**Topology-Driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks**

**Abstract**

关于影响力最大化的研究往往必须应对与向特定用户传播信息相关的营销需求。然而，很少有人注意到这样一个事实，即信息传播活动的成功不仅取决于要检测的最初影响者的数量，还取决于他们的多样性，即活动的目标。我们的主要假设是，如果我们学习的种子不仅能够影响用户，而且与更多样化的(群体)用户相关联，那么影响触发因素也会多样化，因此目标用户将获得更高的参与机会。我们定义了一个新问题，叫做对多样性敏感的目标的影响最大化(DTIM)

**Introduction**

在线社交网络（OSN）现在很火。一个核心问题是在OSN中识别有影响力的个人，这样，从他们开始，就可以触发由“口碑”驱动的连锁影响反应，这就允许在初始投资(预算)方面相对较少并且能影响到网络中的很大一部分人。这通常被称为病毒营销原则，这是osn中经典优化问题的根本动机，即影响最大化。

在市场营销中一个组织经常希望将其产品的广告范围缩小到有特定需求或偏好的用户，而不是针对整个人群。此外，在OSN的场景中，一些事件只会引起具有特定品味的用户的兴趣。我们的工作就是对这个问题进行研究，以下称为目标影响力最大化。

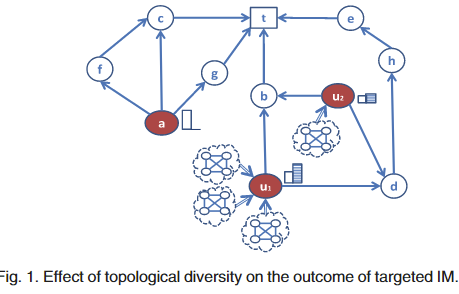
利用多样性去加强影响力最大化。在最大化产品广告的同时，组织也需要最小化目标用户，这显然增加了选择适当数量的种子用户的必要性。令人惊讶的是，一个经常被忽视的重要方面是病毒营销过程的成功不仅取决于种子集的大小，还取决于反映在种子集内或与其相关的多样性。在种类(例如，年龄、性别)、社会文化方面(例如，国籍、种族)或其他特征方面彼此不同的个人，会带来独特的观点、经验和视角，此外，在OSN的背景下，成员自然有不同的知识、社区经验、参与动机和共享信息。值得注意的是，多样性已被普遍认为是数据分析中一个关键的因素，对于提高生产率、提高用户对基于新奇和意外发现的内容推荐的满意度至关重要

将这张图片带入目标IM场景，让我们关注用户参与的问题。没有积极参与社区生活的用户通常会因不同的背景和动机而欢呼雀跃，并就不同的主题进行交流，这使得他们很难参与进来。一个有效的用户参与策略应该考虑到社区中年长的活跃成员的支持，因此，通过识别最多样化、最活跃的成员，触发刺激也将是多样化的，由于不同的个人倾向于联系许多不同类型的成员，有效参与的可能性会更高。

目标IM中的多样性挑战。现有的目标IM方法中不考虑多样性的问题。在这项工作中，我们的目标是克服这一限制，使用无监督的方法。也就是说，我们的研究依赖于一个视角，该视角不假设任何关于用户属性的附带信息或先验知识(例如，个人简介、热门话题偏好、社区角色)，从而能够实现用户之间的多样化。我们假设用户在社交图中的多样性可以基于与其邻居相关的拓扑性质来确定。

我们的主要假设是，如果我们学习的种子不仅能够影响用户，而且与更多样化的(群体)用户相关联，那么我们会期望影响触发因素也会多样化，因此目标用户将获得更高的参与机会

例子1：为了支持上述假设，考虑图1中所示的示例社交图，其中节点表示个人，边表示影响关系。



该图与扩散过程相关，为了简单期间我们省略了影响概率和节点的活跃属性。在图中，节点t为目标节点，彩色节点a、u1、u2为候选种子，对于这些种子，我们知道个体对t的相关传播影响以及根据某种多样性函数的个体拓扑多样性；在图中，这些分数分别由与每个候选种子相关联的最左边的条和最右边的条显示。传统的目标IM方法会将节点a添加到种子集中，因为它在候选种子中具有最高的传播能力；然而，a的位置有两个特征，正如我们将在后面解释的那样，将意味着较差的拓扑多样性:1它没有入边，2并且它会扩散到具有t作为宿的相同子图中的所有节点。相比之下，节点u的位置在拓扑多样性方面是很好的，因为它们可能受到一组或多组节点的影响，从而有可能获得更广泛的各种信息和观点。因此，选择节点u将可能更有效地增加节点t的参与度。

这里出现了两个主要的研究问题，即如何利用用户的社会多样性来提高目标IM任务的性能: R1当我们对用户属性没有先验知识时，如何确定多样性；如何通过将多样性作为目标集来确定种子用户。

贡献：我们提出来DTIM这个概念。更具体地说，为了回答R1问题，我们提供了两种为目标IM建模拓扑驱动多样性的替代方法，这取决于采用何种方法来利用特定于给定目标节点的扩散子图中的结构信息。第一种方法被称为local diversity，第二种方法叫做global diversity。为了回答R2，我们利用本地多样性和全球多样性的定义来开发DTIM问题的替代算法，被称为L-DTIM和G-DTIM。

**2. 相关工作**

**信息传播的多样性。**大多数现存的多样性概念都是围绕网络的结构特征发展起来的，或者基于用户简档属性。[25]的作者提出了一种可控制性的度量，定义为能够在整个网络中传播观点的节点数量。在【29】中介绍了节点多样性的IM问题。

**带目标的影响力最大化**。在这个问题中，一些研究假设目标是唯一的，并且是先前指定的。在[30]中，作者解决了在IC模型下为特定目标用户找到前k名最有影响力的节点的问题。在[31]，作者研究了影响目标用户的最佳传播策略。在[32]，作者考虑了接受概率最大化的问题。与这些研究不用，我们的DTIM方法的目标是通过一个种子集合将激活目标集合的概率最大化。其中，目标集合可以很大，不是单一的节点；种子集合不是固定的也不是单一的也没有限制。

在《[21》中，作者描述了一种基于关键词的目标IM方法，名为KB-TIM。这假设每个用户都与加权术语向量相关联，以捕捉她/他对广告的偏好。关键字与广告相同的用户将属于目标集。

**3.基于拓扑驱动的多样性的目标影响力最大化。**

3.1 问题声明。

是一个带权的有向图代表信息传播模型。V是节点结合，e是边的集合，b是变权重的函数，l是节点权重的函数。边缘加权函数b对应于线性阈值模型的参数，在本工作中我们采用该模型作为信息扩散模型。在LT模型下，如果每个节点的总影响权重超过与该节点相关联的阈值，则该节点可以被其活动邻居“激活”。对于一个边（u，v）它的权重b（u，v）可以理解为节点u对节点v的影响大小。并且规定节点v的入边上的权重之和小于等于1。在最开始的时候，每个节点都会被随机分配一个阈值，这个阈值是0-1之间的。给定一个集合S作为初始活跃节点的集合，如果一个非活跃节点v的活跃邻居的入边权重大于它的阈值。我们使用来代表最终的活跃集合。

给定，节点权重函数l将每个节点的状态确定为目标，即信息扩散过程指向的节点。对于一个用户指定的阈值L，我们将G中目标集合TS定义为：

定义1：Capital。给一个集合S，与最终活跃集合相关联的capital定义为：



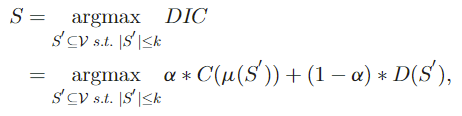
它与种子集激活的目标节点的累积状态成正比

在这里面我们不考虑属于种子集合S的节点

定义2 Diversity：给定一个集合S，它的多样性D（S）与目标集合的关联可以定义为：



定义3：多样性敏感的目标影响力最大化：给定一个扩散图G，一个预算k，一个阈值L，找到一个种子集合S，|S|<=k，要使得对多样性敏感的Capital最大（DIC）

其中a是一个参数。可以看到这个方程其实就是上述两个定义的一个线性组合。多样性敏感的目标影响力最大化保留了IM问题对策复杂性，他也是NP-hard。

论点1：capital函数在LT模型下是有单调性和子模性的

论点2：diversity函数也是具有单调性和子模性的

上述两个论点的证明是另一篇文献中提到的。

**3.2拓扑驱动的多样性**

我们对用户多样性建模的观点是只利用由社交网络图的拓扑结构给出的结构信息。因此，我们利用完全无监督的过程来避免需要任何关于用户属性的辅助信息或先验知识，这可以保证用户的多样化。们从社会科学中获得灵感，因为在OSN，用户与他人的联系方式(又称社会嵌入性)被认为是个人在网络社会环境中多样性的表现。这也与boundary spanning理论相关，该理论基本上指出，OSN用户可以从他们的一些社交联系人那里获得知识，然后通过社交图的一个或多个组件(例如，局部诱发的社交)将其传播(部分)到其他联系人。

定理1：社交图中用户的多样性可以基于她的/他的邻居的拓扑属性来确定。

定义4：特定目标的信息传播子图：给一个传播图G=<V,E,b,l>，社交图G=<V,E>，一个目标节点t属于TS，和一个时间步骤tao，我们将特定目标的信息传播子图定义为一个有向无环图以t为根，对应于G在时间tao向t扩散的展开部分。

定义5：边界集合：给定一个特定目标的信息传播子图，它的边界结合定义为，在子图中的节点至少有一个不在子图的节点指向它，这些节点的集合成为边界集合。



定义6：目标特定扩散子图的扩展：给一个子图，一个时间tao，在tao+1时刻包含了那些可以在tao时刻到达上一个子图的边界集合的节点。

我们为目标IM提供了两种可选的拓扑驱动多样性建模方法，这取决于构造Gt所采用的策略:

1.对于给定的目标t，在信息传播子图中的每一步计算它的节点多样性。由于该方法不需要t的完全扩展的扩散子图的信息，所以它被称为local diversity。

2.第二种是global diversity，在完全扩展的子图中计算节点多样性。

**3.2.1 局部多样性**

我们关于节点局部多样性的概念被设计成考虑给定目标节点的信息扩散图的渐进扩展。

给定一个Gt和Bt中的一个节点v，我们要计算Nin（v）中的每个节点u的局部多样性，遵循以下两个准则。

定理2：节点u的多样性应该与从当前展开的特定于目标的扩散子图Gt之外的节点到达它的可能性成比例，即与G中但不在Gt中的u的入邻居的数量成比例。

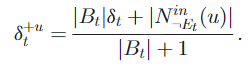
定理3：节点u的多样性应该与该节点对Gt中尚未包括的传入链路数量的增量成比例。

定义7：集合的边界多样性：给定一个Gt，它的边界多样性定义为边界集合中结点的入度的

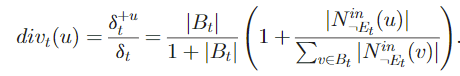
平均值



对于节点u是v的入邻居，如果u加入了Gt，边界多样性就会发生变化



定义8：局部多样性：u的局部多样性定义为Gt中包含u的边界多样性与实际边界多样性的比率



**3.2.2全局多样性**

我们的拓扑驱动多样性计算的第二种方法依赖于完全扩展的目标特定扩散子图的结构信息的可用性。在这节中，Gt代表完全扩展后的目标扩散子图。

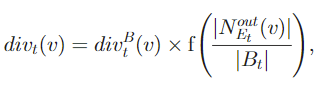
定理4：边界跨越应被视为特定目标扩散过程的外源，因此直观地与来自社交图其余部分的外部影响源相关联。

定义9：节点的边界多样性：给定一个v属于边界集合，v的边界多样性定义为他对边界多样性的贡献。不在Bt中的节点的边界多样性被定义为0



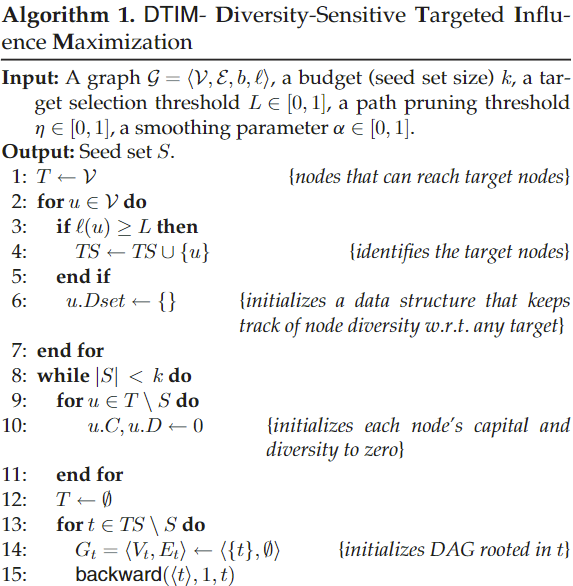
此外，我们观察到，从向给定目标传播的节点的多样性最大化的角度来看，节点多样性的总体度量不仅应当明显与其边界多样性成比例，而且应当与其向外的内部跨度成比例。

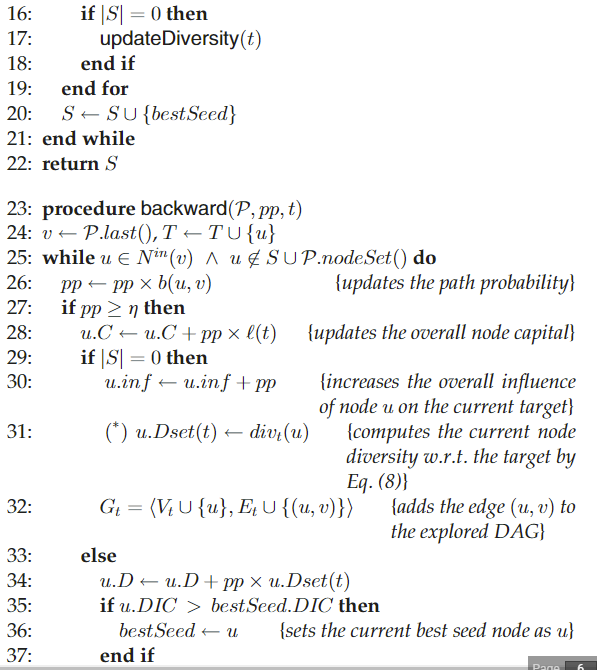
定义10：全局多样性：一个节点v的全局多样性定义为：

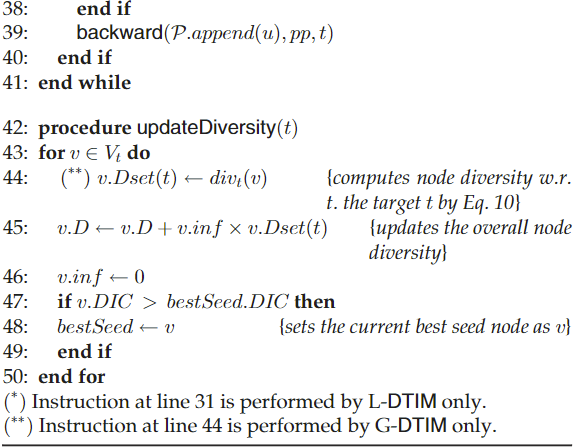
其中使用平滑函数f给outward internal跨度分配至多等于边界多样性的权重

**3.3 DTIM算法**

根据局部多样性和全局多样性，我们提出两个算法分别是L-DTIM和G-DTIM。







L和G都是从被识别为目标的节点开始对扩散图进行反向访问的。为此，首先检查所有节点以计算目标集。为了在主循环(第8-21行)的每次迭代中产生最佳种子集，算法1的两个变体计算到达目标节点的节点集，并在变量最佳种子中跟踪具有最高边际增益的节点。最佳种子节点在每次迭代结束后产生

**4使用DTIM算法使得不活跃用户加入**

我们通过一个特殊的用户参与在线社交网络的案例来评估我们的目标IM框架，这个案例指的是如何将沉默的用户变成社区生活中更积极的贡献者。

所有大型的社区都有一个现象就是参与的不平等性：人群在与其他成员的互动中并不扮演积极的角色，而是扮演沉默的角色。他们被认为是潜伏者。只获取信息却不给社区反馈。社会科学和人机交互研究社区已经广泛调查了解释潜伏行为的主要原因，包括贡献社区智慧的主观理性(而不是恶意动机)，或者认为通过浏览收集信息就足够了，而不需要进一步参与社区。此外，潜伏是可以预料甚至鼓励的，因为它允许用户(尤其是新来者)学习或提高他们对在线社区礼仪的理解。

不管他们的动机如何，潜伏者在社会capital方面可能有很大的潜力，因为他们从OSN获得知识。他们可以意识到不同观点的存在，并且可以利用这些观点来形成自己的观点，但是他们不太可能让其他人知道他们的价值。在这方面，可能需要让这类用户参与进来，或者欺骗他们，即制定一系列战略，鼓励潜伏者通过更积极地参与社区生活，归还他们获得的社会资本。

针对沉默用户的参与行为可以分为四种类型：基于奖励的外部刺激，提供鼓励信息，提高系统的可用性和可学习性，来自资深用户的指导，帮助潜伏者尽快熟悉平台。值得强调的是，我们的方法独立于所采用的欺骗策略。这里的目标是如何在用户参与场景中实例化我们的DTIM算法，其中潜伏者被视为扩散过程的目标用户。所以我们的目标变为：给定一个预算k，找到一个包含k个节点的集合是的对多样性敏感的capital最大，也就是通过不同种子用户激活目标沉默用户的概率最大。

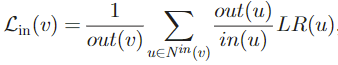
**4.1通过LurkRank找到目标用户**

最初在[[39]和[[41]中提出的潜伏者排名方法旨在挖掘网络中沉默的用户行为，并因此将用户与指示她/他潜伏状态的分数相关联。潜伏者排名方法依赖于拓扑驱动的潜伏定义，该定义仅基于网络结构。假设潜伏行为建立在节点接收的信息量基础上，关键直觉是用户潜伏状态的强度可以基于三个基本原则来确定: 过度消费、所收到信息的权威性、所产生信息的非权威性。上述原则构成了三种不同的排序模型的基础，这三种模型分别考虑了节点的入邻居和出邻居的贡献。根据基于pagerank和AlphaCentrality的公式，提供了潜伏者排名模型的完整规范。下面来介绍LurkerRank

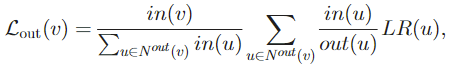
给一个有向社会图G0=（V，E），任意一条边（u,v）意味着v是从u那接受了信息，对于一个节点v的LR值为：



Lin（v）是入节点驱动的潜伏函数

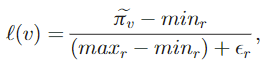


Lout（v）是出节点驱动的潜伏函数

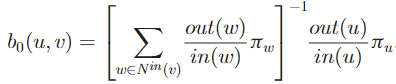


**4.2对扩散图进行建模**

在3.1中，我们使用l（v）来代表节点v的权重用来判断是否为目标。在本章中，如果一个节点的LR值越高，那么他的l（v）就越大。我们定义了节点加权函数，对G0上的LR算法产生的平稳分布进行缩放和归一化。他的缩放弥补了这样一个事实，即LR产生的潜伏者分数，尽管分布范围非常广，但是数值会很低。此外，我们引入了一个小的平滑常数，以避免最高潜伏分数精确地映射到1。对于每一个v，定义一个节点潜伏值l（v）

，表示潜伏者排名分数(pai)除以pai中最小值数量级的基数-10的幂的平稳分布.

为了定义边缘权重，以便它们表达从一个节点到另一个节点的影响强度的概念, 我们再次利用了从LurkerRank获得的排名解决方案以及社交图的结构属性中获得的信息。我们的主要想法是计算边上的重量(u,v)与其近邻u给出的v的原始潜伏分数的分数成比例。

，根据这个公式，我们定义边权值为

一个节点的潜伏值越大，激活该节点所需的相邻节点就越活跃。

**5.评估方法**

**5.1DTIM设置**

需要一些参数，种子集合大小k，目标选择阈值L，路径修剪阈值n，和参数a。我们设定L-perc为，n为

**5.2对比模型**

我们的对比方案为TIM+和KB-TIM，这些都是目前最新的模型。将DTIM算法与非目标IM算法(如TIM)进行比较，该算法需要评估目标场景下竞争算法产生的种子集的质量。为此，我们简单地让TIM在整个图中计算一个大小为k的种子集，然后我们根据DTIM的设置估计不同目标集上的captital。我们考虑了TIM中的主参数的两个相反设置:1，默认值设为0.1，它提供了强有力的理论保证，但却不利于算法的内存消耗。2，设为1.0，这相反地不提供近似保证，而是提供高的实际效率；请注意，TIM+的作者在[3]也使用了后一种设置来与SimPath进行比较。我们使用TIM+中其他参数的默认设置。

关于KB-TIM，我们修改了基于关键字的目标选择阶段，使其等同于DTIM采用的目标选择。需要两个主输入文件来驱动目标选择:1一种文档术语稀疏矩阵，使得图中的每个节点(文档)被分配一个关键字列表。2关键字查询列表，以便每个查询对应于图表中节点子集的选择。了准备这些输入文件，我们定义了与设置相对应的三个查询

**5.3数据**

我们使用FriendFeed,GooglePlus,Ins网络数据集，注意，为了对后一网络中的信息扩散过程有所帮助，我们选择了与原网络图的最大强连通分量相对应的诱导子图，以下简称为Instagram-LCC。

**6.结果**

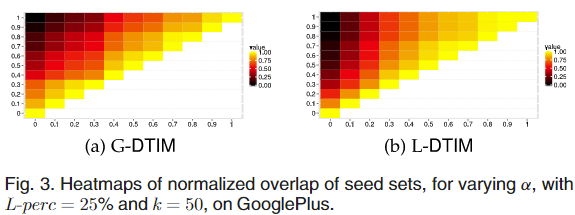
我们根据三个主要目标给出了对我们提出的DTIM算法的评估结果:对识别的种子节点的分析(第6.1节)、对激活的目标节点的分析(第6.2节)和效率分析(第6.3节)。

**6.1 已识别种子节点的评估**

**6.1.1种子集合重叠**

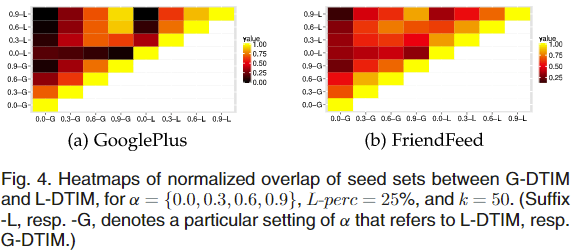
为了研究考虑多样性对种子识别过程的影响，我们初步分析了两种DTIM方法（使用不同的参数a）产生的种子集之间的匹配。这种对种子集的分析有两个方面:(1)通过改变α对特定DTIM方法产生的种子集之间的重叠进行成对评估，以及(2)对特定α值由G-DTIM和L-DTIM产生的种子集之间的重叠进行成对评估。种子集大小为50.

**标准化种子集重叠**。



归一化重叠值跨越整个范围0-1，在对应于G-DTIM的热图中，对于不同子区间中的α值，观察到了50%以上的重叠，在L-DTIM中种子集中的多样性更加均匀。当α为0时，两种方法都没有重叠集合。当使用L-DTIM时，无论目标集大小如何，这些意见通常都适用，而当使用G-DTIM时，较低的L-PERC更有可能出现零重叠的意外情况。

**G-DTIM和L-DTIM种子集的比较**。



当时，在googleplus上(图4a)，对于目标函数中更高的多样性贡献(a < 0:3)，种子集似乎彼此显著不同，而归一化重叠值在0.5-1的范围内当α值较高。可以为FriendFeed(图4b)绘制类似的观察值，但是对于0.6-0.9范围内的a值，重叠值也较低(即归一化重叠约为0.75)。

**比较TIM+和KB-TIM**。我们还分析了由DTIM算法产生的种子集和竞争种子集之间的匹配(结果未示出)。我们这里将α设置为1，因为TIM和KB-TIM在它们的公式中没有整合任何多样性概念。在所有情况下和所有数据集上，DTIM产生的种子集与KB-TIM模型的重叠是最小；但是DTIM和TIM+确有较高的重叠。

**总结**。种子集重叠分析表明，在种子鉴定方面，考虑多样性可以在DTIM方法的行为中产生显著差异。事实上，通过在其全部值范围内改变a，导致标准化种子集重叠的值范围很广。特别是，当在较低的范围下改变a时，重叠的变化更明显，因此表明多样性的贡献（capital）导致更显著多样化的种子集。值得注意的是，当比较分别用a=0和1获得的两个种子集时，重叠可以接近于零，即，当仅考虑目标IM目标函数中的多样性时，可以识别完全不同的种子节点。这两个被提出的多样性概念被证明是非常不同的:事实上，L-DTIM和G-DTIM分别产生的种子集的归一化重叠通常低于50%，对于较低的a值进一步减少

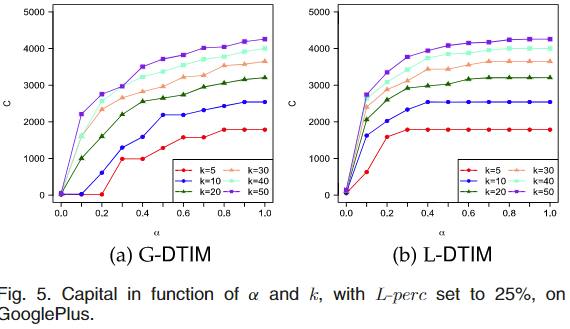
**6.1.2 种子的结构特征**

我们分析了已识别种子的拓扑特征，重点讨论了节点中心性的基本度量，即外度、中间性和核心性。由于篇幅有限，我们在此提供了主要发现的摘要，并请读者参阅在线补充材料中的附录，了解详细结果。突出的一个主要观点是，在DTIM方法中考虑多样性产生了选择种子节点的效果，该种子节点可以不同于根据所选拓扑标准以其他方式获得的那些节点(即，在目标函数中仅使用capital)。 这一结果，再加上之前从集合重叠方面得出的关于多样化的类似考虑，因此加强了在目标IM过程中考虑多样化的重要性。

**6.2激活目标节点的评估**

**6.2.1 Capital**

我们讨论由给定种子用户集合激活的目标用户的预期capital的结果。除了预期的资本增加a(这意味着在目标函数中加权的多样性比资本少)，a对DTIM算法行为的影响是显而易见的，特别是对于k>10的情况，capital值可以变化多达三个数量级. 在所有数据集上，L-DTIM在所有k值上都比比G-DTIM观测到的更高的capital值。在图5中用GooglePlus数据集，但使用G-DTIM，capital’值迅速上升，直到80%，在L-DTIM中我们观察到一个更迅速的增加当a很小的时候



相似的行为在FriendFeed上显示，尽管k<30的增长趋势不是单调的。在Instagram-LCC上，这种网络的相对较小的规模和较高的连通性使得资本价值在15%左右对于所有的a。

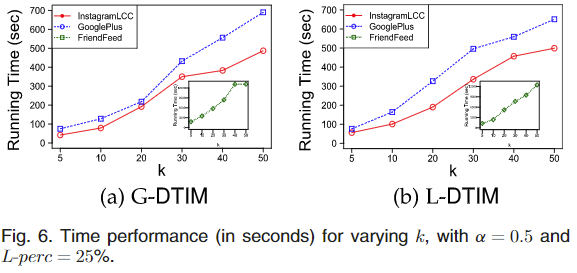
与TIM+和KB-TIM的比较。在DTIM中得到的capital值比在这两种方法里得到的在任何时刻，任何参数下都要高。

**6.2.2目标激活概率**

进行了进一步的评估，以了解α和k的不同设置如何影响DTIM方法所针对的节点的激活可能性。我们把一个节点的激活概率看作是它被激活的次数除以蒙特卡罗模拟的运行次数。由于篇幅有限，我们在此总结了与本次评估相关的主要发现，并请读者参阅在线补充材料中的附录，了解详细结果。对于两种DTIM算法，激活概率随着α增加而遵循非下降趋势。获得高激活概率的可能性随着α的增加而增加，即具有高激活概率的目标节点的数量随着α的增加而增加。然而，在两个最大的数据集上，我们也观察到选择相对较大的k会导致相当一部分目标节点在α=0.1的情况下也有中高概率被激活，从而表明即使具有高度不平衡的资本和多样性，目标节点也可以被激活。相比之下，当选择k较小时，a值的微小变化会显著影响更可能激活的目标节点的数量。

**6.3性能分析**

图6显示了G-DTIM和L-DTIM在不同网络中的的时间性能



FriendFeed在里面的小框里因为它的网络太大，运行时间太长，小图的纵坐标最底下是3000.

TIM+和KB-TIM毫无疑问的比两种DTIM运行效率要高，但是没有给出图例。