Why approximate when you can get the exact? Optimal Targeted Viral Marketing at Scale

---

背景

信息通过说词的方式在OSNs上广泛传播，这被称为病毒营销，影响传播可以从一些关键用户到数十亿用户。为了确定这些关键用户，IM问题被提出，即寻找k个种子节点结合，最大化它的影响节点数量的期望值。之后，CTVM被提出，它兼顾了选取节点的成本和影响节点的收益两个方面。

对于IM和CTVM问题的解决方法都存在可扩展性问题和性能保证有限问题。

解决的具体问题

- 解决了传统随机规划用于CTVM问题时的可扩展性问题，即不适应大型网络。
- 解决了目前方法给出的解的性能保证有限问题，即不能精确的求解IM或者CTVM问题

解决方法

TipTop

作者提出了TipTop来解决传统随即规划的可扩展性问题，并保证了$(1-\epsilon)$的近似率。

TipTop的大致思想为：

1. 设置一个小样本集合$\mathcal{R}$的大小Λ，同时通过增加样本的方式动态的判断样本大小，确保用足够的样本去寻找一个接近最优的候选解。
2. 努力让ILP的规模尽可能小，这需要尽可能避免执行增加样本的过程。如果候选解的值离期望值过远，调用增加样本过程允许加入大批量样本，以便扩大查找空间寻找更好的候选解。
3. TipTop使用BSA来选取随机RR集合来作为样本集合，因为BSA选择节点源的概率与利益成正比，这符合CTVM考虑任意利益的特点。

ILP_{MC}

TipTop用ILP来求解集合最大覆盖问题而不是以往的贪心算法。这主要是对于近似率保证和处理更多样本的取舍。

整个ILP可以看作是SAA，企图在随机生成样本上寻找最优解，但它和SAA最大的不同是，ILP不需要考虑所有边的状态，只是在局部节点源的RR集合上执行，减少了每个样本的大小，这对ILP有重大影响。

Verify and IncreaseSample Procedures

验证过程大致思想是：持生成RR集合去评估候选解的值，直到在误差范围内或者达到最大生成样本数量为止。

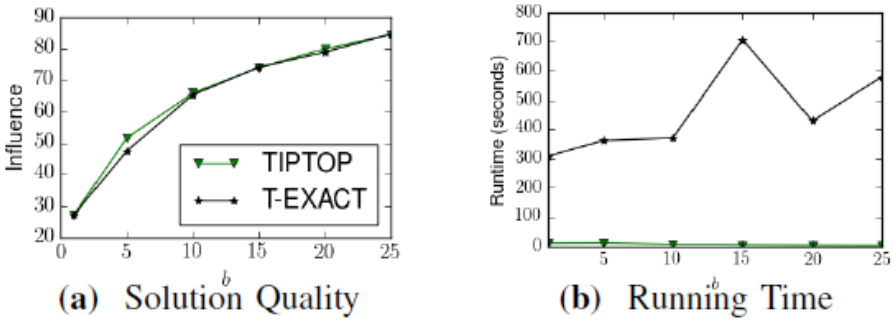
如果候选解没有通过验证过程，TipTop就继续调用增加样本过程来增加查找空间RR集合的大小以便找到更好的候选集继续验证。它不像SSA那样每次将$\mathcal{R}$的大小翻倍，而是通过参数小心的决定增加样本后样本集合的大小。

数据集

Dataset	Network Type	Nodes	Edges
US Pol. Books [21]	Recommendation	105	442
GR-QC [22]	Collaboration	5242	14496
Wiki-Vote [22]	Voting	7115	103689
NetPHY [23]	Collaboration	37149	180826
Twitter [24]	Social	41M	1.5B

实验

1. Comparison to the T-EXACT IP



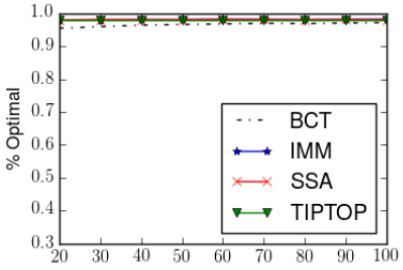
(a) Solution Quality (b) Running Time

由于T-EXACT的可扩展性问题，只在105节点的US Pol数据集上实验，通过图可以看出TipTop解的质量要比T-EXACT好，而且运行速度快了100倍。

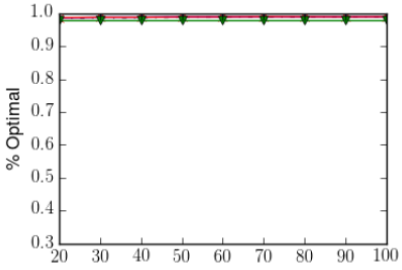
2. Benchmarking Greedy Methods

在独立级联模型下，边权重$w(u,v) = 1/d_{in}(v)$。涉及比较的算法：BCT，IMM，SSA。

首先考虑IM问题，不对算法作任何改动。



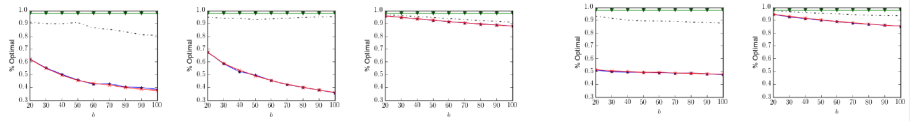
(a) Unweighted NetPHY



(b) Cost-Aware NetPHY

可以看出在IM问题上。贪心的算法表现要比他们的近似率高很多。而且在CTVM问题上，得到的结果都很相似。这表明了十年来的改进起了显著作用。

然后在不同情况下，用不同的方式设置成本，得到结果。分别为0-1均匀随机分布和基于粉丝数量的线性和对数函数情况。



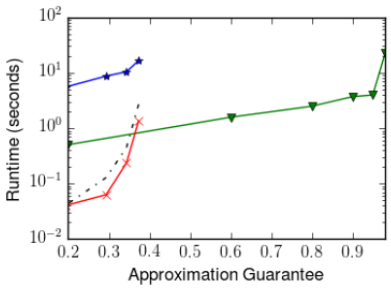
(a) GR-QC (b) Wiki-Vote (c) NetPHY (a) Uniform Random (b) Log-Outdegree

可以看到在任何情况下，TipTop的表现性能一直很高并且随着cost的增加表现十分稳定。

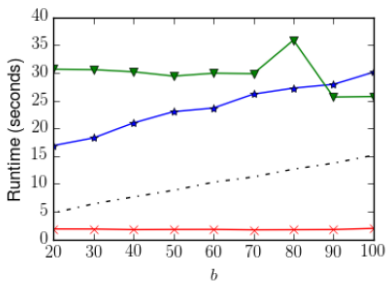
Coverage	Unweighted	Cost-Aware	CTVM
IMM	6.94×10^6	3.96×10^6	8.59×10^6
SSA	2.01×10^6	2.81×10^6	3.42×10^6
BCT	9.34×10^6	4.15×10^6	6.23×10^6
TIPTOP	2.08×10^3	1.30×10^4	6.04×10^3
Verification	Unweighted	Cost-Aware	CTVM
SSA	4.02×10^7	8.44×10^7	4.02×10^7
TIPTOP	6.25×10^6	1.86×10^9	2.80×10^7

从该表格中可以看出TipTop在覆盖问题上用的样本数量要比其它方法少很多，而且在验证时需要的样本在IM和CTVM问题上比SSA要少，在Cost-ware上要多一些。

3. Runtime and Solution Quality

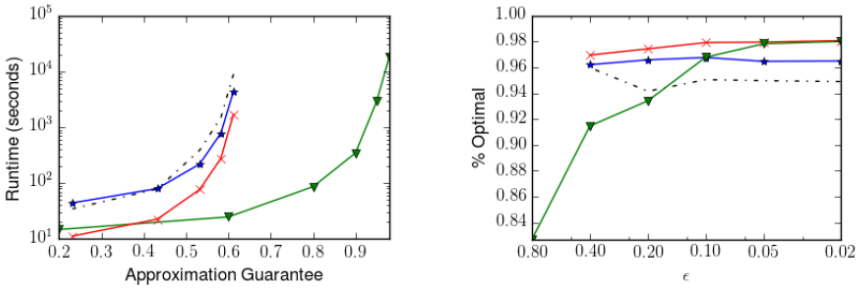


(a) Runtime as ϵ is increased. b is held constant at 50.



(b) Runtime as b is increased. ϵ is held constant at 0.02.

从图看出在CTVM问题上，TipTop运行时间性能上要比其它方法好很多。



(a) Running Time **(b)** Solution Quality

上图是在非常大的Twitter数据集上的性能测试，随着近似率保证的增加，其它方法比TipTop的运行时间要更早达到高峰，而且TipTop能够得到$(1-\epsilon)$的解。

其它解决方法

首先是IM问题被提出，并针对该问题设计了$(1-1/e-\epsilon)$的近似算法在IC和LT模型上求解。后来一系列方法被提出，聚焦于改进算法时间复杂度，比较有名的是偷懒估值法(CELF)。

之后有研究者提出新颖的采样方法，RIS算法。由于其样本生成会变的任意大，SSA算法被提出来解决这个问题。这些算法都有$(1-1/e-\epsilon)$的近似率保证。

随后CTVM被提出，并给出了$(1-1/\sqrt{e}-\epsilon)$近似率的算法BCT，BCT主要思想是最大权重覆盖和BSA采样方法。

创新之处

- 对现有方法的改进：提出了准确求解CTVM问题最优种子集合的方法：TipTop
- 致力于减少用于ILP寻找候选解的RR集合的数量。
- 使用了一种新的Verify的方式只需足够的样本就能评估候选解的质量。