## **Leave or not leave? Group members’ departure prediction in dynamic information networks（5.29 李昊展）**

离开还是不离开？小组成员在动态信息网络中的离职预测



### **基本信息**

期刊：《Information Science》

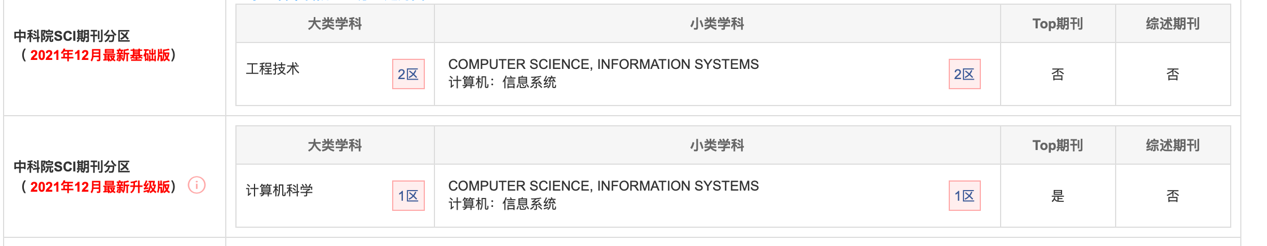
发表时间：2021

作者：Xinrui Wang, Hong Gao, Zhipeng Cai, Jianzhong Li

作者信息：

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Department of Computer Science, Georgia State University, Atlanta, GA, USA



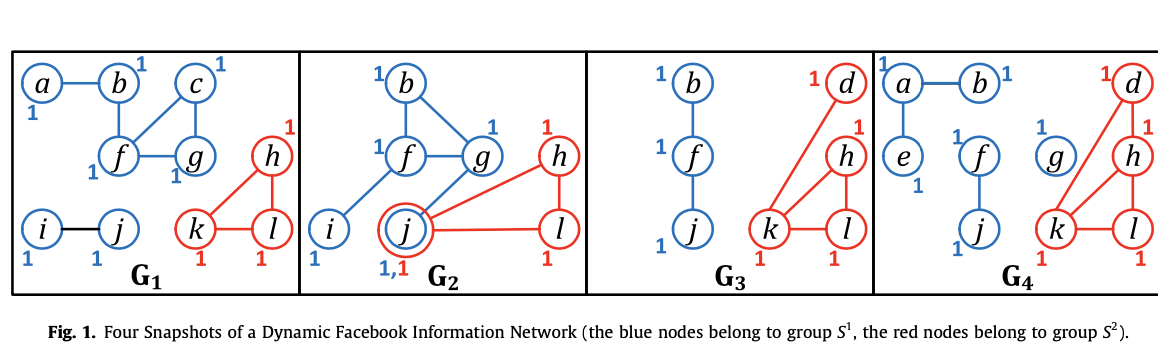
### **文章主要研究内容**

包含不断发展的节点和链接的动态信息网络存在于各种应用中。例如，在Facebook网络中，节点代表用户，链接代表朋友关系，用户通常组成不同的组。随着时间的推移，一些用户会离开一些群组。因此，对于用户和组来说，预测哪些用户将离开哪些组是有意义的。现有研究表明，如果大多数朋友离开，用户更有可能离开。然而，在现实中，低度节点通常占所有节点的大多数，并且存在一些用户几乎没有朋友，但一直呆在群组中。因此，如果只使用邻域信息，可能会导致精度损失，但之前的工作忽略了这一点。为了提高预测精度，本文首先分别引入了群体活跃度和用户自身活跃度的新定义。然后，我们提出每个用户的分组组合活动评分，以便分数较低的用户更有可能离开。之后，提出了一种无监督预测算法，以持续预测组成员在动态信息网络中的离开行为。对几个真实数据集的实验表明，与最先进的竞争对手相比，算法的有效性、效率和超现实性。

### **文章内容**

1.小组成员的离职示例。

Facebook信息网络由大量用户组成。如果他们是朋友，他们之间就存在联系。用户通常组成不同的小型在线聊天组，以便他们可以更方便地分享一些他们都感兴趣的东西。图1显示了带有四个快照（G1到G4）的Facebook信息网络示例。蓝色节点属于S1组，红色节点属于S2组。请注意，用户可以同时属于多个组，例如快照G2中的用户j。每个节点上标记的数字表示用户在快照时间戳下在组中发送的信息数量，相同的颜色对应于同一组。根据之前的工作，在快照G4中，用户g预计将离开S1组，因为它的度为0（这意味着他/她在第四个时间戳的S1组中没有朋友）。然而，实际上，由于他/她对S1中其他人共享的信息的长期兴趣，g在第四次时间戳后一直留在S1中。观察G1到G4，我们发现g在三个快照上在S1中发送了消息，并且S1组从第一个时间戳到第四次时间戳保持活跃，因为其成员在此期间一直在聊天。出于这个动机，本文建议利用用户在群组中的活动信息和群组自身的活动性来预测用户是否会离开群组，而不仅仅是他/她的邻居信息。

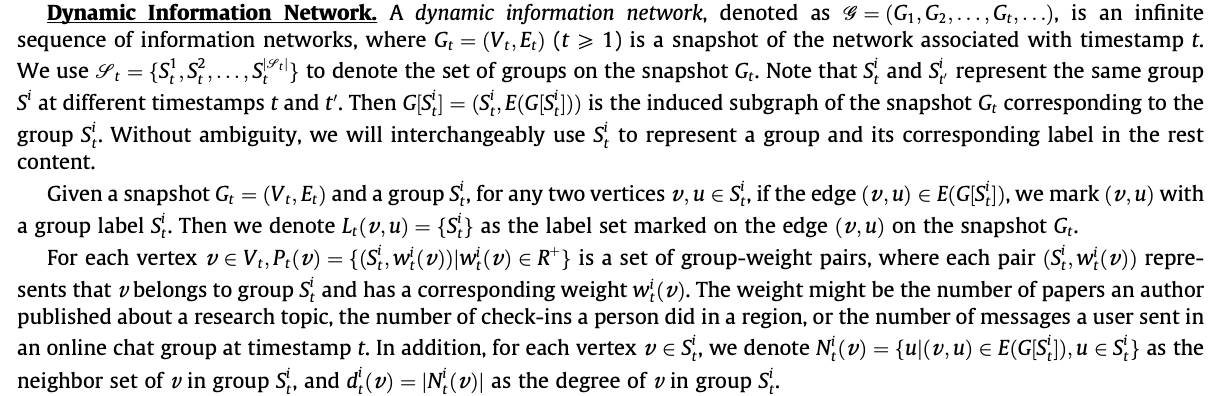


另一个例子是DBLP信息网络由许多作者组成。他们之间的联系代表着合作关系的联系。每个小组对应于一个研究主题。如果作者发表了一篇关于该主题的论文，则他/她属于主题组。有一些作者多年来一直在发表关于某个主题的论文，但很少有合作者。如果只使用作者的合作者数量来预测他/她在某个主题上的放弃，可能会导致一些准确性损失。在本文中，预测作者是否会根据自己的出版物和随着时间的推移该主题的发展放弃研究主题。

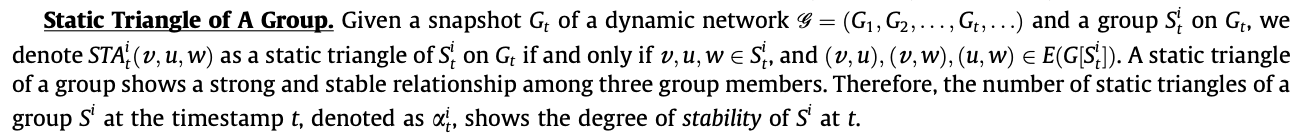
针对这个问题，第一个挑战是如何定义团队的活跃度和用户的活跃度，以及如何认真计算这些测量结果。另一个挑战是如何将它们组合到用户组组合活动评分的定义中，以便得分较低的用户更有可能离开群组。

2.问题概念

（1）动态信息网络

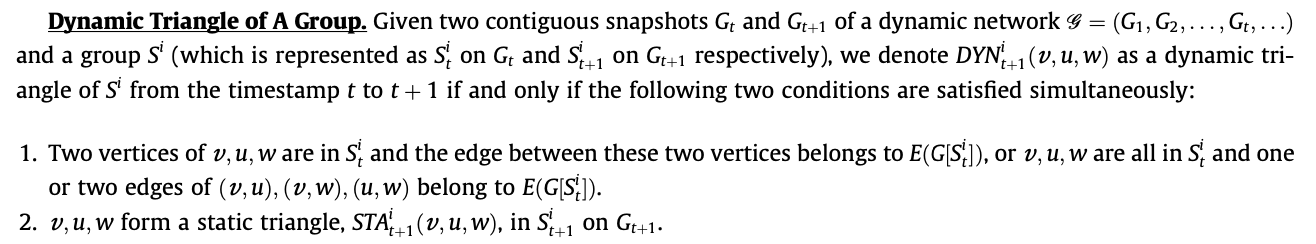
****

（2）静态三元组

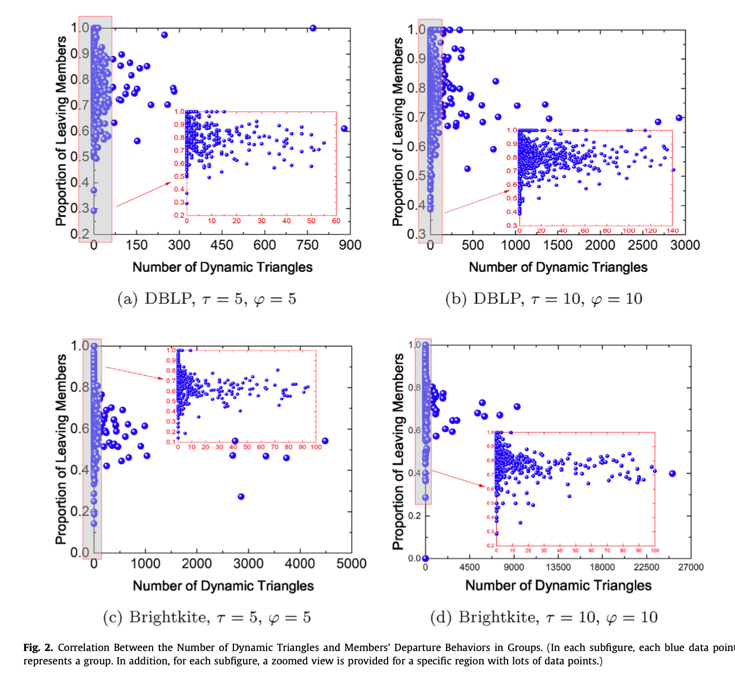
****

一个组的三个成员互有连边，体现三个小组成员之间牢固稳定的关系。

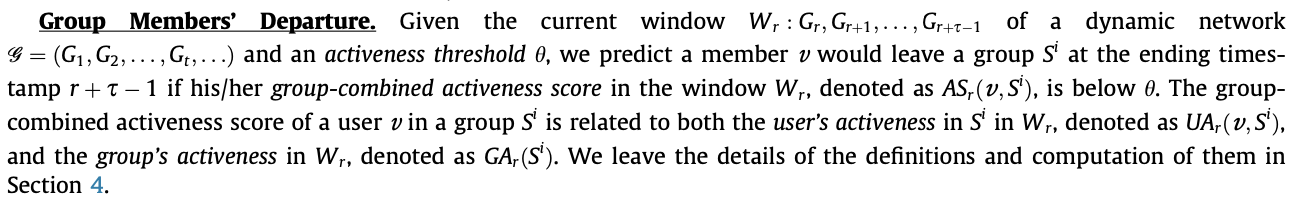
（3）动态三元组



两个连续快照之间的组的动态三角形表明，这三个组成员之间的关系从t到t+1不断增强，显示了Si从t到t+1的连续性程度。图2展示了两个现实世界动态网络DBLP和Brightkite的一些实验结果。在每个子图中，蓝色数据点代表一个组。如所有子图所示，一个组包含的动态三角形越少，其离开成员的比例就越大。也就是说，动态三角形数量较少的组中的成员更有可能离开组。



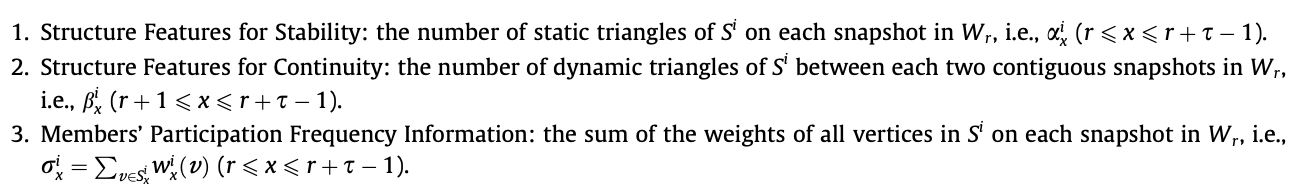
（4）组成员离开

****

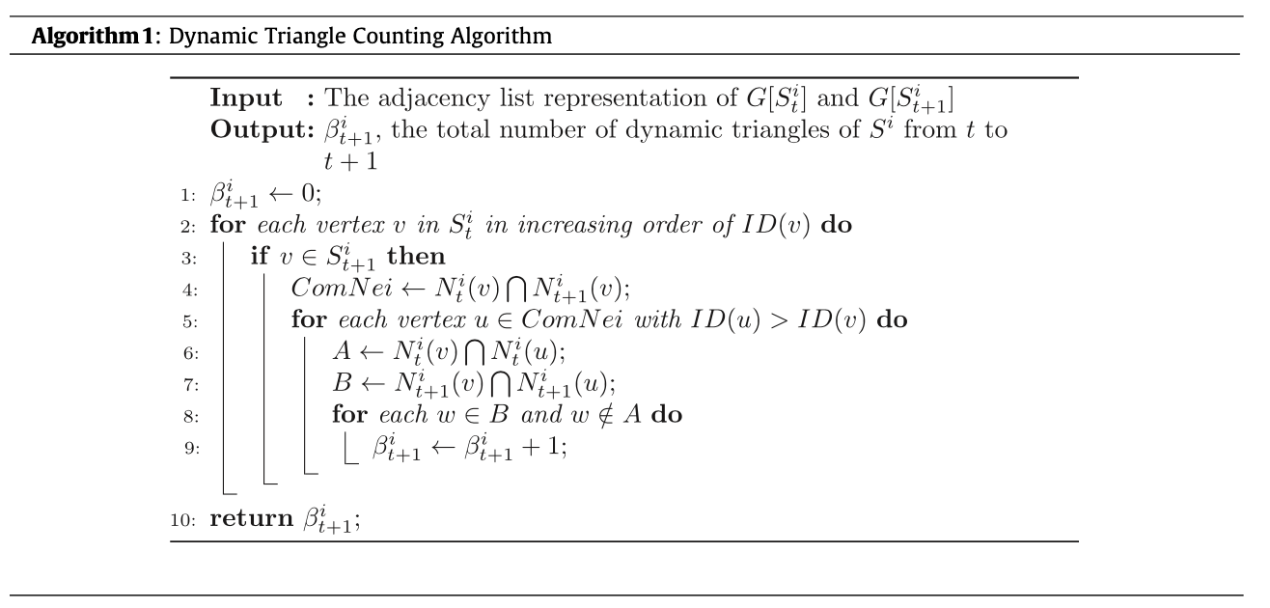
3.组成员离开预测

（1）组活跃度影响因素

为了定义一个团体的活跃性，应该考虑其结构特征和小组成员的活动。首先，如果一个群体在历史快照上包含更多的静态三角形，那么在此期间，组成员之间的关系会更稳定，并且该组现在和将来解散的可能性相对较低。其次，如果一个组在两个连续的历史快照之间存在一些动态三角形，这意味着该组吸引了新用户加入，或者一些成员在组中建立了新的关系，以便该组有能力保持其活力。第三，如果在几个时间戳期间，成员在群组中的参与频率总是很高（例如，在线聊天组中发送的消息数量保持稳定或正在增加），则该组在现实中更有可能具有重大意义，不容易解开。



（2）动态三元组计数算法

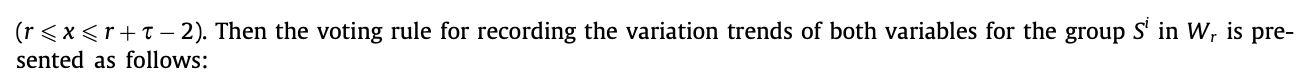
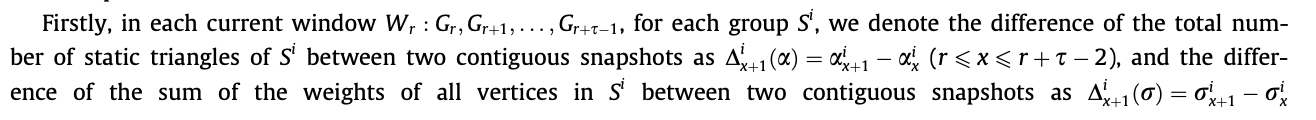
****

首先寻找两个相邻时刻的动态网络快照中组Si的所有共存边，接着对每条共存边，计算其在Si,t中存储在集合A中的两个端点的共同邻居，以及存储在B中指定的Si,t+1中的端点的共同邻居。之后，每个顶点（在B中，但不是在A中，具有共同边缘的两个端点，形成从时间戳t到t+1的群Si的动态三角形，并计数。

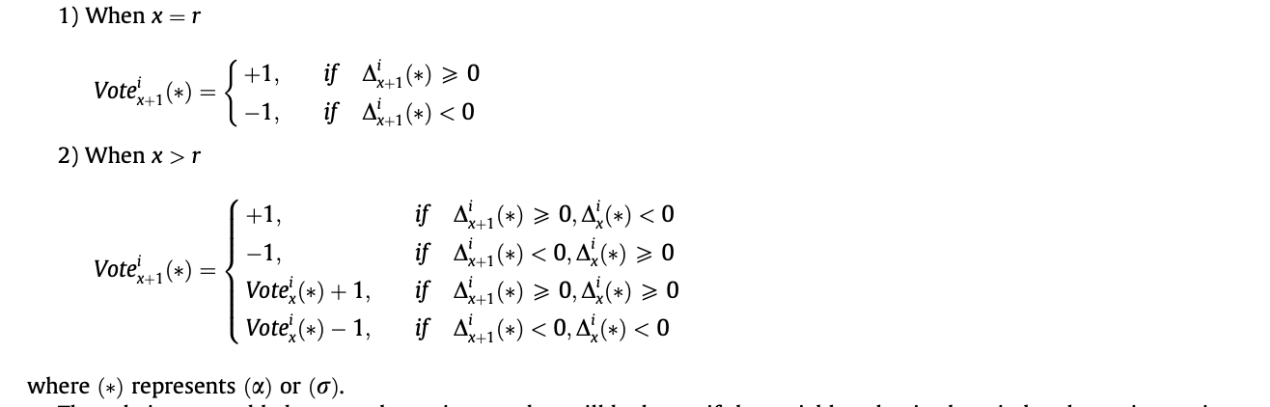
（3）投票规则

在获得上述所有信息后，使用投票规则来记录一个群体的变化趋势。为了反映变量的变化趋势，通常使用特征曲线。然而，在现实世界中的动态网络，在每个当前窗口中绘制静态三角形的数量和所有顶点的权重之和的曲线是非常昂贵的。因此，我们引入了一个投票规则，与特征曲线相比，它的成本很低，可以描述一个变量的变化趋势。

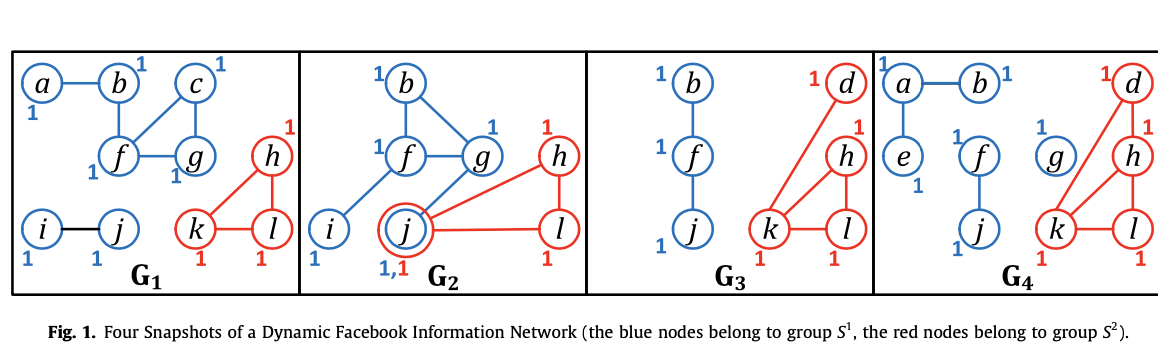
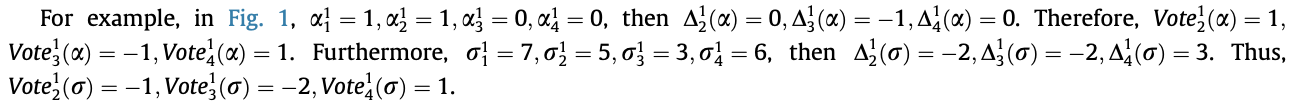
计算各时间戳间各组的静态三元组和顶点权重和的差值。



设定投票值，随着网络快照组中变量值的增加而增大，变量值的减少而减小。

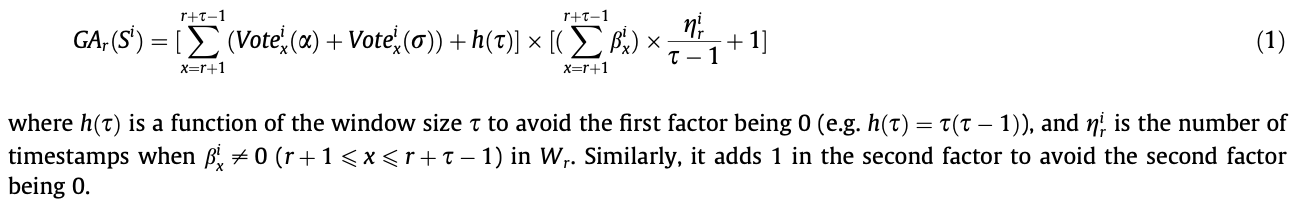


以图1为例：



（4）组活跃度

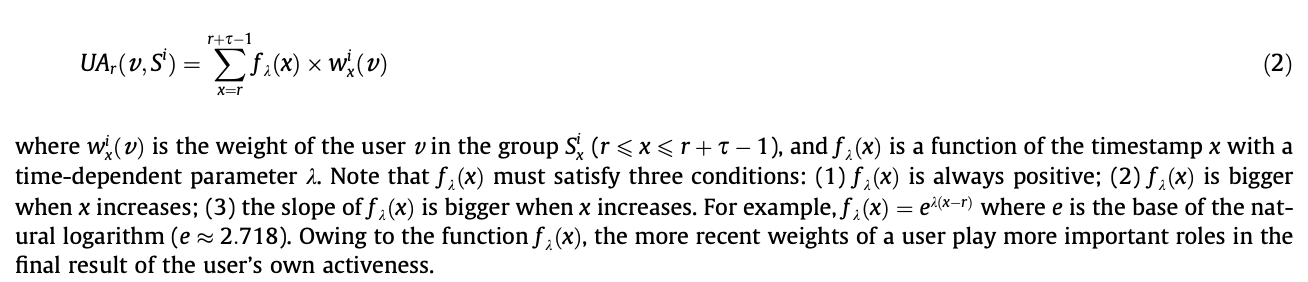
考虑到结构特征和成员的活动信息，我们在当前窗口Wr中定义组Si的活跃性如下：



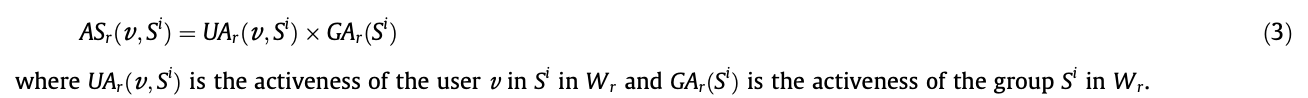
这个定义是合理的，因为如果（1）它在每个快照上包含的静态三角形数量保持不变或随时间增加；（2）它在每个快照上的所有节点的权重之和保持不变或随时间增加；（3）它在窗口的每两个连续快照之间包含更多的动态三角形，那么这个组就比较活跃。

（5）个人活跃度

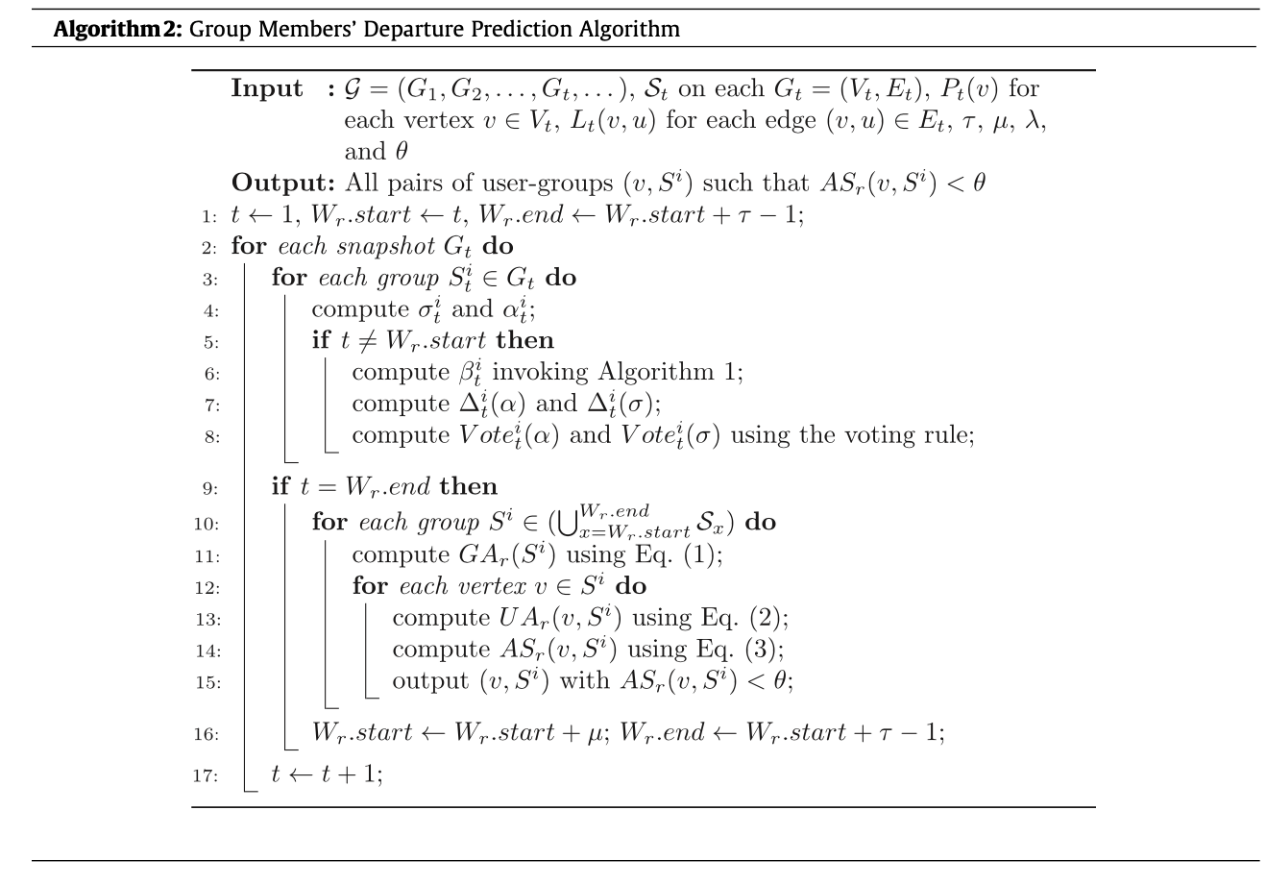
显然，用户在群组中的活动信息直接反映了他/她在群组中的活跃程度，且用户在群组中的活跃程度与他/她在一段时间内的最新活动更相关。以在线聊天组为例，如果用户在最近的时间戳发送消息，他/她可能会更活跃，离开群组的可能性更小。因此，我们引入了用户与时间相关的活动性的定义，以便将时间因素集成到活动性定义中。



（6）用户的组组合活动度得分



（7）组成员离开预测算法



4.算法表现评估

（1）实验设置

对比方法：the Core Number-based method (CoreN)、the method based on neighbors Out- side and In groups (Out-In)

数据集：Brightkite、Gowalla、US Patent、online bibliographic

http://snap.stanford.edu/data/loc-brightkite.html

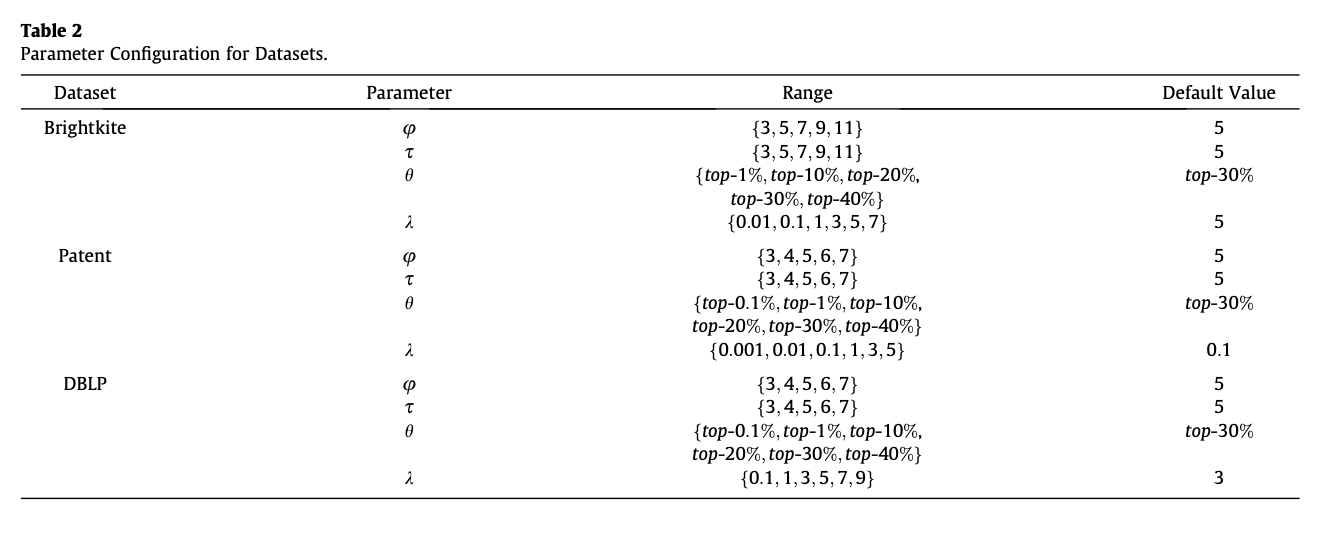
http://snap.stanford.edu/data/loc-gowalla.html

http://www.nber.org/patents/

https://dblp.dagstuhl.de/xml/release/

离开行为界定：在设定时间段（real size）内，该成员不在组内，则判断为离开。

参数：real size、window size、activeness threshold、time-dependent parameter



度量指标：

