# 基于节点删除的不确定源负面影响最大阻塞算法

## 文章信息

文章：Node deletion-based algorithm for blocking maximizing on negative influence from uncertain source

期刊：《Knowledge-Based Systems》

时间：2021.8

作者：Weijia Ju, Ling Chen, Bin Li, Yixin Chen2, Xiaobing Sun

作者信息：扬州大学信息工程学院；美国圣路易斯华盛顿大学计算机科学系

分区：



### 摘要

在社交网络中，流行病、谣言、虚假信息、计算机病毒等负面影响的传播可能会导致严重的后果。**负面影响阻碍最大化**的问题引起了研究者的浓厚兴趣。然而，在现实的社交网络环境中，负面影响的确切来源通常是未知的。在大多数情况下，我们只知道负种子的分布，也就是每个节点成为负种子的概率。在这项工作中，我们研究了最大限度地阻塞来自不确定来源的负面影响的问题。针对这一问题，我们提出了竞争影响线性阈值传播模型(CI-LTPM)。基于IC-LTPM模型，我们定义了不确定负源影响阻塞最大化问题(UNS-IBM)。我们使用live-edge (LE)子图中的传播树来估计影响传播。提出了一种基于LE子图传播树的正种子块增量计算算法。我们观察到正种子的阻塞效应是将正种子及其相关边从LE子图中删除后对负影响的减少。在此基础上，我们提出了一种基于节点删除的算法NDB (node- delete -blocking)来解决UNS-IBM问题。实验结果表明，与其他方法相比，NDB可以在较短的计算时间内屏蔽更多的负面影响。

本文主要贡献有：

1. 针对正面和负面影响同时传播的竞争性社交网络，我们定义了一个扩展的LT传播模型CI-LTPM。基于CI-LTPM模型，我们正式定义了不确定负源的影响阻塞最大化问题(UNS-IBM)。
2. 为了避免对影响传播过程的模拟耗时，我们在LE子图中提出了传播树来估计影响传播。
3. 提出了一种基于LE子图中传播树的负影响阻塞增量计算算法。
4. 我们提出了一种基于节点删除的算法NDB (node delete-blocking)来解决UNS-IBM问题。

## 2.2文章内容

符号网络主要代表了类似竞争网络的社交网络形式，其所面对的IM问题可以归结为：要寻一个种子节点集使得网络的积极影响最大化。

### 对IBM和UNS-IBM的影响传播进行建模

**IBM问题：**

由于社交网络的开放性和高速性，各种**虚假信息**和谣言可以在用户中迅速而广泛地传播。这种虚假信息往往伴随着网络热点或引人注目的信息，往往在网络中引起用户的广泛关注。谣言会降低企业的信誉，也会损害一个人的声誉。此外，假新闻的传播可能会导致社交网络的声誉和安全性的严重损失。因此，必须控制这些错误信息，使社交网络在信息交换中更加可靠和安全，即必须找到有效的方法来限制负面影响的破坏性影响。为了有效阻断负面影响在社交网络上的传播，常用的方式是选择能够传播真实新闻的有影响力的个体，将虚假新闻的影响降到最低。如果用户收到了真实的信息，并纠正了错误的信息，用户将不再受到有害信息的影响。因此，积极信息的传播可以阻断消极信息在社交网络中的传播这个问题的**关键是如何发现有说服力的个体，他们可以传播积极的信息，以最大限度地抵消消极信息的影响。**这是负面影响阻塞最大化的问题(IBM，influence blocking maximization)

**定义1. 消极影响阻塞量：**

用D、C分别表示消极影响种子集和积极影响种子集，表示网络中最终被D**消极影响**（激活）的节点数目；在C、D共同作用下，网络中被消极影响的节点数目。而积极种子集C的消极影响阻碍量表示为：

**定义2. IBM：**

研究已证明IBM的目标函数是单调的、具有子模性，且是NP-hard问题。

**UNS-IBM问题**

然而，在现实社会网络环境下的应用中，由于社交网络中的隐私规则，**很难获得负面影响的确切来源。**在很多情况下，我们只知道负种子的分布，也就是每个节点成为负种子的概率。这种负面影响源的分布用来描述负面种子的不确定性，可以通过经验知识或对用户行为的观察来估计。

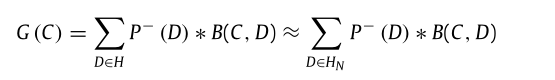
例如，在流行病的传播中，我们不知道流行病的原始来源，但我们可以从被感染者的信息中大致知道它的分布。与感染者最密切接触者成为传染源的可能性较大。再比如，在谣言传播过程中，我们不知道谣言的来源，但是我们可以根据谣言接收者的信息得到它的大致分布。与受众沟通频繁的有影响力的人成为谣言来源的可能性更高。在计算机病毒的传播过程中，我们可以利用被感染计算机的信息来估计病毒源的分布。与被感染计算机交换信息越多的网络节点成为病毒来源的可能性就越大。

在这种情况下，很难检测出能最大限度地阻断不确定源的负面影响的积极种子集。**将寻找一个群体能最大限度抵消不确定来源的消极信息传播的问题定义为UNS-IBM**（uncertain negative source influence blocking maximization）

在消极影响源不确定的情况下，为每个节点赋予表示节点称为消极影响种子节点的概率。函数表示消极种子节点在网络中的分布。节点集D被选为消极种子节点的概率为：



设G(C)为积极种子集C的期望阻塞效应，为了近似G(C)，我们构造了一个由N个消极种子集组成的子集。当N足够大时，有：

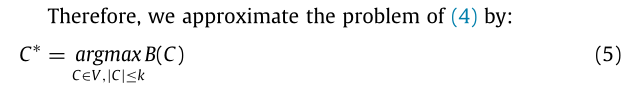


设中采样的不同种子集为， 在中的出现次数为。那么在中的出现频率为。用B(C)表示积极种子集C对阻断的预期负影响，有：



其中B (C, d)为正种子集C对种子d的负影响的阻塞值。定义函数：







为了克服UNS-IBM问题中未知负种子的困难，文章采用(5)来估计阻塞对每个可能消极种子的影响。如果对于所有候选的积极种子集C，我们知道它们的阻碍量B(C,d)，我们可以选择B(C,d)值的最大加权总和的集合C作为积极种子集。因此，关键问题是如何估计候选积极种子集C和可能的消极种子集d的阻碍量B(C,d)。我们提出了一种在扩展LT模型下基于LE子图的抽样方法来计算阻碍量B(C,d)。

### LT模型

1. 假设网络中不同节点有不同的激活阈值，可以更准确地反映节点在社会网络中传播的影响下的状态。
2. 一个节点是否被激活并不取决于一个邻居的影响，而是取决于它所有邻居的综合影响。
3. 在LT模式下，影响传播可以通过“live-edge”(LE)子图来模拟，文章可以在LE子图的传播树上执行节点删除来估计阻塞增量。

传统的LT传播模型仅适用于网络中只有正向影响的情况。为了解决负面影响阻塞问题，我们将LT传播模型扩展到包含正面和负面影响的竞争网络中。我们定义竞争影响线性阈值传播模型(CI-LTPM)如下。

**CI-LTPM(Competitive Influence Linear Threshold Propagation Model):**

设G = (V, E, P)是一个积极和消极影响同时传播的竞争网络。G中的每条边(u, v)都有一个传播概率P(u, v)。在区间[0,1]中，每个节点v被分配一个激活阈值。设C和D分别为消极种子集和积极种子集。每个节点处于三种状态之一：未受影响、受积极影响、受消极影响。其传播规则为：

1. 开始时，所有积极、消极种子节点分别处于积极、消极影响状态，其他节点均处于未影响状态。
2. 指向节点邻居的积极影响超过的阈值，则节点受积极影响激活
3. 指向节点邻居的消极影响超过的阈值，则节点受消极影响激活
4. 如果积极、消极影响都超过了的阈值，根据社会心理学中的“负面影响支配”原则，节点还是主要收到消极影响
5. 当被积极、消极影响激活后，不能再受到相反的影响。
6. 无新的激活节点，传播结束。

### 构建live-edge图

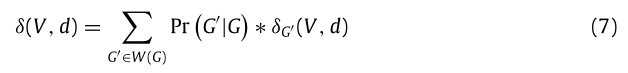
在LT模型下，live-edge (LE)子图可以用来模拟影响传播。构造LE子图时，根据传播概率P(u,v)，选择每个节点v的一条入边(u,v)加入LE子图。在网络G = (V, E)中，假设选取的活动边集合为，则LE子图为。

在CI-LTPM模型下，我们通过生成多个LE子图来模拟负面影响的传播。通过从每个LE子图中删除一些节点，我们可以计算被删除节点的负面影响阻塞效应。从而，选取阻塞效应最大的节点集作为积极种子集。

为了解决UNS-IBM问题，关键问题是如何估计Eq.(5)中B(C,d)的值，即积极种子集C对消极种子集d的消极影响的阻塞效应。根据Eq.(1)，有:

其中，是消极种子集d在积极种子阻断作用下传播的消极影响（的节点数目）。**在CI-LTPM模型下，如果不存在积极的种子集，等于所有LE子图中d可达节点数的期望。**

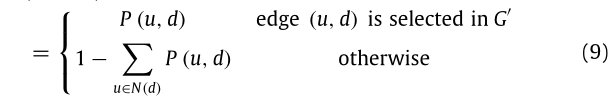
设W (G)是G的所有LE子图的集合，是G的LE子图，是G '中d可达的节点数，则有:

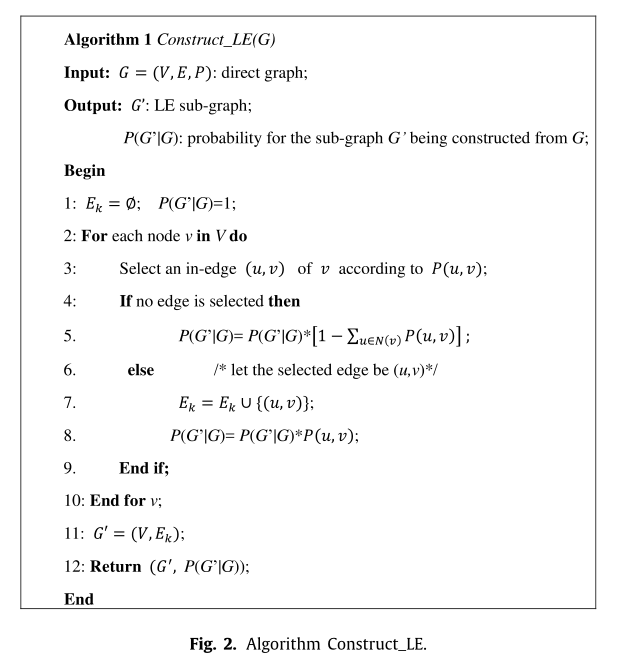


式中表示G '在G中被构造的概率，其计算公式如下:



这里，是节点d与G '中至少一条边连接的概率。设N(d)是d的内邻居的集合，P(u, d)是边(u,d)上的传播概率。那么可以用下式计算：





### 基于传播树计算阻碍量

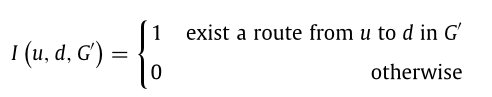
UNS-IBM来解决这个问题,我们必须估计B(C, d)。我们知道B(C, d)的值可以用进行估计。

我们注意到，如果我们从网络中删除一个节点c和它的相关链接，负面影响力就不能再通过节点c传播了。设b(c)为删除节点c后负面影响力的减少量。假设c是一个正种子，那么c的阻断效果为b(c)。因此，每个积极种子节点c所阻断的负向影响等于删除c及其相关链接后的负向影响的减少。**基于这样的观察，我们提出了一种基于节点删除的算法来阻断不确定来源的消极影响。**我们通过生成多个LE子图来模拟消极影响的传播。**在每个LE子图中，我们定义了以每个负向种子d为根的传播树，并通过从传播树中删除每个候选正向种子c来估计其影响阻断值的增量。**最后，可以根据每个节点的阻断效果增量来选择正向种子集。

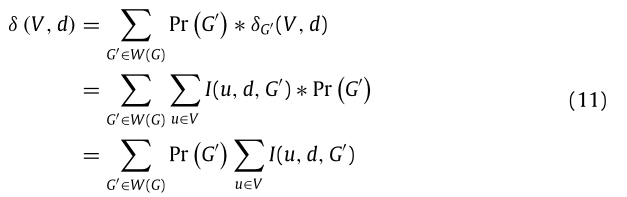
**1）传播树及其覆盖范围**

通过(6)来计算阻塞效应B(C,d)，必须估计和的值。通过(7)我们知道可以基于来计算，可以在LE子图G '中通过以下公式计算：





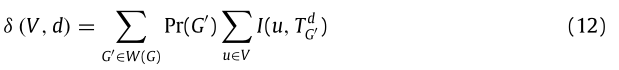
用W (G)表示图G中所有LE子图的集合，通过(7)与(10)结合，我们可以得到：

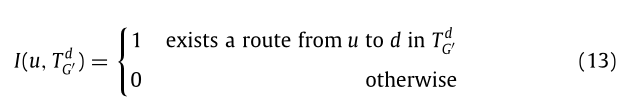


为了估计(11)中对于每个可能的负种子d的，我们在LE子图中定义了以d为根的传播树。

**定义5. 传播树：**

LE子图G '中节点d的传播树，记为，定义为G'中以d为根的广度优先搜索树。由于LE子图中的每个节点只有一条入边，因此这种以d为根的传播树在G '中唯一存在。基于以负种子d为根的传播树，我们可以估计传播函数为：





**定义6. 中节点u的覆盖范围：**

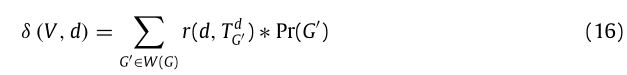
中节点u的覆盖范围记为，是中以u为根的子树的节点数。



如果我们把u用d代入(14)，我们可以看到r (d, )就是中的节点数：



结合(12)(15)式可得：

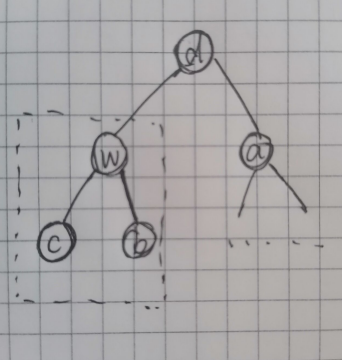


**2）估计阻碍增量**

假设一个候选的积极种子节点w被添加到部分积极种子集C中，为了计算新种子w的阻断值增量，我们从G中删除它和它的相关边，得到一个子图，这个子图被表示为。然后，我们就可以估计中的消极影响扩散。

对于G的每个LE子图G′，存在一个唯一的的LE子图，对应于G'。设d是一个消极节点，是G'中以v为根的树，也有一个以v为根的传播树，对应于G'中的。那么在LE子图中，由d传播的负面影响是 。和之间的差异是w对d在G'中传播的消极影响的阻挡作用。（增量）与传播树相比较，我们可以看到中只有节点w和以w为根的子树是缺少的。因此，很明显：

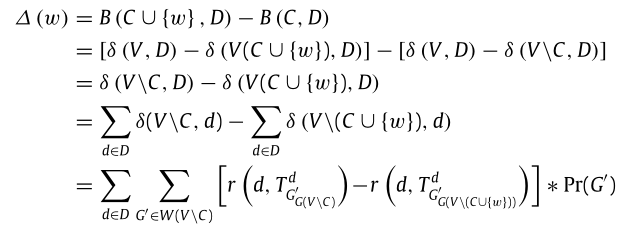




假设C是一个部分正集，d是一个消极的种子，在C中加入一个候选的积极种子w后，对d的负影响的阻断增量为：



给定正、负种子集C和D，以及：一个候选的正种子w∈V\D\C，将节点w加入C后，阻断增量值∆（w）为：



同样，在不确定的负种子集的情况下，对于一个正种子节点集C，在向C中加入一个候选的正种子w后，可以通过以下方式估计阻断增量∆(w)：



从上式可以看出，候选积极种子w的阻塞增量∆（w）可以通过从G构建LE子图G′G(V\C)来计算。阻塞增量∆（w）是w在所有传播树T d G′G(V\C)中的覆盖率的加权和，用于d∈V\C。具有最高阻塞增量∆(w)的节点将被选为新的积极种子。

**3）构建传播树**

为了计算每个节点w的阻塞增量∆(w)，我们提出了一个名为Construct\_PT（构建传播树）的算法，以构建LE子图G'中每个节点v的传播树。基于传播树，可以计算出的值和中w的阻塞增量∆（w）。该算法以每个节点v为根，用广度优先搜索法生成传播树。传播树是通过记录树中每个节点的父和子来表示的。在构建了传播树之后，我们计算中每个节点u的覆盖率，也就是u在中的子代数。∆(u)在G'中的部分值是。在O(m)时间内，使用广度优先搜索方法来构建传播树。因此，Construct\_PT算法的时间复杂度为O(m,n)。

**4）估计LE子图的数量**

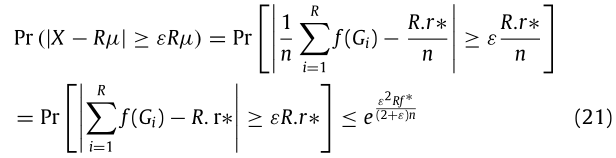
从（20）中可以看出，我们需要生成W（G）中的所有LE子图，以计算每个候选正种子w的阻塞增量∆（w）。然而，在所有LE子图中构建传播树需要大量的时间。为了克服这一障碍，我们提出了一种抽样方法，随机选择部分LE子图来构建W（G）的子集。用H(G)表示从W(G)中随机选择的LE子图的集合。我们可以通过用H(G)代替(20)中的W(G)来近似计算∆(w)的值。显然，我们建立的LE子图越多，我们可以得到更准确的阻塞增量，但需要更多的计算时间。为了平衡计算时间和结果的准确性，我们预先定义了一个误差阈值和一个概率。为了将误差限制在概率下的内，我们使用Chernoff bound来估计H(G)中采样的LE子图的数量。

**理论1. Chernoff Bound**

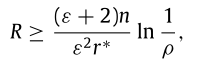
给定R个独立且相同分布的随机变量在[0,1]中，平均值为μ，让X是的总和。对于一个给定的误差阈值ϵ>0，则有：



根据定理1，我们可以计算出在给定的误差阈值下，被抽样的LE子图的数量：



由此知道LE子图的采样数量满足：



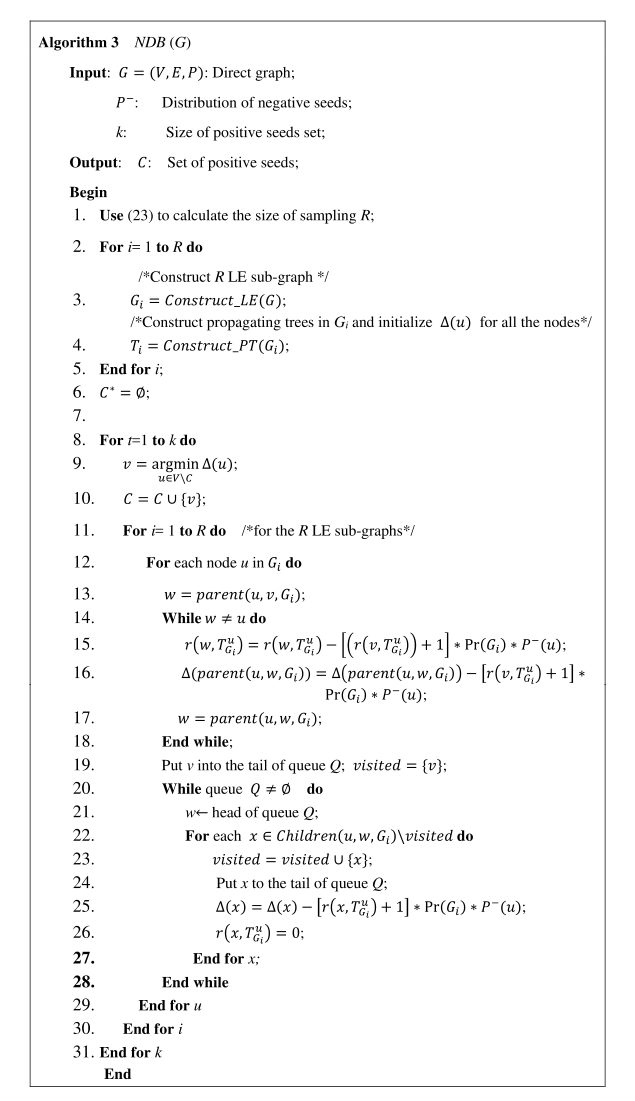
在概率为的情况下，误差将小于。

**定理2.**

设负数种子集的最大大小为。给定一个概率和一个误差阈值，如果采样的LE子图的数量满足：



则在概率为的情况下，误差将小于。

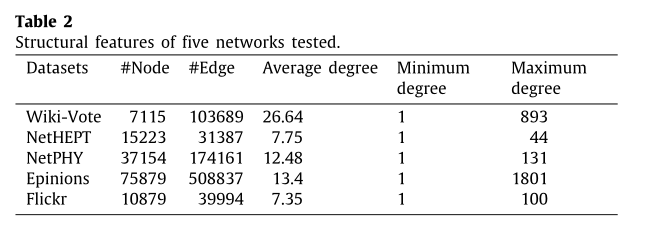


## 实验解析

为了评估算法NDB的有效性，我们在五个真实的网络中进行了实验。在实验中，我们测试了我们的算法NDB的阻塞效果，并与其他影响阻塞最大化算法进行了比较。

**Datasets：**

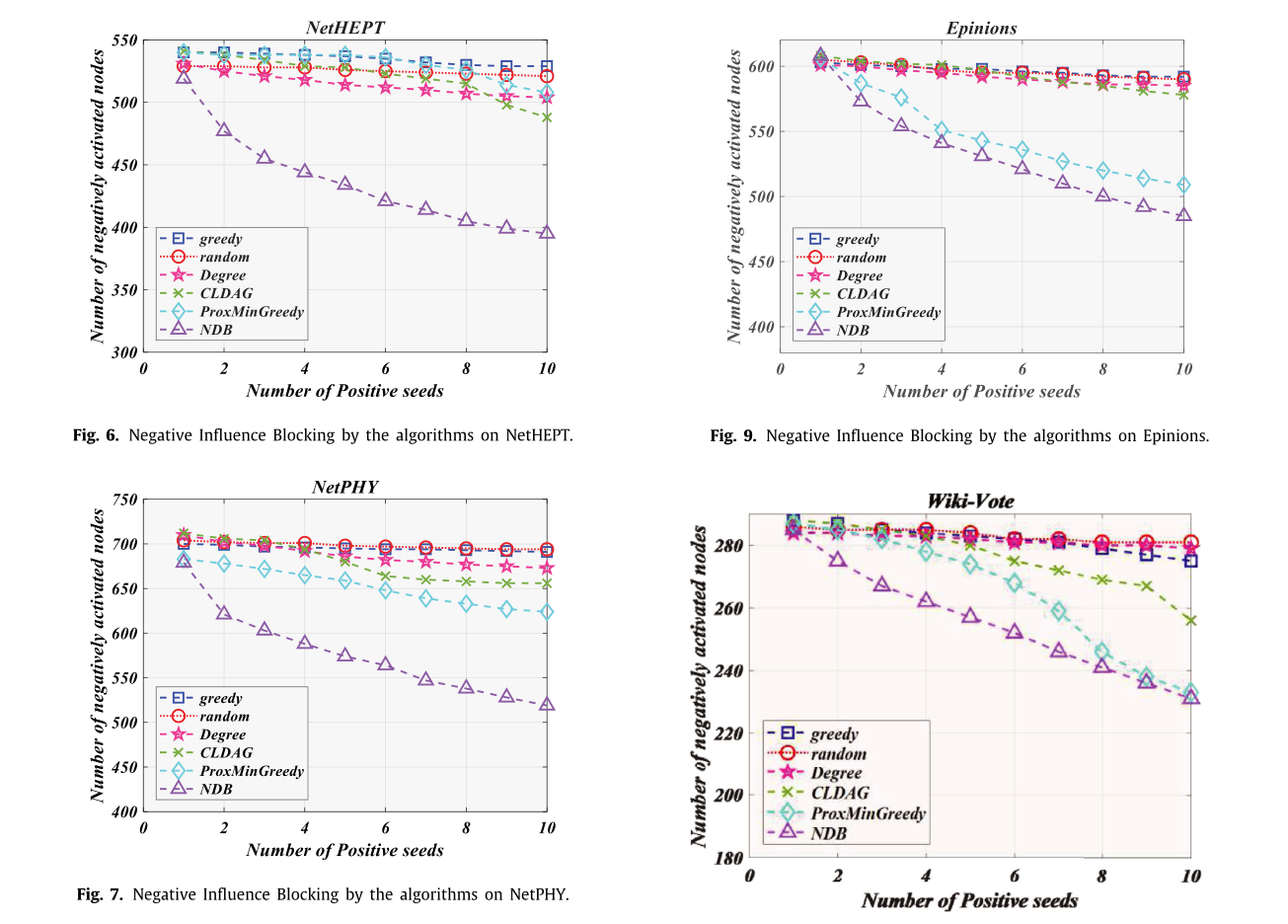
|  |  |
| --- | --- |
| Epinions | 一个消费者可以讨论商品的网站。他们也可以对其他用户的评论给出自己的意见。根据他们的观点，可以定义他们之间的积极和消极关系。 |
| Wiki-Vote | 它在维基百科上记录了投票历史。网络中的节点代表维基百科的用户，从用户u到v的签名链接表示用户投票赞成或反对向管理部门推荐v。 |
| Flickr | 这个网络是通过在Flickr中使用公共元数据的图像之间添加边来构建的。在网络中，每个节点代表一个图像。如果两个节点的图像来自相同的位置、图库、组或集，则会有一条边连接它们。 |
| NetHEPT and NetPHY | 这两个数据集代表了从arXiv网站获得的物理领域的学术研究网络。NetHEPT数据集由高能物理理论部分的作者和文章组成，而NetPHY数据集由物理部分的作者和文章组成 |

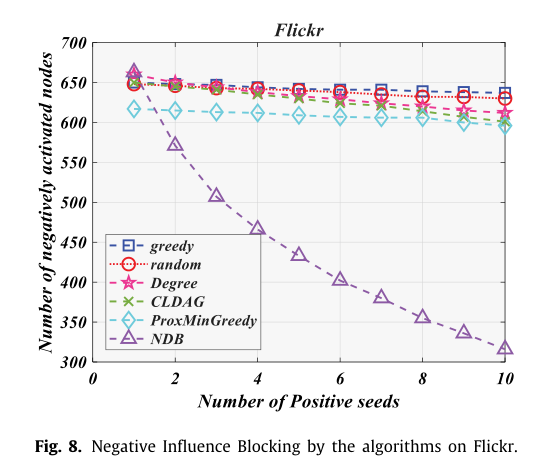


**Algorithms：**

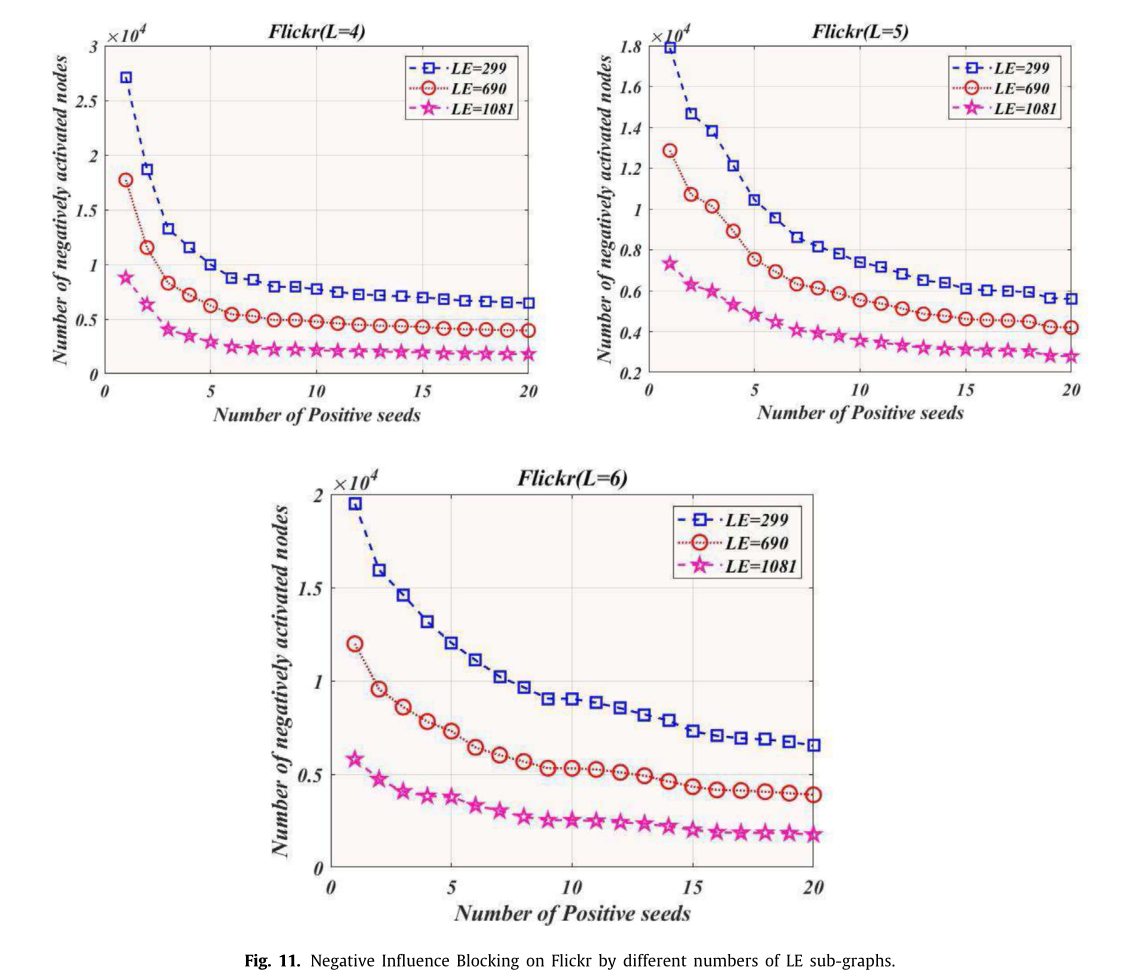
|  |  |
| --- | --- |
| Random | 这是一种随机方法，随机选择节点作为正种子。在测试中，随机种子选择进行了1000次，1000个种子集平均阻塞为最终结果。 |
| Degree | 该方法将中心度最高的节点作为影响最大的节点，采用启发式方法选择影响最大的节点作为正种子。 |
| Greedy | 该算法将种子集初始化为空集，并反复选择一个潜在的候选节点加入种子节点。在每一轮种子选择中，该方法选择负影响阻塞增量最大的节点作为候选种子。通过蒙特卡洛模拟估计节点的负影响阻塞增量。 |
| CLDAG | 该算法是在一种扩展的竞争性线性阈值模型CLT (competitive linear threshold)下设计的。它利用CLT模型的性质，基于LDAG(局部有向无环图)结构，通过计算正、负激活概率来选择正种子。 |
| ProxMinGreedy | 这是在2020年提出的一种针对社交网络中负面影响遏制的方法。该方法采用扩展的单方向状态转移线性阈值模型(LT1DT)来描述网络中竞争信息的传播。采用贪心方法来最小化虚假信息在社交网络中的传播。 |

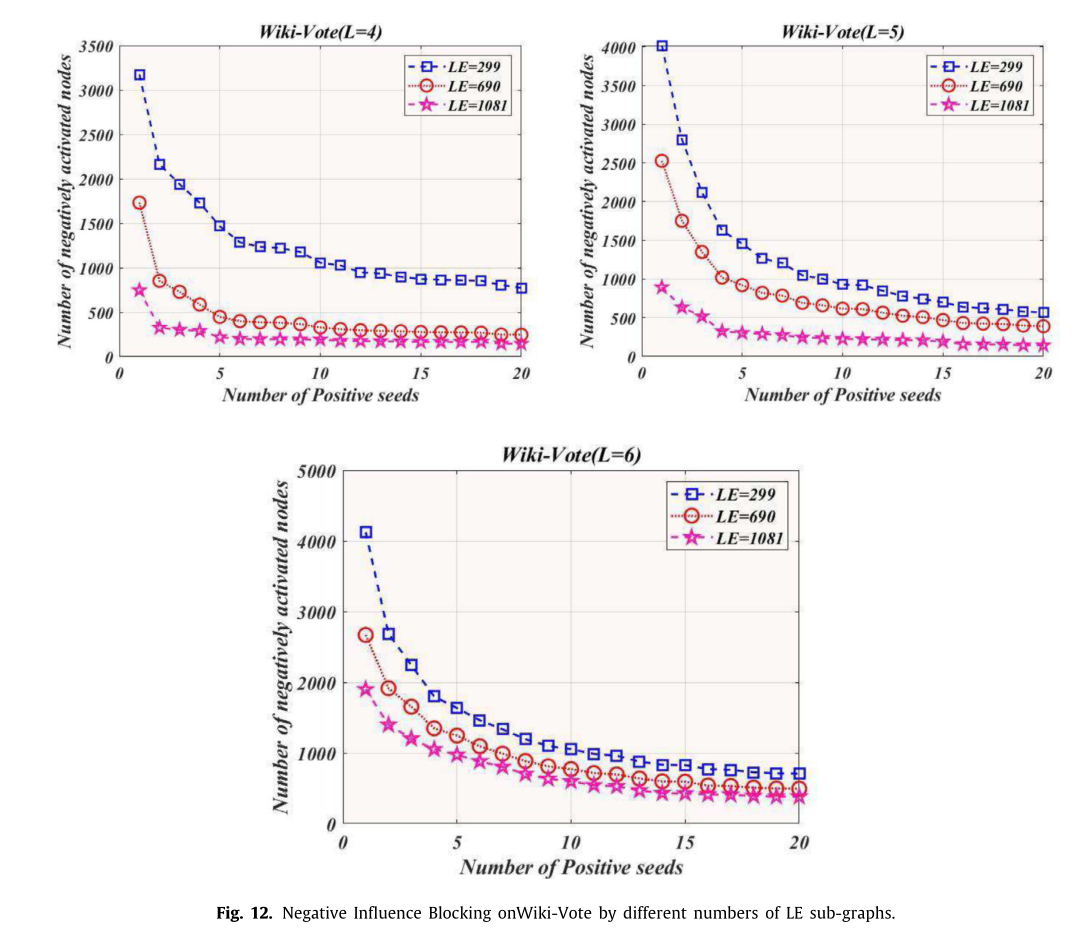
### 负影响阻碍量比较

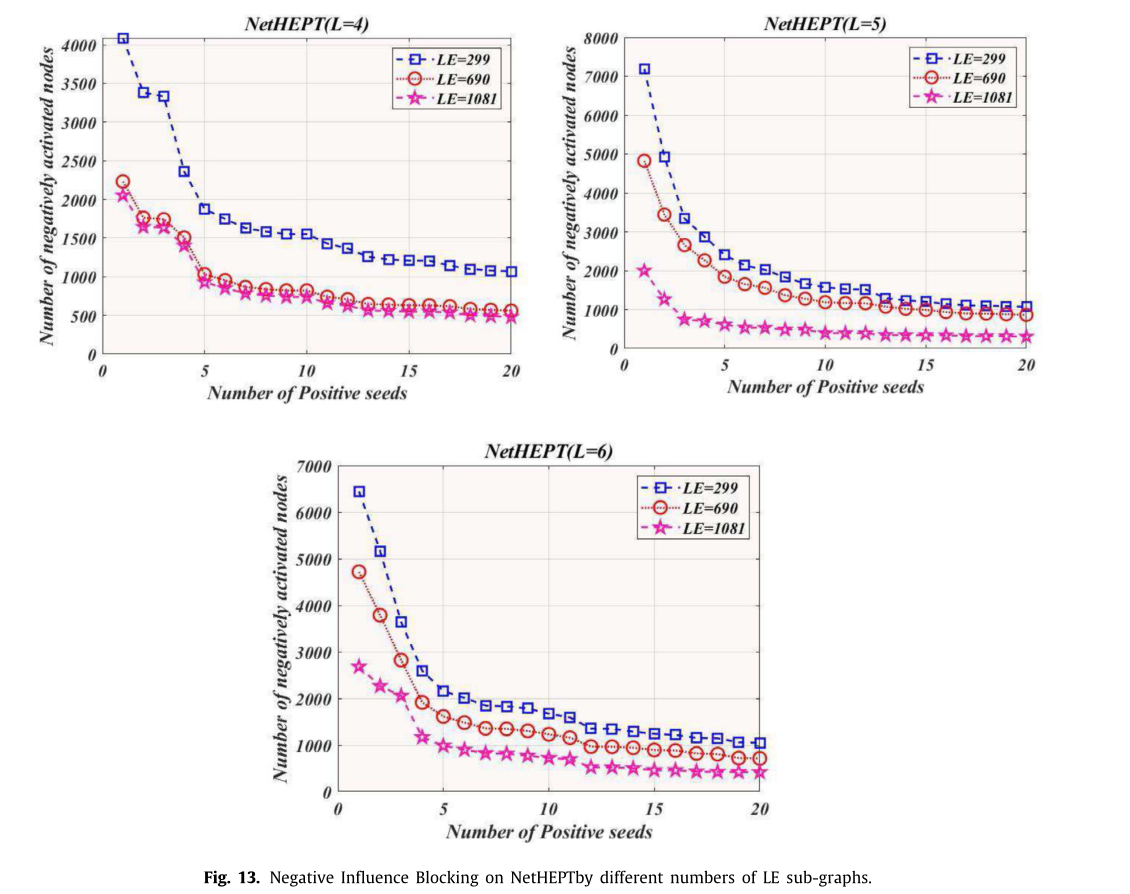




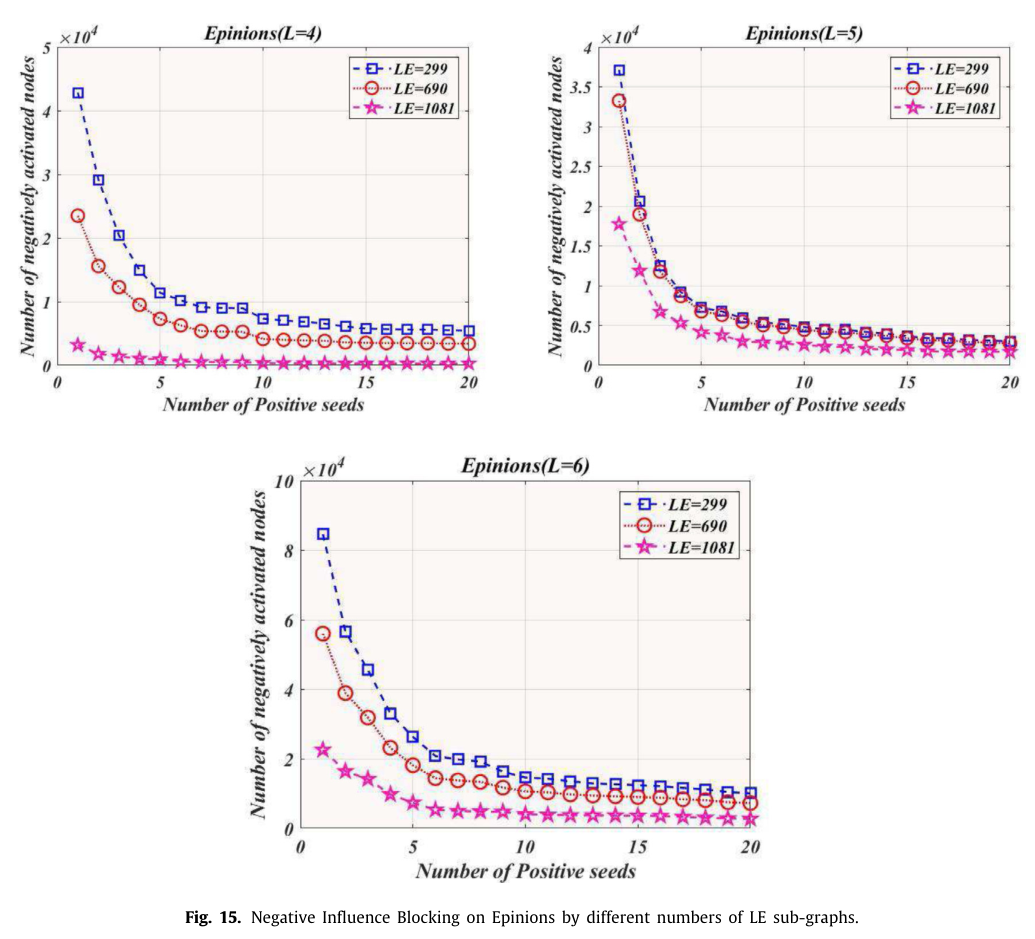
### 通过不同的采样大小来影响阻塞





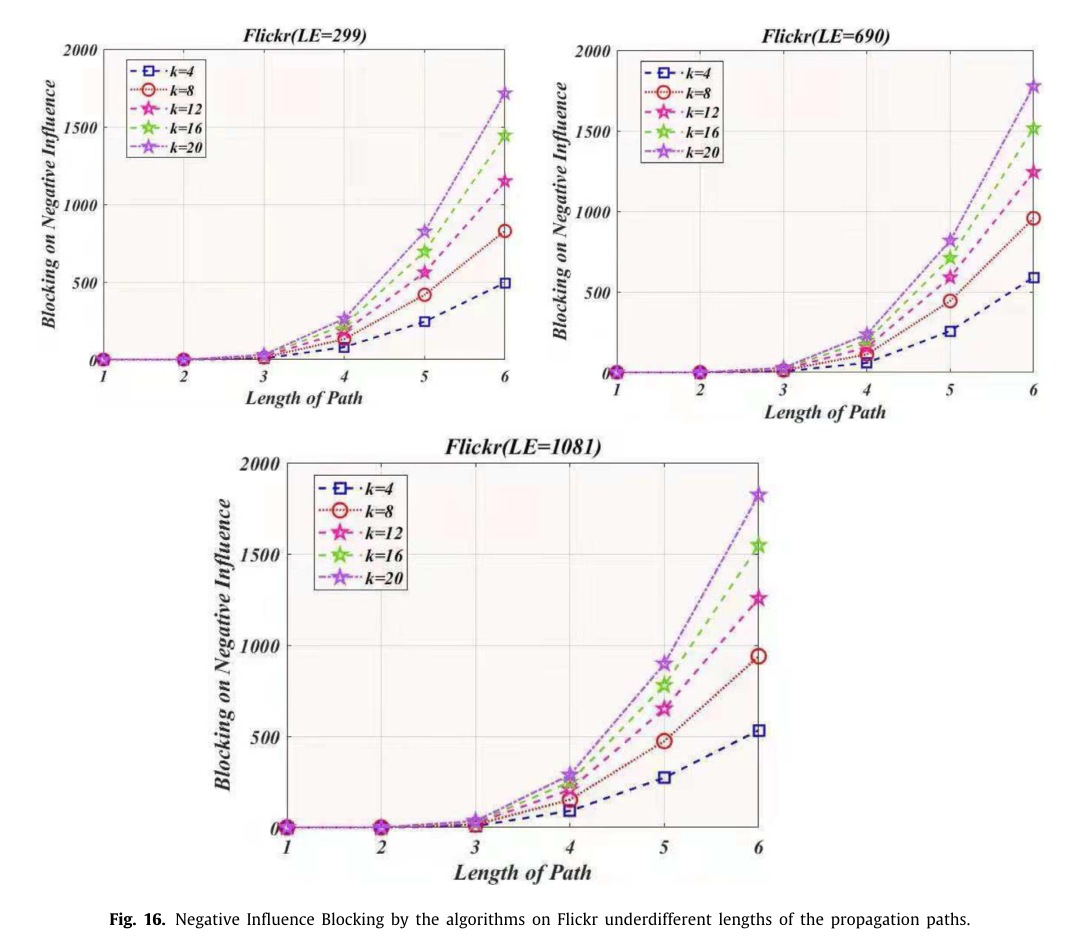
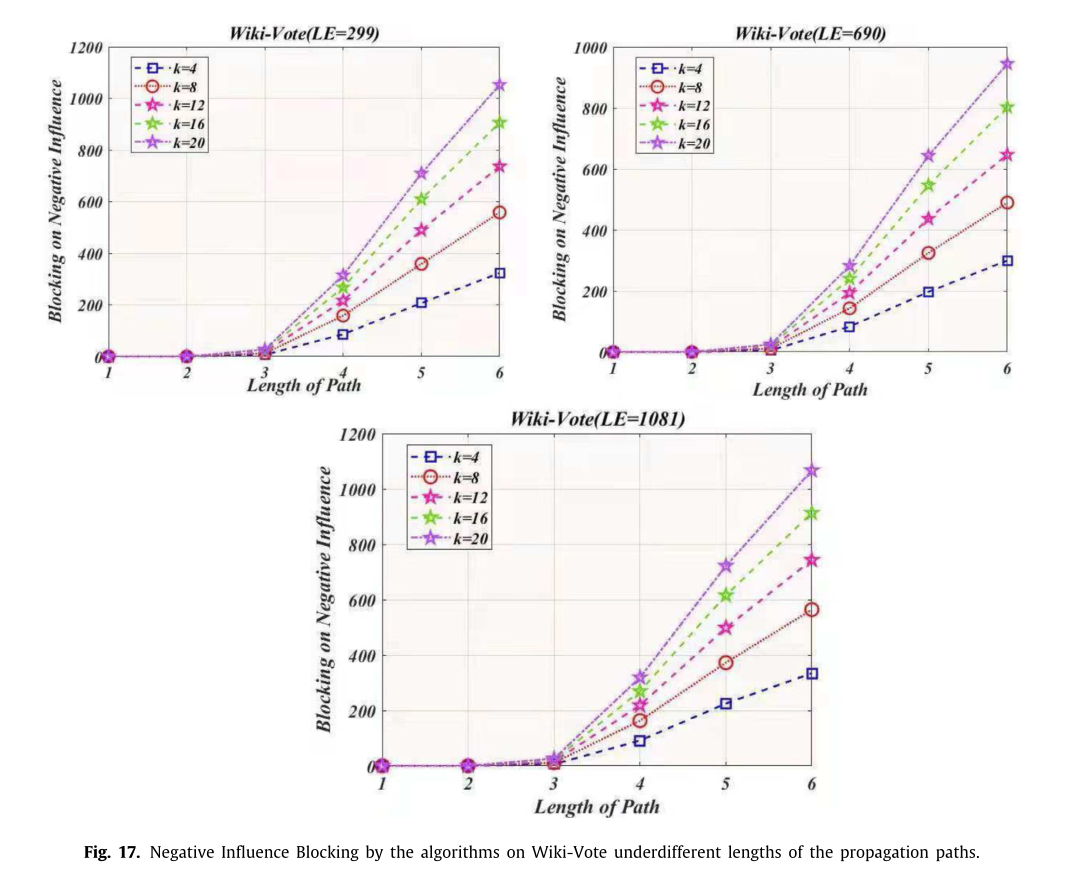


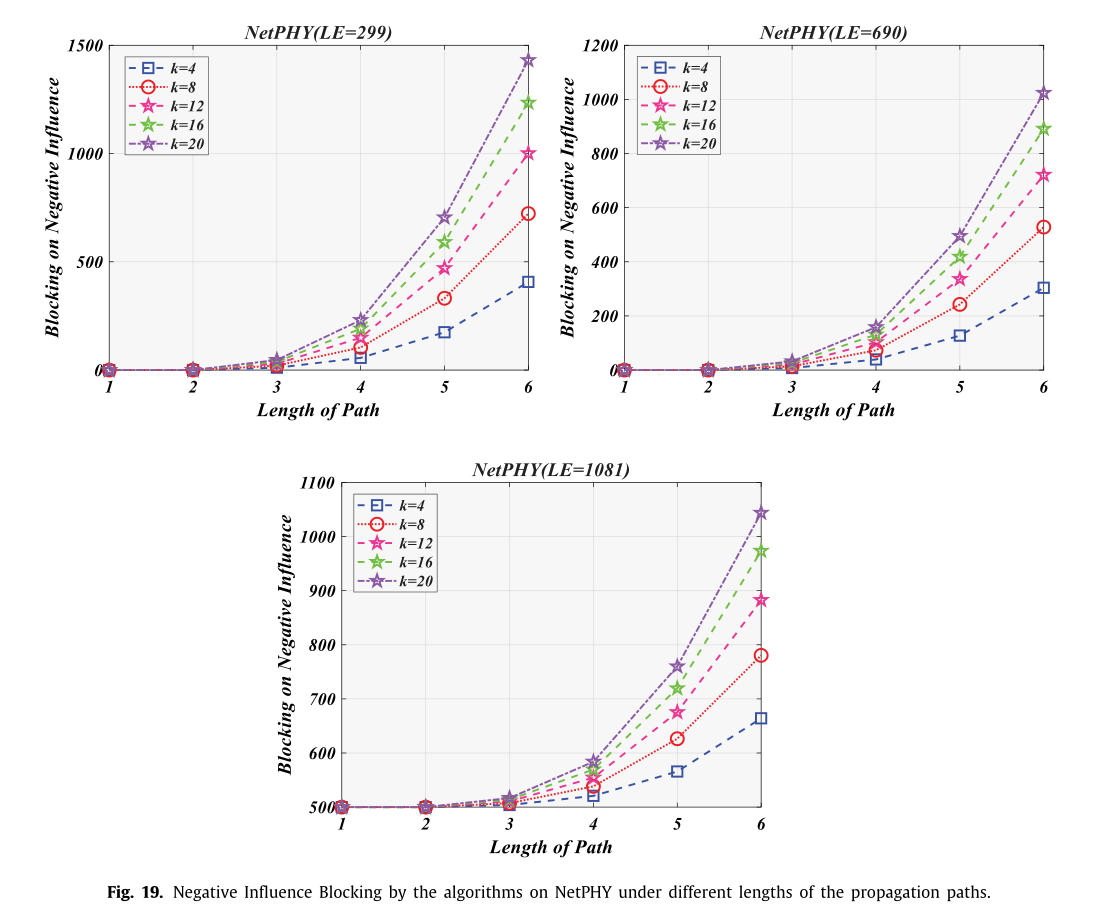
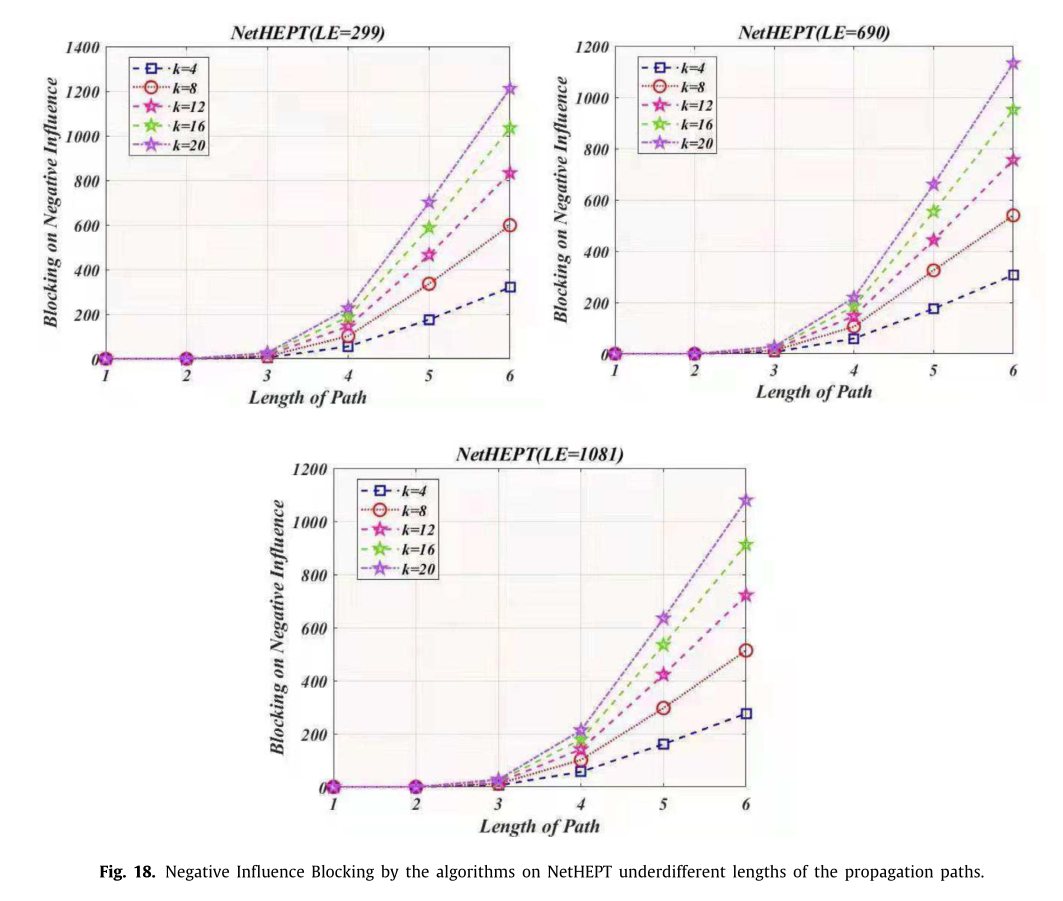


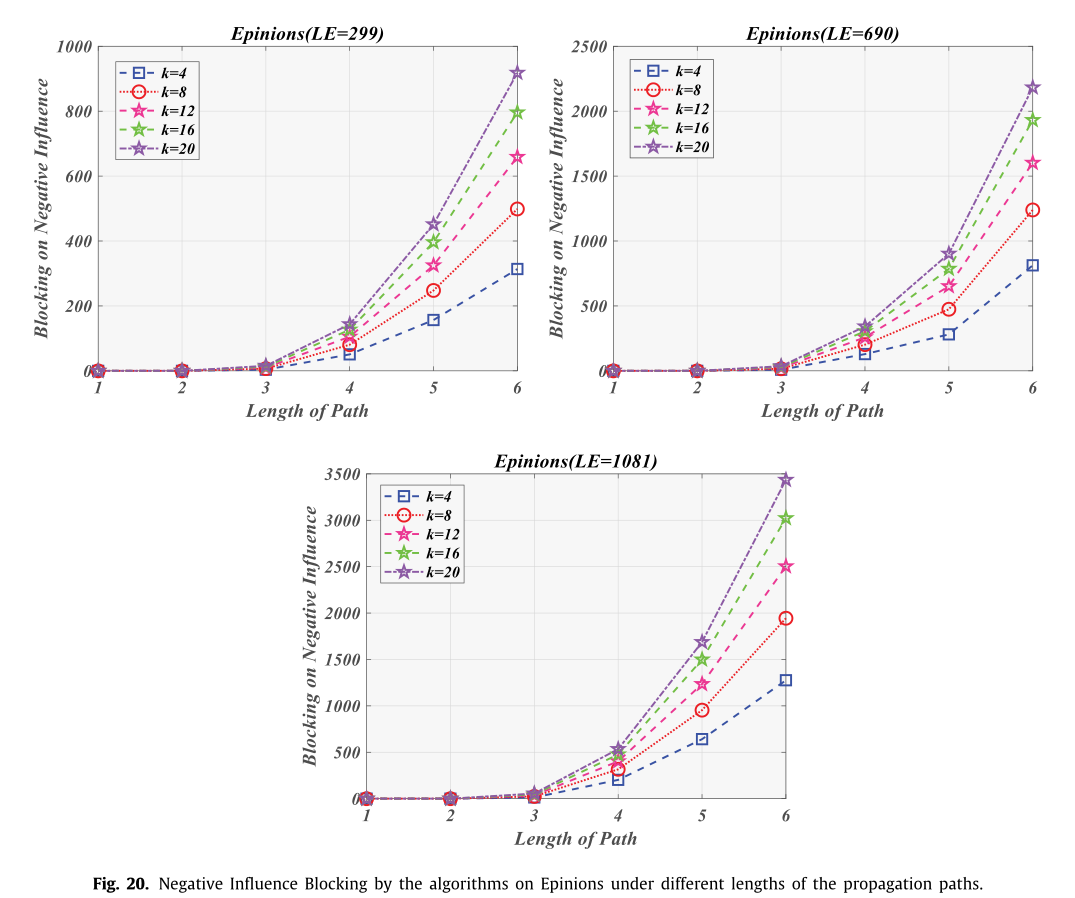


### 通过不同长度的传播路径影响阻塞

我们还在五个数据集上用不同传播路径长度L的NDB算法对影响块进行了测试和比较。在实验中，我们通过设置误差阈值ε = 0.1、0.08和0.05，分别得到了299,690和1081个采样的LE子图。阳性种子数设k = 4、8、12、16、20。图16 ~图20显示了不同传播路径长度和不同正种子数量下对数据集的影响块。这种影响阻塞就是负性激活节点的减少。从图中可以看出，传播路径越长，所覆盖的传播范围越广，阻塞效应越大





### 时间复杂度

