreedy algorithm	$O(snRm)$

2.3 影响力的传播模型

对于复杂网络中每个节点有两种状态, 激活状态和未激活状态: 若一个节点已经接受了信息, 则称为激活节点, 否则为非激活节点. 激活节点对于未激活节点存在影响, 如果某节点邻居的激活节点越多, 则该节点被激活的可能性就越大. 新激活节点又会影响其他处于未激活状态的邻居节点. 在网络环境中, 最主要的交互活动就是信息的发布、分享和扩散, 所以影响力在社会网络中的作用过程和信息的扩散过程有内在紧密的联系和十分相似的机理, 因此传播模型在影响力传播问题的研究过程中发挥着非常重要的作用, 独立级联模型和线性阈值模型是信息传播过程进行建模的重要方法.

2.3.1 独立级联模型

IC 模型 (independent cascade model) [14] 是基于相互粒子系统设计的一个信息扩散模型, 可以描

述为: 在复杂网络 $G=(V,E)$ 中, 对于 V 的每一个顶点 u 和它的邻居节点 v, 有一条连 $e(uv)$ 存在, p_{uv} 表示在传播过程 u 对邻居节点 v 的影响力的概率, p_{uv} 的取值是独立的. 如果在时刻 t, u 是激活的状态, 并且其邻居 v 是未激活的, 那么 u 将尝试以概率 p_{uv} 去激活 v. 如果这个过程成功了, 那么在 $t+1$ 时刻 v 就成了激活状态. 但是不管成功与否, u 再也不能试图去激活 v. 如果 v 在 t 时刻同时有多个邻居都处于激活状态, 他们尝试激活 v 的顺序是任意的. 系统从初始态开始传播, 直到没有新的节点可以被激活为止.

2.3.2 线性阈值模型

LT 模型 (liner threshold model) [14] 是诸多阈值模型的核心, 可以描述为: 在复杂网络 $G=(V,E)$ 中, 定义 $N(u)$ 为节点 u 的邻居节点集合. 被激活的节点 u 对邻居节点 v 存在影响为 b_{uv}, 一个节点 u 的所有邻居节点对 u 的影响力总和小于等于 1, 即 $\sum_{v\in N(u)} b_{uv}\leqslant 1$. 每个节点 u 有一个特定阈值 $\theta_u\in[0,1]$, 如果 $\sum_{v\in N(u)} b_{uv}\geqslant\theta_u$, 则 u 被激活. LT 模型中, 当一个激活节点 u 尝试激活它的未激活邻居 v 没有成功时, 节点 u 对节点 v 的影响力 b_{uv} 被积累起来, 这样对后面其他邻居节点对 v 的激活是有贡献的, 直到节点 v 被激活或传播过程结束. 这就是 LT 模型的"影响积累"特性, 这与 IC 模型是不同的.

3 算法模型

由于影响力最大化问题是 NP-hard 的, 时间复杂度高, 目前算法都必须以了解网络的全局信息为前提, 然而了解网络的整体结构关系也是难以做到的. 如表 2 所示, 随机启发式 (random heuristic) 算法执行时间复杂最好, 度中心化启发式 (degree heuristic algorithm) 算法时间复杂度及结果综合较优, 但必须了解网络全局信息. 结合目前社会网络基本都符合为幂律 (power law) 特征分布的无标度网络 [1−3], 其度分布是呈集散分布: 大部分的节点连接较少, 而少数节点有大量的连接. 且网络中任何节点之间连接的度数满足六度分隔理论 [26] (six degree of separation), 且已在社交网络中被证实, 如 Facebook 为 4.74 度分隔 [27], Twitter 为 4.67 度分隔 [28]. 同时, 现实社会网络中, 我们知道任何人 (节点) 至少知道他朋友的基本信息这一常理.

3.1 RMDN 算法模型

基于以上知识, 我们提出了随机节点最大度邻居 (RDMN) 算法模型. 基本思想是: 从具有 n 个节点的复杂网络中随机选出一个节点, 再从此节点及其邻居节点中选出一个度最大的节点作为种子节点, 一直到选择 s 个不同的传播源种子节点为止. 简单描述过程如算法 2 所示.

算法 2 RandomMaxDegreefNeighbor(G,s)

```
initialize S = ∅
while len(S) < s do
  random choise vertex u ∈ V\S
  select u = argmaxDegree {u ∪ neighbor(u)}
  S = S ∪ {u}
end while
output S.
```

3.2 模型理论分析推导

对于具有 n 个节点的无标度网络, 其度分布满足幂律特征, 即度为 k 的节点出现的概率为 p_k 正比于 $ck^{-\gamma}$ [29]. 系统中节点度的最小值为 $k_{\min}$, 由 $\int_{k_{\min}}^{\infty} p(k)\mathrm{d}k=1$, 得出 $c=(\gamma-1)k_{\min}^{\gamma-1}$, 假设节点度的最大值 $k_{\max}$ 为一个节点, 那么

$$\int_{k_{\max}}^{\infty} p(k)\mathrm{d}k=\frac{1}{n},\quad p(k)=ck^{-\gamma}. \tag{1}$$

将常数 $c=(\gamma-1)k_{\min}^{\gamma-1}$ 代入 (1) 式, 推得

$$\begin{aligned}\int_{k_{\max}}^{\infty} p(k)\mathrm{d}k&=\int_{k_{\max}}^{\infty} ck^{-\gamma}\mathrm{d}k\\&=\int_{k_{\max}}^{\infty}(\gamma-1)k_{\min}^{\gamma-1}k^{-\gamma}\mathrm{d}k\\&=(\gamma-1)k_{\min}^{\gamma-1}\int_{k_{\max}}^{\infty}k^{-\gamma}\mathrm{d}k=\frac{1}{n}.\end{aligned} \tag{2}$$

可得度最大值为 $k_{\max}=k_{\min}n^{\frac{1}{\gamma-1}}$. 同理可以得到整个网络中度数最大的前 s 个节点为 $k_{\text{top-}s}=k_{\min}(n/s)^{\frac{1}{\gamma-1}}$.

3.2.1 网络中节点度的分布

根据 Newman [30] 提出的生成函数 (generating function), 网络中度为 k 的节点分布生成函数 $G_0(x)$ 可以用表示为

$$\begin{aligned}G_0(x)&=\sum_{k_{\min}}^{k_{\max}}p_kx^k=\sum_{k_{\min}}^{k_{\max}}ck^{-\gamma}x^k,\\c&=\frac{1-\gamma}{k_{\max}^{1-\gamma}-k_{\min}^{1-\gamma}}.\end{aligned} \tag{3}$$

p_k 是网络中度数为 k 的节点出现的概率, 是 c 满足归一化条件 $G_0(1)=1$ 的一个常量. 可知

$$G_0'(x)=\sum_{k_{\min}}^{k_{\max}} kp_k x^{k-1},$$

$$G_0'(1)=\sum_{k_{\min}}^{k_{\max}} kp_k=\langle k\rangle.$$

3.2.2 任意节点邻居度的分布

对于数度为 k 的节点 u, 其被随机选中的概率为 P_k, 那么由边到达节点 u 的概率为 kP_k(注意随机选择节点与随机从连边选择节点是不同的), 其生成函数可以表示为 $\sum\limits_k kP_k x^k$, 标准归一化表示为

$$\frac{\sum\limits_k kP_k x^k}{\sum\limits_k kP_k}=x\frac{G_0'(x)}{G_0'(1)}. \tag{4}$$

当随时选择一节点 u, 设其任意连边的邻居节点 v 的度数为 m, 参照示意图 1, v 被选中的概率为 P_m, 除去 $u\to v$ 的边时, 节点 v 的度数变为 m', $m'=m-1$. 整个网络节点数为 n, 那么 u 被随机选中的概率为 $1/n$, 那么 v 除去 u 被选中的概率变为 $P_{m'}=P_m-1/n$, 当 $n\to\infty$, $1/n\to 0$, 因此 $P_{m'}=P_m-1/n\approx P_m$. 节点 v 的分布生成函数为

$$\sum_{m'}P_{m'}x^{m'}\cong\sum_{m'}P_{m'}\frac{x^m}{x}\cong\sum_{m'}P_m\frac{x^m}{x},$$

与 (4) 式只差 x 的一次方, 因此 $G_1(x)=\dfrac{G_0'(x)}{G_0'(1)}$, 即对于任意节点连接的一度邻居离散型生成函数可以表示为

$$G_1(x)=\sum_{k_{\min}}^{k_{\max}}p_m x^m=\sum_{k_{\min}}^{k_{\max}}bm^{1-\gamma}x^m,$$

$$b=\frac{2-\gamma}{k_{\max}^{2-\gamma}-k_{\min}^{2-\gamma}}. \tag{5}$$

同理可得对于随机选择任意节点其二度邻居连接分布可以表示为

$$G_2(x)=\sum_k p_k[G_1(x)]^k=G_0(G_1(x)).$$

我们可以推导出对任意选取节点的 m 度邻居表达式:

$$G^{(m)}(x)=\begin{cases}G_0(x), & m=1,\\ G^{(m-1)}(G_1(x)), & m\geqslant 2.\end{cases}$$

本节只对一度邻居进行分析, 推导分析过程同样可适用于多度邻居.

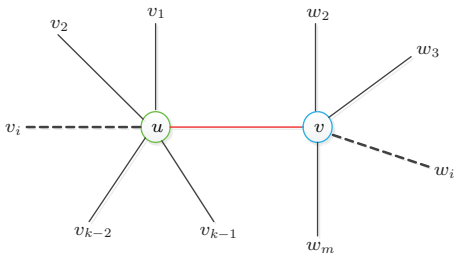

图 1 随机选择节点及其邻居节点度的分布分析示意图

Fig. 1. Schematic drawings of randomly chosen u and which neighbor nodes.

3.2.3 算法模型时间复杂分析

我们重点讨论的无标度网络是带有一类特性的复杂网络, 其典型特征是在网络中的大部分节点只和很少节点连接, 而有极少的节点 (称为中枢节点 Hubs) 与非常多的节点连接. 结点度数是自然数 k 的概率: $P(k)=ck^{-\gamma}$ [29]. 分析我们提出的 RMDN 算法, 其运行时间复杂度为 $O(k\log(n))$. 分析过程如下: 由于

$$\langle k\rangle=\sum_1^n kp(k)=\sum_1^n kck^{-\gamma}=c\sum_1^n\frac{1}{k^{\gamma-1}},$$

$$P(k)=ck^{-\gamma}, \tag{6}$$

当 $\gamma>2$ 时,

$$\langle k\rangle=c\sum_1^n\frac{1}{k^{\gamma-1}}\leqslant c\sum_1^n\frac{1}{k}=c\ln(n),$$

$$n\to\infty, \tag{7}$$

其中 $\langle k\rangle$ 表示网络节点的平均度数. 因此根据算法模型选取大小为 s 的初始节点集, 而对于每一次随机选取的节点我们都要查询其邻居节点的度数, 所以运行时间规模可表示为 $s\langle k\rangle=sc\ln(n)$, 那么 RMDN 算法的时间复杂度为 $O(s\log(n))$.

3.2.4 任意节点邻居的度数为 top-k 的概率

设定 $p_{\text{top-}k}$ 是从任意一条边出发遇到的节点度数大于等于 top-k (常称为中枢节点 (Hubs)) 的概率, $p_{\text{top-}k}$ 可以用积分形式表示为

$$p_{\text{top-}k}=\int_{k_{\text{top-}k}}^{k_{\max}}p_m\,\mathrm{d}m=\int_{k_{\text{top--}k}}^{k_{\max}}bm^{1-\gamma}\,\mathrm{d}m$$

$$=\frac{k_{\max}^{2-\gamma}-k_{\text{top-}k}^{2-\gamma}}{k_{\max}^{2-\gamma}-k_{\min}^{2-\gamma}}. \tag{8}$$

那么选取 s 个节点是 top-k 的 Hubs 节点的概率为 $1-(1-p_{\text{top-}k})^s$. 根据以上理论推导, 我们选

取网络节点数为$n = 10000$, $k_{\min} = 1$, 种子节点集合大小$1 \leqslant s \leqslant 30$, $2 < \gamma < 3$进行了模拟实验分析(结果如图2所示), 网络随着γ增大, 随机选取的s个传播源种子节点为整个网络中度数为top-k的概率逐渐下降; 随着节点个数s的增大, 命中top-k的概率也在增加.

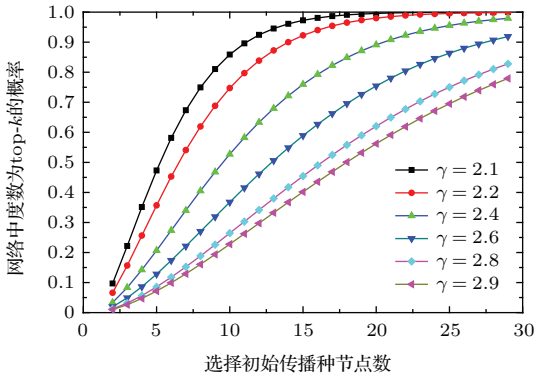

图2 随机选择一节点, 其邻居节点在最大度数节点为top-k的概率随不同γ的复杂网络分布情况

Fig. 2. The probabilityof a randomly chosen u being a hub node when size of seed is k.

3.3 RMDN++算法模型

由上3.2.4节点理论推导分析可知, 随着选取源种子节点集合的增多, RDMN算法选中top-k的Hubs节点的概率也明显增大, 在不增加算法的时间复杂度的情况下, 我们进行一步提出了RMND++算法模型(如算法3所示), 基本思想是

算法 3 RandomMaxDegreeNeighbor++(G, k, α)

initialize $S = \emptyset$
while len$(S) < \alpha s$ do
  random chose vertex $u \in V \backslash S$
  select u = argmaxDegree $\{u \cup$ neighbor$(u)\}$
  $S = S \cup \{u\}$
end while
T = argmaxDegree_$K\{S\}$ and $|T| = s$
output T.

在基于RMND的基础上, 首先, 扩大$\alpha(\alpha \geqslant 1)$倍可选择预备源种子节点集合, 然后再从$\alpha s$集合中选出$s$个度数最大的节点作为最终传播源种子节点, 算法的时间复杂度为$(O(s\log(n) + \alpha s))$, 与RMND相近, 为了保持文章的统一性, 本文中算法取$\alpha = 2$.

4 实验与结果分析

考虑到不同的社会网络类型代表不同的网络拓扑结构特性, 我们选取具有不同γ的实际社会网络中进行了相关算法的比较分析, 实验结果显示即使我们不了解整个网络信息的情况下, 只需知道选择节点及其邻居节点信息, 通过在IC模型与IT模型实验, 最大影响力效果与现有效果较好的典型贪心算法运行效果近似, 且有时还略优. RDMN与RDMN++算法的时间复杂度仅为$(O(s\log(n)))$, 比目前时效性最优的度中心化启发式算法$(O(m))$有了显著提升, 运行时间随着网络规模的扩大速度提升呈线性增长$(m/s\log(n))$; 而且算法模型简单, 可适合用性, 可操作性更强.

本文所有实验运行的硬件环境是: 处理器Intel®Core™i5 CPU M430 @2.27 GHz, 内存(RAM) 3 GB.

4.1 实验数据

由于不同类型的社会网络通常具有相似的网络结构特征, 我们选取了4个具有不同γ的实际社会网络中进行实验分析比较, 表3给出各个网络的属性特征: 1) 全美航空网的网络拓扑数据来自于文献[31], 它是一个典型的无标度、小世界网络. 网络中的一个节点代表一个城市的航空港, 如果两个航空港之间有航班往来, 则两个节点之间有一条连边; 2) 全球最大的社交网络Facebook中部分用户关系网络[32]; 3) Blogs网络数据[33], MSN博客空间中交流的关系网络, 4) Twitter用户签到数据[34].

表3 现实社会网络的属性情况

Table 3. The basic topological features of the four real networks.

Network Name	n	m	$\langle k\rangle$	$k_{\max}$	$k_{\min}$	d	γ
USAir97	332	4252	25.61	139	1	2.738	1.821
Blogs	3982	6803	3.42	189	1	6.227	2.453
Facebook	4039	88234	43.69	1045	1	3.692	2.510
Twitter	554372	2402720	4.33	11443	1	9.827	2.638

其中 n 是网络中节点数, m 为边数, $\langle k\rangle$ 表示网络中平均度数据, $k_{\max}$ 为节点中最大度数, $k_{\min}$ 为节点中最小度数, d 为节点之间最短路径的平均数, γ 为网络中度分布的幂指数值.

4.2 实验效果

我们知道无标度网络中节点度分布的幂律特征在对数坐标系中, 一般将会是一条斜率介于 -2 到 -3 之间的直线. 根据 Clauset [35] 和 Barabasi [36] 提出的幂律拟合极大似然估计方法和 KS 统计量拟合幂律分布 γ 指数, 图 3 给出了 4 个真实社会网络度分布情况的幂律拟合分布情况. 很明显地看到网络中的大部分节点只和很少节点连接, 而有极少的节点与非常多的节点连接这一特征, 具有 20/80 特性 [37].

我们应用 IC 模型和 LT 模型, 对本文所提出的 RDMN, RDMN++ 算法与已有的典型算法在 4 个真实社会网络的传播范围情况进行了分析比较. 为了保持实验的易读性, 实验模拟传播过程中, 每次选取传播种子节点 $1 \leqslant s \leqslant 30$ 作为传播源的种子集合, 传播概率取 $p = 0.01$(如果节点的传播能力很强, 很难区分单个个体的重要性), 传播范围取 10000 次迭代的均值, 横轴为根据各算法中传播影响力最大的节点进行排序所得的集合大小, 纵轴为相应所选传播种子节点的传播范围大小. 时间复杂度分析统一取 $s = 30$, 时间取迭代运行 10000 次的平均运行时间.

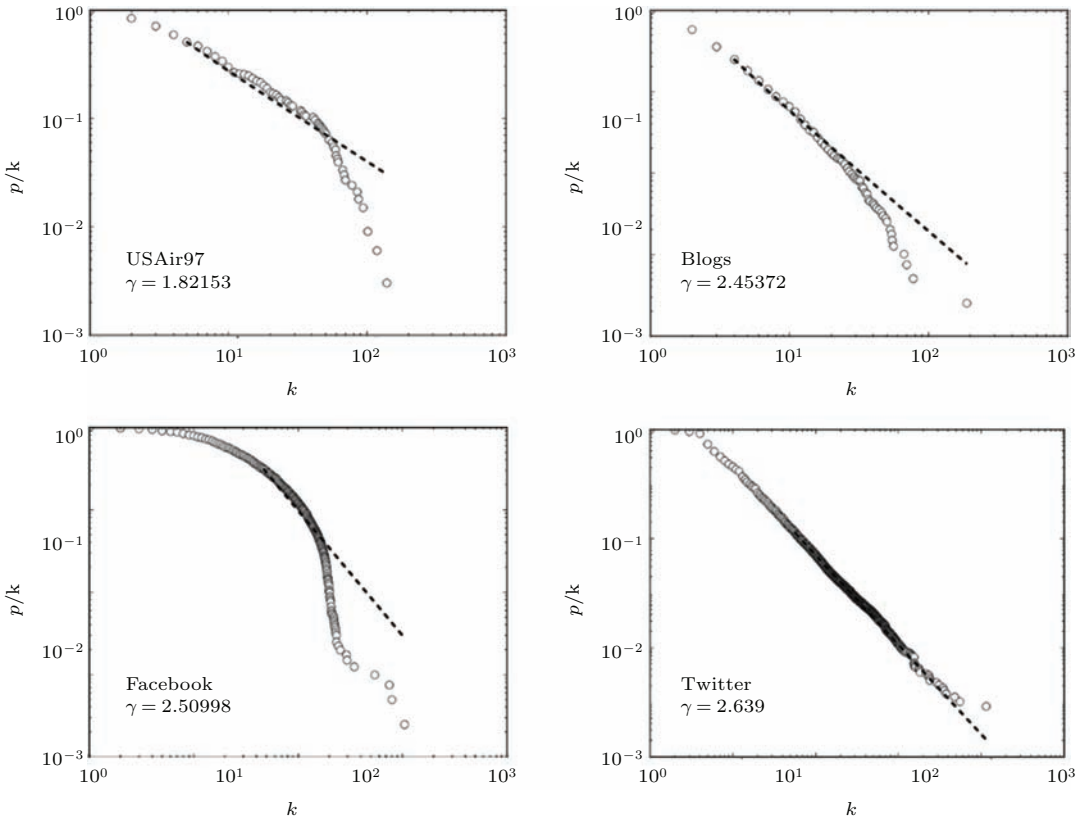

图 3 美国航空线路网络 USAir97 ($\gamma = 1.82$)、博客 Blogs($\gamma = 2.45$)、社交网络 Facebook($\gamma = 2.509$), Twitter 签到数据 ($\gamma = 2.638$) 的互补累积分布函数和拟合幂律指数及对数正态分布情况

Fig. 3. Estimation the value of γ with datasets in Table 3.

从图 4 到图 8 显示可以看出, RandomHeuristic 算法运行影响力传播最大化效果是最差的, 但是运行时间是最快的; DegreeDiscountIC, SingleDiscount 算法虽然在传播范围比 DegreeHeuristic 算法要稍优一点, 但是运行时间也较 DegreeHeuristic 算法长, 整体而言 DegreeHeuristic 算法的时效性较优. 从图 4、图 8 (a) 可以看出我们提出的 RDMN 算法与几种算法运行效果接近, 有时还略优; 但运行时间比 DegreeHeuristic 算法快了 3.4 倍, 特别指出的是在 USAir97 美国航空网络中 General Greedy

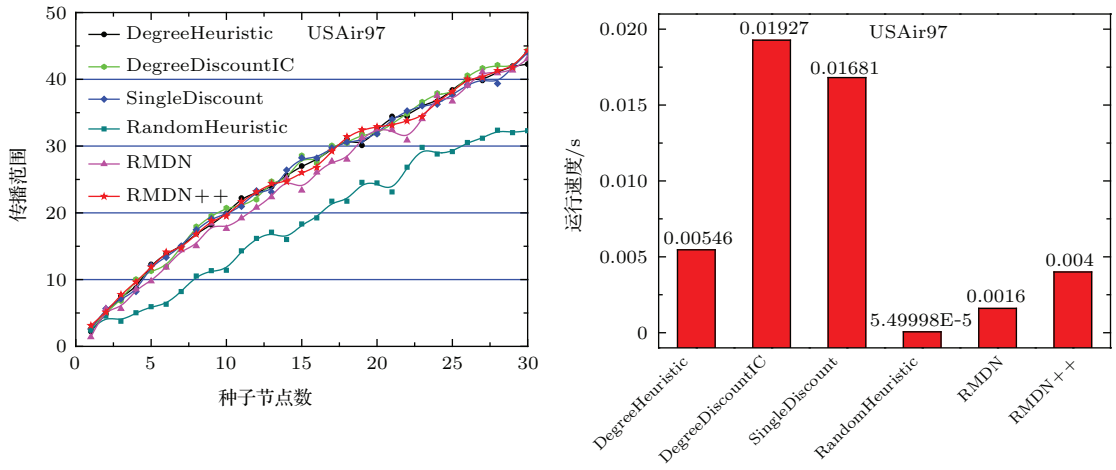

图 4 (网刊彩色) IC 模型下不同算法在美国航空线路网络 (USAir97) 中运行时效分析比较; 其中左图代表传播影响力最大化分析比较, 右图代表运行时间复杂度比较 ($n = 332$, $m = 4252$, $p = 0.01$, $1 \leqslant s \leqslant 30$)

Fig. 4. (color online) Influence spreads and running times of different algorithms on the collaboration graph USAir97 under the independent cascade model ($n = 332$, $m = 4252$, $p = 0.01$, $1 \leqslant s \leqslant 30$).

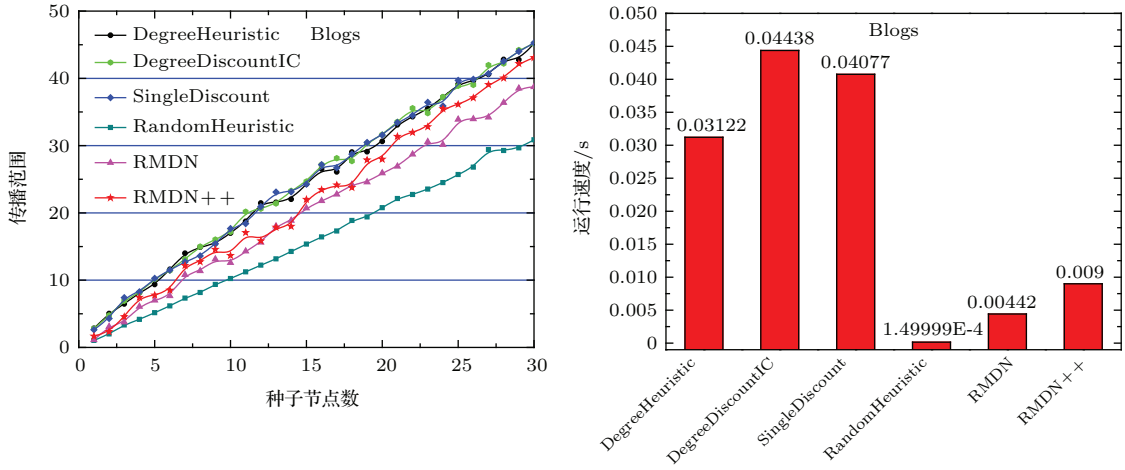

图 5 (网刊彩色) IC 模型下不同算法在博客网络 (Blogs) 中运行时效分析比较; 其中左图代表传播影响力最大化分析比较, 右图代表运行时间复杂度比较 ($n = 3982$, $m = 6803$, $p = 0.01$, $1 \leqslant s \leqslant 30$)

Fig. 5. (color online) Influence spreads and running times of different algorithms on the collaboration graph Blogs under the independent cascade model ($n = 3982$, $m = 6803$, $p = 0.01$, $1 \leqslant s \leqslant 30$).

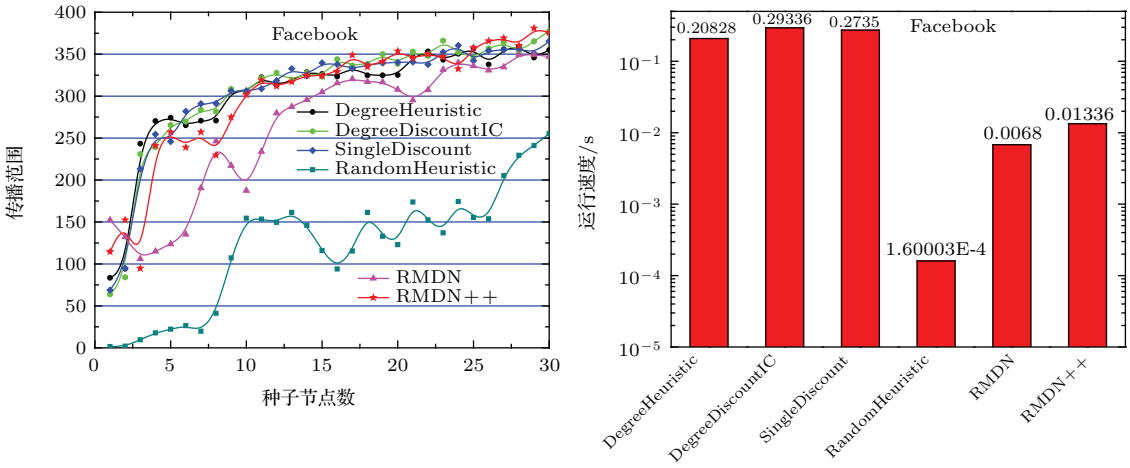

图 6 (网刊彩色) IC 模型下不同算法在社交网络 (Facebook) 中运行时效分析比较; 其中左图代表传播影响力最大化分析比较, 右图代表运行时间复杂度比较 ($n = 4039$, $m = 88234$, $p = 0.01$, $1 \leqslant s \leqslant 30$)

Fig. 6. (color online) Influence spreads and running times of different algorithms on the collaboration graph Facebook under the independent cascade model ($n = 4039$, $m = 88234$, $p = 0.01$, $1 \leqslant s \leqslant 30$).

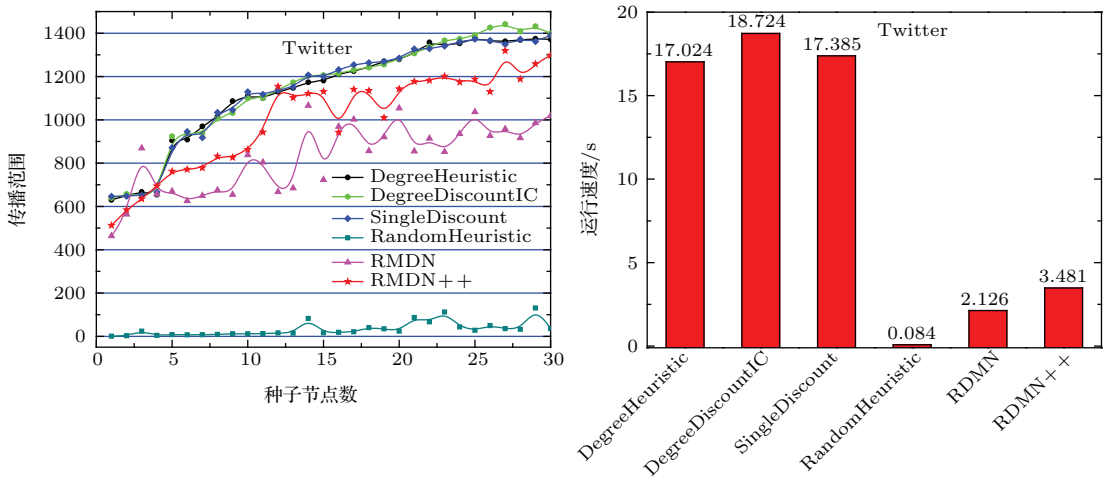

图 7 (网刊彩色) IC 模型下不同算法在社交网络 (Twitter) 中运行时效分析比较; 其中左图代表传播影响力最大化分析比较, 右图代表运行时间复杂度比较 ($n = 554372$, $m = 2402720$, $p = 0.01$, $1 \leqslant s \leqslant 30$)

Fig. 7. (color online) Influence spreads and running times of different algorithms on the collaboration graph Twitter under the independent cascade model ($n = 554372$, $m = 2402720$, $p = 0.01$, $1 \leqslant s \leqslant 30$).

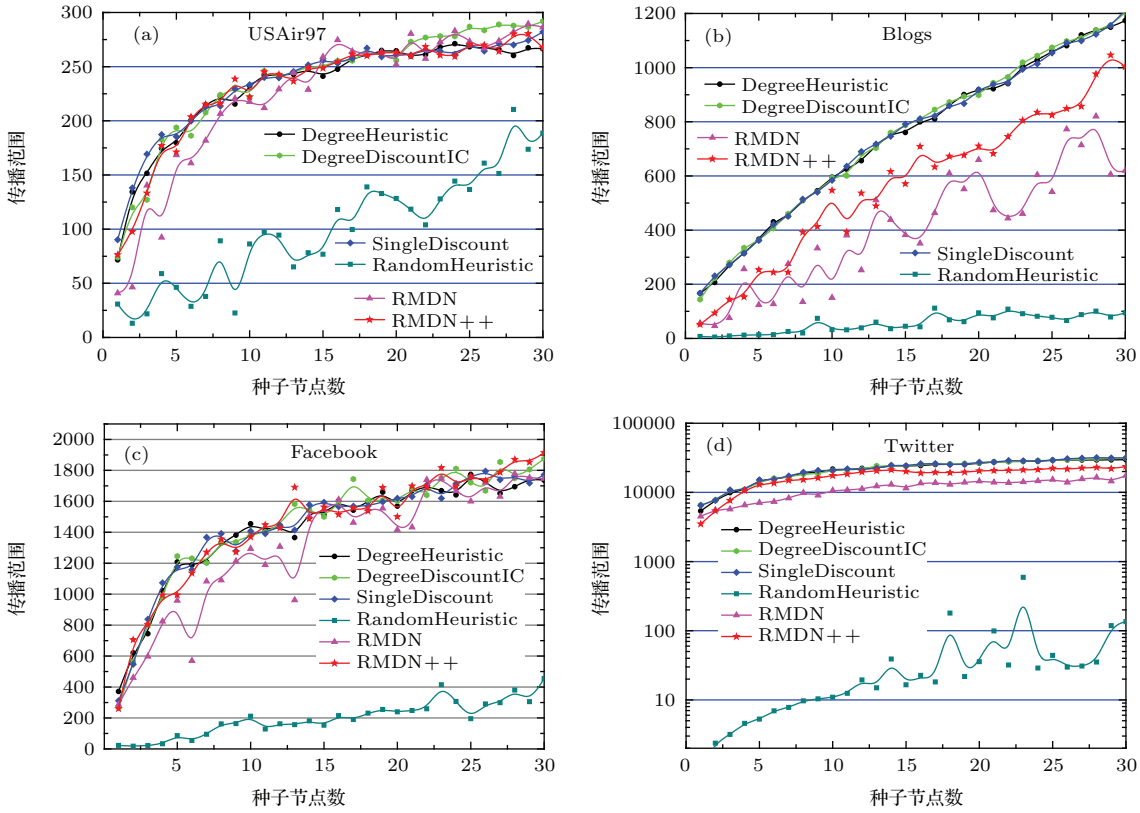

图 8 (网刊彩色) LT 模型下不同算法在 4 个网络中运行传播范围分析比较 ($1 \leqslant s \leqslant 30$) (a) 美国航空线路网络 (USAir97) 运行结果; (b) 代表博客网络 (Blogs) 运行结果; (c) 代表社交网络 (Facebook) 运行结果; (d) 代表社交网络 (Twitter) 运行结果

Fig. 8. (color online) Performance under the linear threshold model with four datasets.

算法的影响力最大化运行时间为 355.575 s, 而 RDMN 算法是 0.0016 s, 速度提高了 2.2×10^5 倍以上. RDMN++ 算法整体优于 RDMN 算法, 有时运行效果比其他经典算法还要好, 因为对整个网络进行随机选择符合随机抽样原理, 所以所得的种子节点也更能代表整个网络, 而不至于陷入“富人俱乐部现象”局部传播之中; 且运行时间只是 RDMN 的 2.5 倍.

同样, 为了保证我们对所提出的算法有较好的应用性, 下面给出了在 LT 模型下进行分析比较, 从

图8可以看出, 整个算法运行效果与IC模型下运行结果相似, 可以看出我们所提出的算法的适应性较好.

从图4到图8可以看出我们所提出的RDMN, RDMN++算法在不同γ特征的网络中, 只是在了解局部的节点信息, 算法选出的影响力最大化种子节点在IC模型、LT模型下运行效果接近于必须全面了解网络整个信息的DegreeHeuristic算法, 但运行时间复杂度随着网络规模的扩大, 运行时间从几倍到几十倍地提高, 由时间复杂度$(O(s\log(n))$可知当网络连边的规模达到10^8以上, 运行时间增长倍数理论上为$m/s\log(n)$ (一般来说边$m\sim n^2$, 也就说理论上可以大约可以提升10^{13}右左), 增速非常地明显.

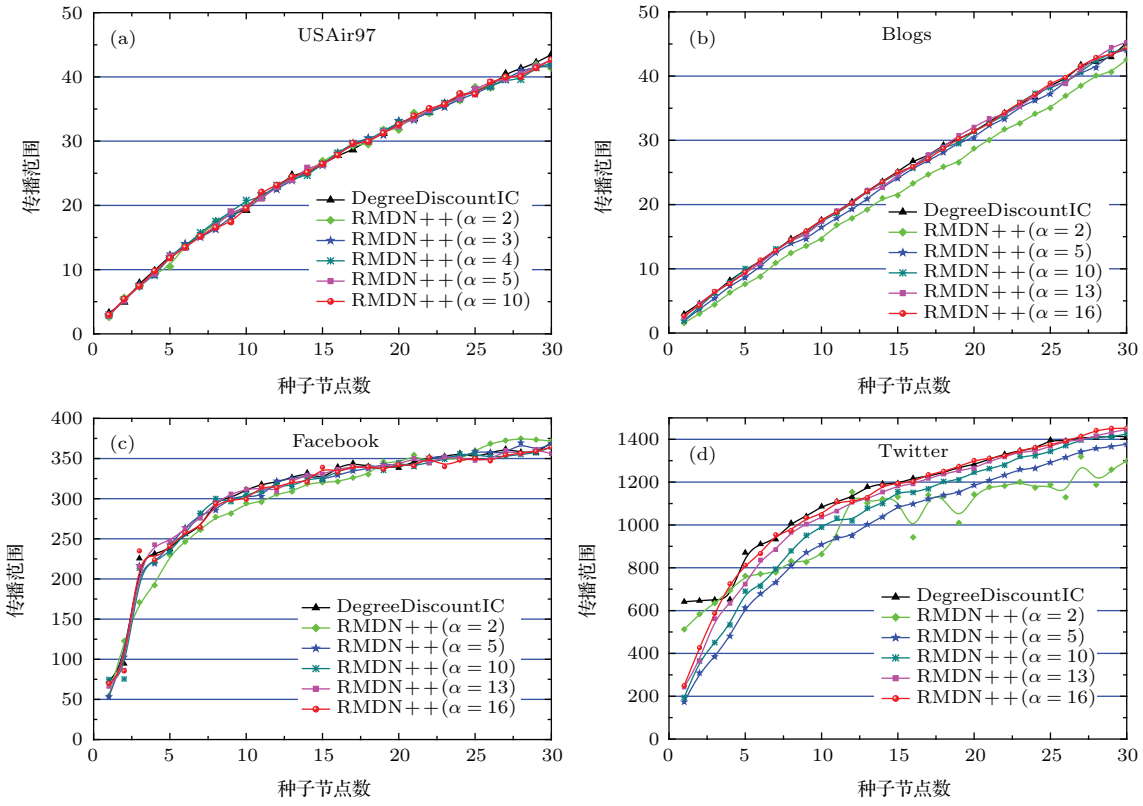

图9 (网刊彩色) RMDN++算法中取不同的α与DegreeDiscountIC算法在4个复杂网络中运行IC模型进行传播范围分析比较

Fig. 9. (color online) Performance under the independent cascade model with different value of α.

图9我们分析了在IC模型中对RMDN++算法中选取不同的α值与DegreeDiscountIC算法运行效果的比较分析结果, 实验显示RMDN++算法在USAir97美国航空网络中$\alpha=2$, Blogs网络中$\alpha=5$, Facebook网络中$\alpha=2$及在Twitter网络中$\alpha=13$时传播范围已很快近似或超过DegreeDiscountIC算法的运行效果. 因此可以看出RMDN++算法模型在不同类型的复杂网络中的只需要较小的经验值α, 就能取得很好的影响力传播效果.

以上我们对所提出的影响力最大化算法在现实社会网络进行了实验分析比较, 1) 充分论证了我们所提出的算法与理论推导高度一致性; 2) 我们只需知道随机选择节点及其直接连接邻居节点信息, 巧妙地避开了必须了解全局节点信息的问题, 且该算法执行结果与现在典型算法接近, 且运行时间复杂度有了明显提高; 3) 算法在IC模型和LT模型两个不同模型中运行效果相似, 可以看出算法适用性较强; 4) 我们提出的算法实际应用极其简单, 可行性、适用性更强.

5 结 论

在现实社会中影响力最大化问题可帮助我们提高新知识、新产品的传播有效范围, 同时也可以有效的预测、分析和控制疾病传播、流言散布、

计算机病毒扩散. 在给定的有限预算前提下, 在复杂网络中找出影响力最大化传播种子集合一直以来都是研究的热点与难点, 我们提出了RDMN, RDMN++算法模型. 我们不但从现实生活中常见的4种领域, 具有不同幂指数γ网络特征的复杂网络上实验证实了所提出算法时效性, 验证了算法的高效性和可行性, 而且给出了相应的理论分析推导证明.

通过实验分析结果与几个典型影响力最大化贪心算法相比, 我们所提出的算法虽然运行效果接近或稍差一点, 但是算法的运行时间随着网络规模的增加, 时间复杂度的优势显著. 且我们只需知道选择节点及其直接连接邻居节点的局部信息, 巧妙地避开了必须知道全局节点信息为前提的问题, 这使模型算法的适用性更广, 可操作性更强. 我们所提出的算法为这项具有挑战性研究提供了新的算法思路.

参考文献

[1] Watts D J, Strogatz S H 1998 *Nature* **393** 440
[2] Barabási A L, Albert R 1999 *Science* **286** 509
[3] Barabási A L, Albert R, Jeong H, Bianconi G 2000 *Science* **287** 2115a
[4] Lü, L, Zhang Y C, Yeung C H, Zhou T 2011 *PloSone* **6** e21202
[5] Hu Q C, Yin Y S, Ma P F 2013 *Acta Phys. Sin* **62** 140101 (in Chinese) [胡庆成, 尹龑燊, 马鹏斐 2013 物理学报 **62** 140101]
[6] Ren Z M, Shao F, Liu J G 2013 *Acta Phys. Sin* **62** 128901 (in Chinese) [任卓明, 邵凤, 刘建国 2013 物理学报 **62** 128901]
[7] Aral S, Walker D 2012 *Science* **337** 337
[8] Liu J G, Ren Z M, Guo Q 2013 *Physica A: Statistical Mechanics and its Applications* **392** 4154
[9] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A 2010 *Nature Physics* **6** 888
[10] Ren X L, Lü L Y 2014 *Chin. Sci. Bull.* **59** 1175 (in Chinese) [任晓龙, 吕琳媛 2014 科学通报 **59** 1175]
[11] Liu J G, Ren Z M, Guo Q, et al 2013 *Acta Phys. Sin.* **62** 178901 (in Chinese) [刘建国, 任卓明, 郭强等 2013 物理学报 **62** 178901]
[12] Domingos P, Richardson M 2001 *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* San Francisco, CA, USA, August 26–29, 2001 p57
[13] Richardson M, Domingos P 2002 *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* Edmonton, Alberta, Canada, July 23–26, 2002 p61
[14] DKempe, JKleinberg, ETardos 2003 *Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining* New Washington, DC, USA, August 24–27, 2003 p137
[15] Leskovec J, Krause A, Guestrin C 2007 *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* San Jose, CA August 12–15, 2007 p420
[16] Goyal A, Lu W, Lakshmanan L V S 2011 *Proceedings of the 20th international conference companion on World wide web* Johannesburg,South Africa Sep 14–16, 2011 p47
[17] Zhou C, Zhang P, Guo J *2013 Data Mining (ICDM),2013 IEEE 13th International Conference on. IEEE* Dallas, Texas, USA Dec 8–11, 2013 p907
[18] Chen W, Wang Y, Yang S2009 *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* Paris, France June 28–July 1, 2009 p199
[19] Kimura M, Saito K 2006 *Knowledge-Based Intelligent Information and Engineering Systems* Bournemouth U K, October 9–11, 2006 p937
[20] Chen W, Wang C, Wang Y *2010 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data miningWashington*DC, USA, July 25–28, 2010 p1029
[21] Wang Y, Cong G, Song G 2010 *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data miningWashington* DC, USA, July 25–28, 2010 p1039
[22] Galstyan A, Musoyan V, Cohen P 2009 *Phys. Rev. E* **79** 056102
[23] Li D, Xu Z M, Chakraborty N 2014 *PloS one* **9** e102199
[24] Zhou S, Mondragón R J 2004 *Commun. Lett.* **8** 180
[25] Bonacich P 1972 *Journal of Mathematical Sociology* **2** 113
[26] Milgram S 1967 *Psychology today* **2** 60
[27] Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S 2012 *ACM Web Science 2012: Conference Proceedings* Evanston, Illinois, USA, June 22–24, 2012 p45
[28] Six Degrees of Separation, Twitter Style *fromSysomos* Apr 30, 2010
[29] Cohen R, Havlin S 2003 *Phys. Rev. Lett.* **90** 058701
[30] Newman M E J, Strogatz S H, Watts D J 2001 *Phys. Rev. E* **64** 026118
[31] Batagelj V, Mrvar A 2006 *Pajek datasetsWeb page http://vlado. fmf. uni-lj.si/pub/networks/data*
[32] Leskovec J, Mcauley J J 2012 *Advances in neural information processing systems* South Lake Tahoe, Nevada, United States, December 3–6, 2012 p539
[33] Xie N 2006 *Social network analysis of blogs M.Sc. Dissertation,* University of Bristol
[34] Li G, Chen S, Feng J 2014 *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* New York, NY, USA, June 22–25, 2014 p87
[35] Clauset A, Shalizi C R, Newman M E J 2009 *SIAM Rev.* **51** 661
[36] Barabási A L, Albert R, Jeong H 1999 *Physica A* **272** 173
[37] Newman M E J 2005 *Contemporary Physics* **46** 323

A new approach for influence maximization in complex networks*

Hu Qing-Cheng† Zhang Yong Xu Xin-Hui Xing Chun-Xiao
Chen Chi Chen Xin-Hua

(*Research Institute of Information Technology, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*)

(Received 8 December 2014; revised manuscript received 10 June 2015)

Abstract

Influence maximization modeling and analyzing is a critical issue in social network analysis in a complex network environment, and it can be significantly beneficial to both theory and real life. Given a fixed number k, how to find the set size k which has the greatest influencing scope is a combinatory optimization problem that has been proved to be NP-hard by Kempe *et al.* (2003). State-of-the-art random algorithm, although it is computation efficient, yields the worst performance; on the contrary, the well-studied greedy algorithms can achieve approximately optimal performance but its computing complexity is prohibitive for large social network; meanwhile, these algorithms should first acquire the global information (topology) of the network which is impractical for the colossal and forever changing network. We propose a new algorithm for influence maximization computing-RMDN and its improved version RMDN++. RMDN uses the information of a randomly chosen node and its nearest neighboring nodes which can avoid the procedure of knowing knowledge of the whole network. This can greatly accelerate the computing process, but its computing complexity is limited to the order of $O(k\log(n))$. We use three different real-life datasets to test the effectiveness and efficiency of RMDN in IC model and LT model respectively. Result shows that RMDN has a comparable performance as the greedy algorithms, but obtains orders of magnitude faster according to different network; in the meantime, we have systematically and theoretically studied and proved the feasibility of our method. The wider applicability and stronger operability of RMDN may also shed light on the profound problem of influence maximization in social network.

Keywords: complex network, influence maximization, information diffusion, greedy algorithm

PACS: 01.75.+m, 05.90.+m, 89.75.–k, 89.70.Hj **DOI:** 10.7498/aps.64.190101

* Project supported by the National Basic Research Program of China (Grant No. 2011CB3023302) and the National High Technology Research and Development Program of China (Grant No. SS2015AA020102).

† Corresponding author. E-mail: hqc10@mails.tsinghua.edu.cn