Cost-aware Targeted Viral Marketing in Billion-scale Networks

背景

在线社交网络对营销和广告具有重要意义，可以通过说词的方式迅速广泛的传播信息，即病毒营销。它可以通过少量关键用户影响近十亿的用户，为了鉴别关键用户，大量的工作聚焦于影响最大化问题，即寻找k个种子用户的集合，最大化影响期望。但是这里存在两个不切实际的假设：

- 任何用户有同样的选择成本
- 所有用户对某一信息兴趣一样

这会导致最终得到不切实际的解或者误导公司去影响错误的用户，这些用户可能对公司产品不感兴趣也没有潜在收益。

解决的具体问题

- 解决了IM拥有不切实际假设的问题。
- 解决了扩展RIS解决CTVM问题不适应大型网络的问题。

解决方法

CTVM

作者提出了新的问题定义，具有成本意识的病毒营销，它兼顾了选择节点具有任意成本和影响节点具有任意收益两个方面。

给定一个图$G = (V, E, c, b, w)$，其中V是社交网络的节点集，E是有向边集，c是选择某个节点的成本，b是影响某个节点的收益，w是边$(u, v) \in E$的影响权重。

CTVM定义为给定一个图$G = (V, E, c, b, w)$和一个预算$B > 0$，寻找一个种子集合$S \subset V$，且$cost(S) \leq B$，最大化收益$\mathbb{B}(S)$。CTVM涵盖了如下病毒营销问题：

- Influence Maximization: IM是CTVM特殊的情况，
  $c(u) = 1\ and\ b(u) = 1\ \forall u \in V$。
- Budgeted Influence Maximization: 寻找成本不超过B的k个节点的集合最大化$\mathbb{I}(S)$，$b(u) = 1\ \forall u \in V$。
- Target Viral Marketing: 在一个目标集合T中寻找一个K个节点的集合，最大化影响节点的数量。
  $c(u) = 1\ \forall u \in V and\ b(v) = 1\ if\ v \in V, b(v) = 0\ otherwise$。

BCT

RIS

1. RIS Framework
    - 随机超边

      $G=(V,E,w)$，一个随机超边ε_j从G生成，通过（1）选择一个随机节点$v\in V$（2）生成一个样本图$g\subseteq G$（3）返回一个在g中能够到达v的节点集合作为ε_j。

    - RIS框架

      随机超边ε_j显然包含能够影响v的节点，如果随机生成多个随机超边，那么具有影响力的节点大概率会出现在随机超边上，所以，如果一个种子集合尽可能覆盖这些随机超边，那么可能取得最大影响传播。

      因此IM问题可以用如下框架来解决：

        i. 从G中生成多个随机超边
        ii. 用贪心算法来解最大覆盖问题以选择覆盖最多数量随机超边的种子集合S作为解。
2. Extending Problem

    如果把覆盖最多的随机超边改为覆盖最多的具有最大权重的随机超边即可用于CTVM问题，但作者证明这种方式并不适应大型的网络。作者提出了自己新的解决方式，和最大权重覆盖相结合构成了BCT。

BCT

1. BSA

    如果说RIS选取随机超边的方式是随机搜索的话，那么BSA更像是爬山法的启发式搜索。BSA的过程是：（1）选取一个利益率最大的节点u，$P(u)=b(u)/\sum_{v\in V}b(v)$（2）尝试选择它的入邻居节点$v$，根据线性阈值模型的传播方式使得$(v,u)$变成活边（2）然后相当于让$u$移动到$v$的位置，继续该过程。直到回到先前访问的顶点或者没有边被选择，算法停止。

    这样的选择方式，好像是说如果选择了一个利益率较大的点，那么他周围的节点利益率也相对较大。这样在选择节点的时候同时考虑了它的利益，而这个思想更像是贪心爬山算法。

2. Weight-Max-Coverage

    通过BSA算法得到了随机超边以后，再通过最大权重覆盖算法选取种子集合S:
    （1）让S为空

    （2）从V/S中选取满足$c(v)\leq B-c(S)$且使得S的最大权重覆盖率最大的节点v加入S

    （3）重复以上过程直到不满足（2）的条件为止

    （4）选择最大权重覆盖的一个节点u

    （5）比较S和u的最大权重覆盖数量，选择最大的作为解返回。

3. BCT

    （1）设置超图$\mathcal{H}=(\mathcal{V},\varepsilon=\phi)$

    （2）遍历设置的停止阈值Λ_L次，每次通过BSA算法生成一个随机超边ε_j加入ε

（3）通过最大权重覆盖算法生成候选解S，如果S的最大权重和覆盖值小于停止阈值，那么让循环次数为停止阈值的两倍并减去$|\varepsilon|$(减少无意义的循环次数)，重复（2），直到候选解的最大权重覆盖值大于停止阈值为止。

4. Complexity

作者证明其时间复杂度为$O((ln(1/\delta)+lnM_k)\epsilon^{-2}n)$

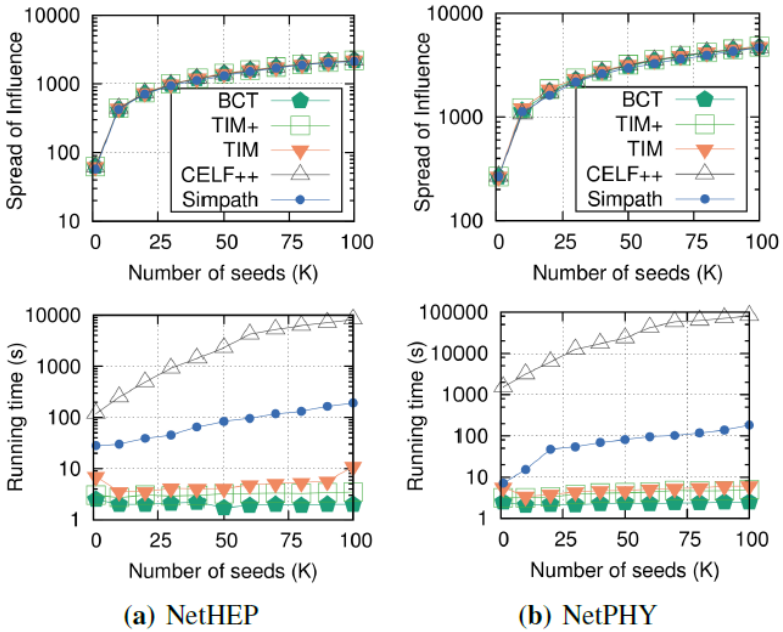
数据集

Dataset	#Nodes	#Edges	Type	Avg. degree
NetHELP [3]	15K	59K	undirected	4.1
NetPHY [3]	37K	181K	undirected	13.4
Enron [22]	37K	184K	undirected	5.0
Epinions [3]	132K	841K	directed	13.4
DBLP [3]	655K	2M	undirected	6.1
Twitter [23]	41.7M	1.5G	directed	70.5

实验分析

1. 影响最大化任务：TIM，TIM+，CELF++，SIMPATH

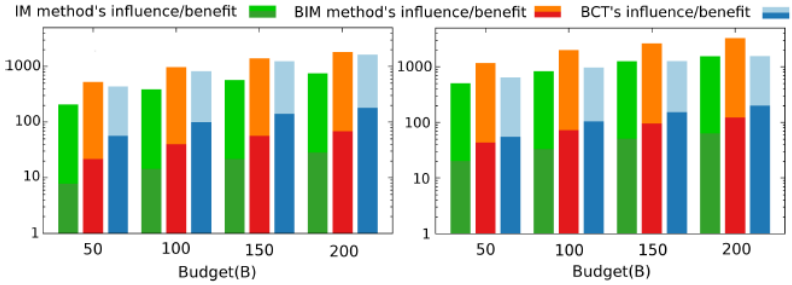
Method	Spread of Influence			Running Time (s)		
	Epin.	*Enron*	*DBLP*	*Epin.*	*Enron*	*DBLP*
BCT	**16320**	**16776**	**108600**	**3**	**2**	**5**
TIM+	16293	16732	108343	6	3	12
TIM	16306	16749	107807	8	4	17
Simpath	16291	16729	103331	23	18	136



(a) NetHEP (b) NetPHY

通过表图可以看出，在IM task上，在不同的数据集上，BCT无论在影响传播还是运行时间上，都要比先进的算法效果要好。

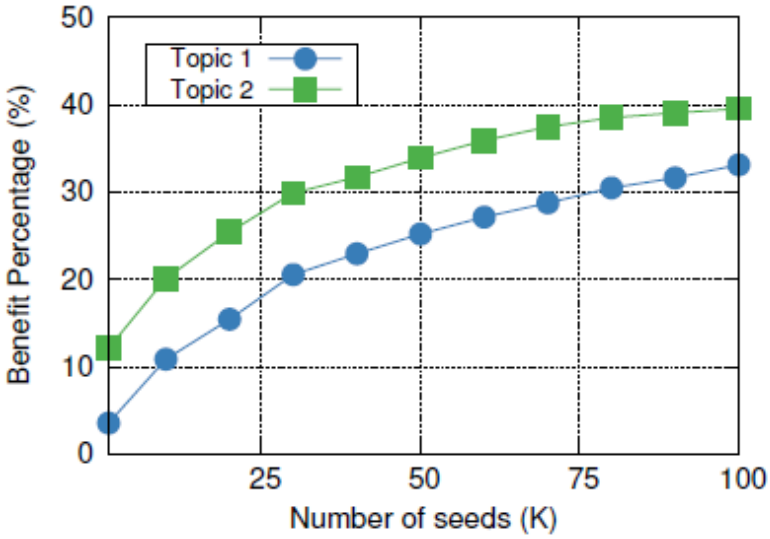
2. CTVM：基于IM的方法，基于BIM的方法和BCT



可以看出，在CTVM问题上，影响传播的效果BIM要好一些，但在收益最大化方面，BCT效果更好。

3. BCT

用BCT针对两个推特热门话题并附带关键词，用推特数据做实验去提取在推文中涉及到话题关键词的用户列表，和他们发出的有关该话题的推文数量。将用户的这些推文数量作为节点的收益。



随着选择用户的总结，收益比逐渐增加，但速率趋缓。

Topic	Keywords	#Users	First 5 selected Twitters
1	**bill clinton, iran, north korea, president obama, obama**	997K	dominiquerdr, stockmarketcash, uncoolbobby, larsthebear, dadashiii
2	**senator ted kenedy, oprah, kayne west, marvel, jackass**	507K	royasmusic, bksmarvelous1, edithayala, capitarecesion, dietmission

这是BCT在两个话题上最终选取的用户和最初选取的用户，开始选取的用户虽然只有几千粉丝，但是他们的推文数量特别多。

其它解决方法

- IM

  固定种子节点个数，存在两个不切实际的假设，导致解会出现一定的错误。

- BIM

  只考虑了任意选择成本。不适应十亿级别的数据集。

- TVM

  用于影响一部分用户的模型，基于启发策略，但不能提供性能保证。

创新之处

1. 问题上的创新：提出了一种兼顾选择节点具有不同成本和影响节点具有不同收益的新病毒营销问题：CTVM
2. 现有方法上的创新：在RIS框架的基础上提出了BCT来解决CTVM问题，并提出了新的抽样方式BSA来解决RIS不适应大型网络的问题，且给出了有效的停止规则。