Seed Set Selection in Evolving Social Networks

1.introduction

在本文中，我们试图解决这个问题，在静态网络中选择大小为k的最具影响力的节点，以便在社交网络进化过程中得到更好的结果，这就是进化社交网络中的种子集选择。

我们可以给出一个实际应用的例子。

Contribution：首先，作者对在演进中网络选种这个问题做了一个定义。**：**显然，我们只能在t = 0时获得网络，而在其他时间戳时网络的演化是未知的，因此，我们提出了基于公共邻居的图预测算法来预测网络的演化，然而现实社会网络的规模一般都比较大，所以需要花费大量的时间对网络中的所有节点进行预测。因此，我们提出了一种有效的选择节点来预测网络演化的算法，通过使用该预测算法，可以在以下时间戳中生成一系列网络快照。最后，提出了从网络快照中选取最具影响力的k个节点的策略。

**2.preliminaries**

**A social network evolution**

一个演变的有向社交网络在t时刻可以用Gt=(Vt,Et)表示。由于网络的不断演进，网络结构就会不断变化，生成一系列网络快照，用表示，T是网络演化的周期，我们假设T是有限的，这个网络叫做进化网络。

**B IC模型**

在时间t时，网络为Gt=(Vt,Et)，每个节点只有激活或者未激活状态。在时间t时，节点w是未激活的，N（w）是节点w的邻居，对于每个节点u属于N(w)，节点u有一次机会去激活w，激活概率为P(u,w)。但是如果w的所有邻居全部失败了，那么w仍处于未激活状态，如果w被激活那么在t+1时刻变成激活状态。

**C IM problem**

给定一个静态有向的社交网络Gt=(Vt,Et)，一个时间t，一个整数k。IM问题要找到一个包含k个节点的集合使得影响范围最广。

**D 蒙特卡洛模拟方法**

蒙特卡罗模拟是一种有效而简单的估计影响扩散和对静态社交网络快照进行影响的算法。给定一个大小为k的初始节点集，它运行该过程数千次，并报告初始活动节点集的平均影响范围，这个值是预期的影响力传播。作者将在提出的算法中使用这种方法。

**E Degree Discount Algorithm (DDA)**DDA算法有以下几个优点：它可以很容易地实现，它是非常快速和可伸缩的。其大意如下。N(w)是节点w的邻居节点，如果节点v属于N(w)并且节点v已经被选择了，那么他不被计算到w的出度中。

**问题定义**：

G0代表在时刻t=0时刻的网络快照。在任何时间戳t>0时，网络结构都会变化，并且这个变化是不知道的，在时间t=T，一系列的网络快照会被生成，该问题的目的是在G0的信息只可用的情况下，选择G0中大小为k的种子集。使用S表示种子集，使得在中获得最好的影响力。

**3.提出的方法**

**A Selecting The Nodes To Predict**

从直观上看，真实的在线网络规模一般都比较大，因此需要花费大量的时间对网络中的所有节点进行预测。一般来说，大多数节点的变化很小，并且在接下来的时间戳中，网络上解决方案的变化最小。因此，我们可以选择部分节点来预测网络，而不是使用所有的节点。一种朴素的选择方法是选择一些概率相等的节点。然而，将真实网络和预测网络的影响最大化差异最小化几乎是不可能的。

作者的策略是制定一个目标函数方程(1)，策略表示将真实网络和预测网络之间的损失最小化的概率。作者提出了最大间隙选择(算法1)优化我们的目标函数，其中MGS算法是从[5]中提出的MaxG算法改编而来的。其中，MaxG算法通过探测一些节点来更新图形快照，然后选择种子集。

**MGS算法**：假设我们有两个在t=0和t= 1时刻进化的社交网络的图快照，它们分别表示为G0和G1，假设节点v是属于G0和G1的交集，G定义为网络，根据G1更新G0中节点v的连接，分别在G0和G中采用度折现算法(DDA)选择最优种子集S0和S，结果的差值可以用来表示，它称为节点v的性能差异。代表集合S的影响范围。

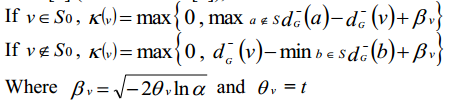
对于每个v属于G0，PGv都应该被估计，并且根据它选择满意的节点。条件是：



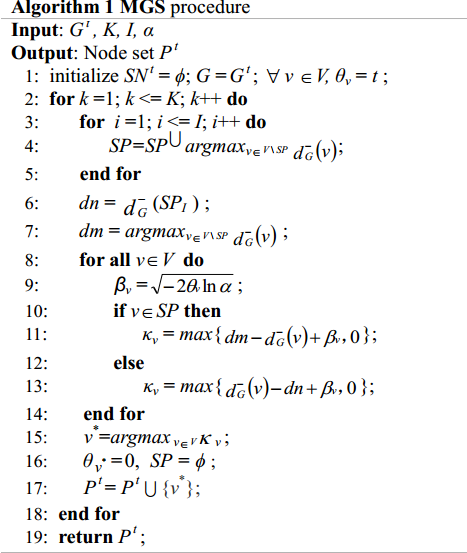
其中α是一个数值，只有大于它的PGv才能被选中。

作者的MGS算法就是要选择可以将k（v）最大化的节点。

很多种方法都可以估计，本文利用种子集S的入度和对其进行估计。如果每个节点的入度分布不被计算的话，k（v）的计算就不准确。，所以作者使用Azuma-Hoeffding不等式来计算k(v)。



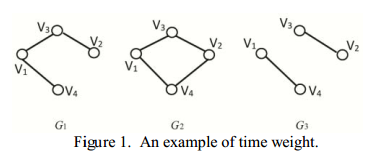
MGS的算法如下：



**B Graph Prediction Based On Common-Neighbors**

**链接预测**：假设现在知道在时间t时的一个静态网络Gt=(Vt,Et)，随着时间的推移，它会不断进化，增加新的节点，增加新的边，失去节点，失去边。然后将生成一系列图形快照并且被表示为，链路预测的目标是计算网络中每对节点之间边缘的出现概率。根据[9]中提出的策略，提出了一种改进的方法。注意我们的方法是基于有向图的，而[9]是基于无向图的。

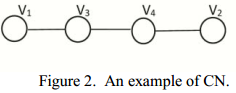
**时间度量**：假设已知三个图快照分别表示为G0, G1, G2。现在可以用几种方法来预测G3。现有的方法通常基于G2预测G3，而不考虑链路的存在时间和频率。在图1中，仅考虑G3时，G3中e(v1,v2)的发生概率相对较低，反之，考虑G0, G1, G3的结构，则G3中e(v1,v2)的接触概率相对较高。



所以作者定义了time weight Tw(t) 突出网络结构在以前时间戳中的重要性。

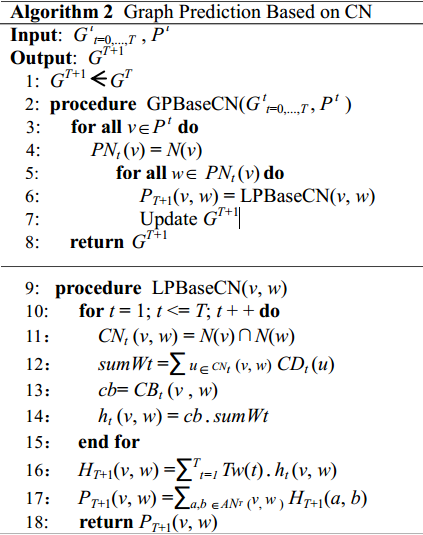


**普通邻居度的改变**：该算法具有简单、高效、性能好等优点。然而，大多数基于CN的算法只考虑一跳内的CN。在图2中，考虑到一跳内的CN，节点v1和v2之间没有CN，但是e(v1,v2)的接触概率并不小。为了提高准确率，本文选择考虑CN的两跳。



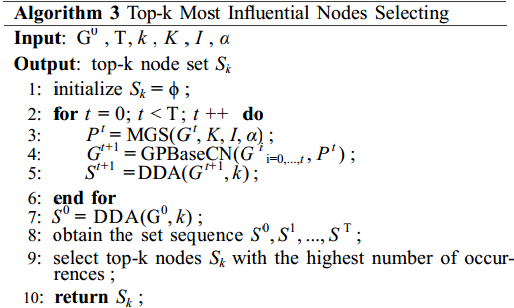
**Closeness between CNs**: 在图3中，图G和G’中节点v1和v2之间的CNs是相同的，同理v3，v4，v5也是相同的。在此基础上，定义了中CNs在t时刻的紧密度

**Graph Prediction Based on CN**：



**C Selecting The Most Influential Nodes**

该算法用于选择一系列网络快照中最具影响力的节点。给定一个时刻为t的静态社交网络Gt使用算法1可以选择一些节点，随后使用算法2来绘制预测图Gt+1随后会有一系列的预测图被绘制。然后，利用度折现算法(DDA)生成在图影响力最大的节点序列。



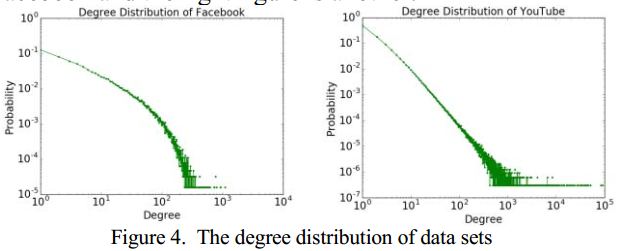
**5 实验结果**

**A 数据集**

两个真实网络，一个虚拟网络

**虚拟网络**：G0中有1000个节点，在每对节点之间建立链接的概率为0.02。然后对于每一个时间戳，我们首先随机选择200条概率相等的边，然后将它们两头的节点进行交换，这些节点随机选择的概率与它们的自变量成正比。

**真实网络**：Facebook，63731点,1634070边；YouTube有3,223,585个顶点18750748边。作者随后将两个网络中度的序列分布画出：



**B 实验建立**

作者使用IC模型，将传播概率设置为：，入度分之一。

为了验证我们提出的方法，对以下算法进行了比较。

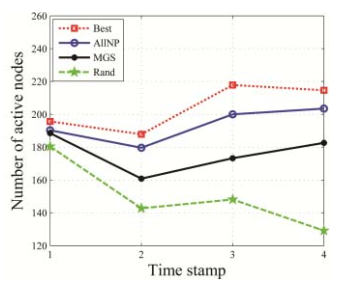
Best：一系列的快照全部都知道，用DDA算法进行选择。

Baseline：按照G0选择的种子当做最佳的种子。

Ours：

**C 实验结果**

1）MGS算法的分析：



下图是在虚拟网络中的结果

