本科毕业论文（设计）开题报告

动态社会网络团体发现与挖掘算法研究

An Approach to Discover and Analyze Communities in Dynamic Social Networks

1. 课题背景

社会中人与人之间的错综复杂的关系，可以抽象成一个非常复杂的网络图，每个人就是这个图中的一个节点，而人与人之间的关系就是这个图的边。人和人的关系有陌生有紧密，紧密的关系使相关人们形成一个团体。研究社会网络中的团体的发现方法和进化规律对于社会学研究和相关应用有非常重要的意义。例如，对犯罪分子的犯罪网络团体的挖掘和研究能够帮助公安机关深层次地挖掘埋藏在人际关系中的重要线索，辅助公安刑侦人员的案件侦破，对公安刑侦具有非常重要的现实意义；研究互联网社区(如微薄、论坛等)团体的进化，有助于相关组织者把握最新的流行趋势，帮助参与者，也能对商业决策提供有力的支持。

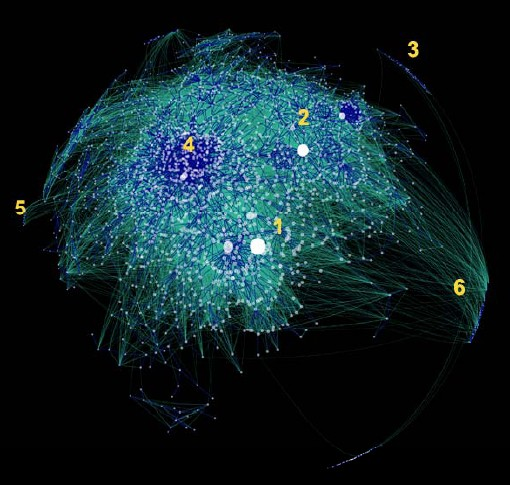


Figure ：Blogosphere博客圈数据网络图。社会网络图的一个示例

而随着时间的变化，社会网络中的团体构成随时都在发生着变化。相关的团体可能发生如分裂、合并、生长、消亡等变化。研究动态的社会网络中的团体变化和其变化规律，可以帮助社会学家研究社会发展的相关规律，帮助商家和受众掌握最新的流行趋势，可以帮助政府部门根据社会的发展制定最优决策……

社会中人与人之间的联系表现为人们的电话通信记录、邮件记录、共同参加一个活动的记录等等。使用一定的算法和数学模型对这些数据进行建模，可以把社会网络抽象为一个网络图。研究基于这个网络图上的团体发现和分析的相关算法，并开发有关的计算机程序，可以自动化地高效地分析数据，挖掘出有用的信息和规律。

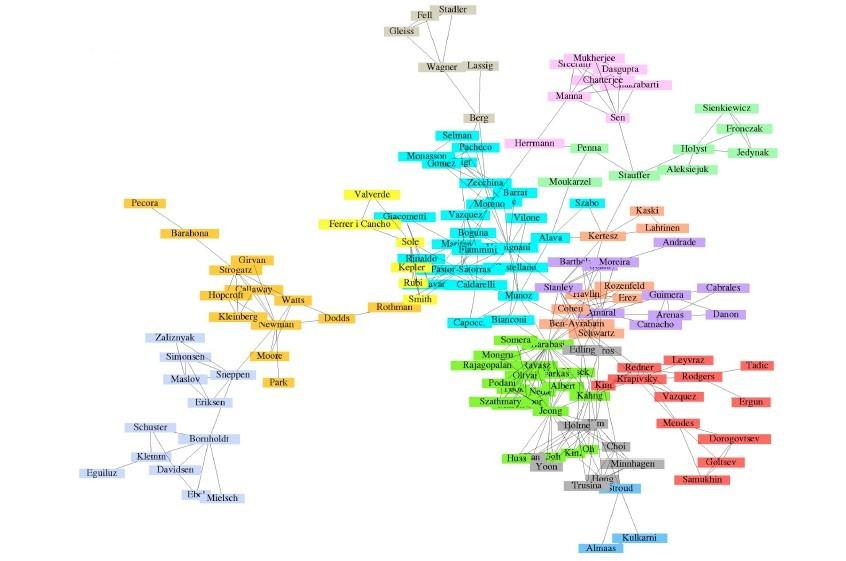


Figure ：研究合作作者关系网络，社会网络的一个示例

1. 目标和任务

对于静态的网络进行团体发现和挖掘，已经有相当多的研究和算法，如层次聚类、k-means聚类、betweenness切片划分算法等等。这些算法能够根据不同的聚类相似度依据对静态网络进行分析并挖掘出其中的团伙。对于静态网络，也有很多的研究和方法来分析团伙内成员、团伙间的联系。关于静态网络的数据挖掘现在已经有非常成熟的知识体系和方法。但是，现有的静态网络的挖掘方法对实际分析的支持还并不充分，因为实际的社会网络，总是处在时刻变化中。如果有一套行之有效的对于动态网络的分析挖掘方法，将为相关领域提供非常大的帮助。

与静态网络不同，在动态网络中，由于节点（人）之间的联系随着时间轴的推进而不断变化，而网络结构也时刻在发生着改变。使用传统的静态分析挖掘方法就难以提供有效地对动态网络进行分析和挖掘。本次研究的目的，就是研究出一套较有效的挖掘方法和算法对动态网络进行分析，挖掘出动态网络中的团伙信息，并应用这套方法分析和研究在动态网络中的团伙进化现象及进化规律。概括之，其主要研究点包括以下几点：

1. 研究并得出一套有效的对动态网络进行挖掘和分析的方法和算法
2. 设计软件实现相关的算法
3. 使用实际数据集对实现的算法进行实验验证，对实验结果进行分析。针对数据集探讨动态网络中团伙的演化现象和演化规律

1. 研究方案和关键技术考虑

（本部分仅列出研究的思路考虑和关键技术的关键内容，相关算法的解读将在文献综述部分作进一步地讨论。）

对于基于时间线的动态社会网络，可以定性为分析在以时间维度为动态线，网络中团体的演变过程和个人的活动过程。以时间维度为基线，一个最行之有效的方法就是对动态网络在各个时间点上的状态做切片，然后分析各切片之间的演化。从这个思路出发，本次研究的两个基本点即研究重点为：1.单时间片上的静态团体挖掘；2.时间片之间的团体演化分析与个人活动分析。



Figure 3：时间切片，从左到右分别是：a)时间切片t=1；b)时间切片t=2；c)时间累积切片t=2

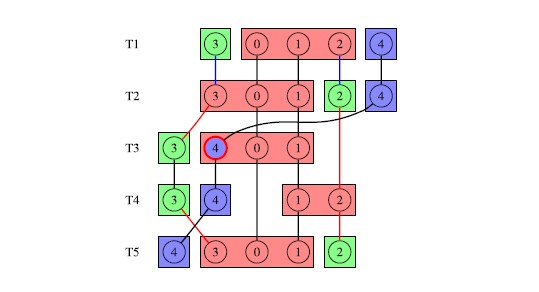


Figure ：随着时间的变化，图的结构发生着变化。图四展示了5和时间片上的图结构

静态的网络挖掘，我们采用一个聚类分析的过程进行团体挖掘。将采用的方法是Michelle Girvan 和Mark Newman提出的Betweenness切边算法。算法基本思路比较简单，网络中的关键路径是团伙之间的联系，通过去掉网络中的关键路径，即可以发现团伙。

Betweenness切边算法的效率关键落在关键路径的查询上了。经典算法中采用各个点两两之间的最短路径的叠加来寻找关键路径。通常的算法中，Betweenness计算的算法复杂度为O(n3)，空间复杂度为O(n2)，其中n为网络中节点的个数。我采用Ulrik Brandes提出的一个优化的betweenness计算方式，其计算空间复杂度为O(m+n)，时间复杂度为无权图O(mn)或有权图O(mn+n2log n)，其中n为网络中节点的个数，m为网络中边的条数。

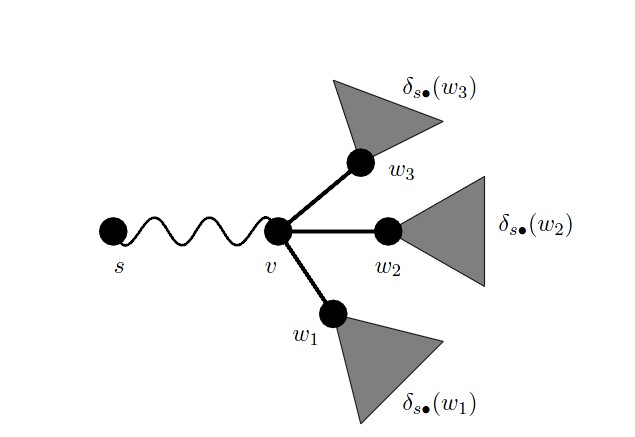


Figure ：改进的betweenness算法示意图

如所示在改进的betweenness算法中，Ulrik运用动态规划的思想，利用已经计算完成的路径的关键度权重来支持后面的计算，减少了重复计算，提高算法效率。

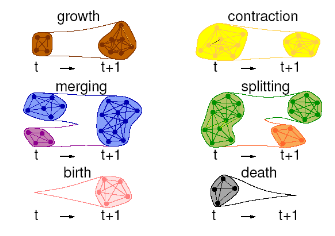
切片算法有个与生俱来的缺陷，就是需要事先指定一个切去的边的条数，或者是最终团伙的个数。同时，对于聚类分析的团伙结果，我们也需要做一个质量检测。因此，我们还需要一个检测团伙聚类质量的度量。M. E. J. Newman 和 M. Girvan在其经典论文[引用编号]中就提出一个聚类效果的度量：Modularity Qualify。Modularity再具体介绍一下？Modularity度量也是本研究中的一个关键技术。

有了度量，我们可以对切边算法做进一步优化，使其自动化起来。通常来说，逐一切片，网络的MQ度量会呈现一个或多个峰值，最高峰值是理论上的最优聚类结果。我们对网络逐一切边，记录过程中的切边过程和MQ度量。以实现两个方式发现最优结果：a.贪心法找到第一个峰值结束；b.逐一切去所有边，找到最高峰。另外切边过程中记录过程值，可供使用者手工调整达到最优。

3.2

3.1中单切片上的团体发现已有一个比较完善的方案。下面是动态网络的分析。

时间维度上的团体演化，可以分为以下几个基本动作:1.团伙出现；2.团伙瓦解；3.团伙分裂；4.团伙合并；5.团伙保持；6.团伙成长；7.团伙衰退。对于个人来说，两个基本动作：1..进入团伙；2.离开团体



//图的注解

//可以分两点讲：1.团伙匹配。2.事件发现

无论是个人行为发现还是团伙演化研究，其中一个基本点也是难点在于切片之间团体的对应关系的发现和保持。简单解释就是在切片t上的团体A0，到了切片t+1上，我们要在分出的团体中找到和A0对应的团体A1。如果发生的是分裂和合并还需要追踪到最为相近的一系列团伙。追踪的过程比较复杂也是动态挖掘的难点所在。研究的思路还是要化纷繁为简单，抓住核心。首先最基础的一个基线是找到最相近的团体。

要找到相近团体，最基本的是相似度的度量，我们使用Jaccard系数进行度量即C1,C2的相似度 = |C1 ∪ C2| / |C1 ∪ C2|

切片t中团体集合St0={P0,P1,P2,…,Pn}，切片t+1中团体集合St1={Q0,Q1,Q2,…,Qm}，两两之间(Pi与Qj间)都会有一个相似度Jaccard系数。对于单团，取出最大的相似度对应团是非常自然的想法。但是不能这样简单地处理，必须全局地看待这个问题。即取出全局上最优的匹配组合。这样，这个问题可以转化为二部图的最佳匹配问题。对于二部图的最佳匹配问题，我将采用Kuhn－Munkras算法来解决二部图匹配问题。

至此，解决个人的行为挖掘应该是比较简单的了。个人的行为无非就是离开和进入一个团体，我们既然在时间片跟踪了团体，个人的行为也就显而易见了。

研究团伙的活动，对于团伙的出现和团伙的瓦解两个动作是比较清晰的了。直接查阅切片间的团伙id就可以知道。对于完全的团伙保持（无任何成员进出），完全的团伙合并（子团完全合并，无成员溢出和加入），完全的团伙分裂（母团完全分裂成子团，无成员溢出和加入），也是比较清楚的。比较相似团体的成员合并关系就可以直接得出结论。

但是，在实际情况中，团伙的行为并不总是完全的行为。还伴随着比较复杂的个体行为。团伙的演化行为虽然是比复杂的，而复杂的活动可能有几个基本活动组合而成，其活动性质介于几种活动之间。这时候，要界定这种活动属于那种活动，最行之有效的方式是引入一个度量活动程度的值。这就是如k-Merge、k-Split的一系列度量方法，来确定该活动应该属于哪种活动。其算法在Sitaram Asur, Srinivasan Parthasarathy等人的[引用编号]一文中被提出来。我们将采用这种处理方式来分析团伙的动态行为。

1. 预期研究结果
2. 研究给出一套有效的对动态社会网络进行挖掘和分析的方法和算法
3. 开发实现出相关的算法，计划使用java为核心语言平台辅助以xml、python、Shell程序及mysql数据库实现相关的算法和数据结构
4. 开发出一套具有可视化界面的程序，把算法程序整合到可视化程序中，便于形象生动的表现动态网络的情况和算法的效果。Java swing实现
5. 使用多类型的实际数据集对算法进行验证实验。在实验中分析团伙的演化现象和规律，得出一些有意义的结论
6. 进度计划

目前开题报告的工作基本接近尾声。接下来的两个月的时间里，主要的工作就是对算法的实现和改进，以及使用实际的数据集进行验证和研究

**3.15-4.1:**

对整个软件进行设计，实现图的描述和基本的设计数据结构主要是动态图的描述。实现betweenness边算法，实现以modules聚类质量度量为依据的自动切边聚类。实现静态部分的可视化软件开发

**4.1-4.15:**

实现二部图的匹配算法，建立模型分析时间切片间的图演化。并使用数据集验证，改进分析的算法。实现动态图分析的软件可视化部分

**4.15-4.30：**

使用数据集进行验证和研究在相关数据集中团伙的演化行为。继续优化算法和软件可视化部分。可扩展研究其他的如core动态分析方法的应用

**5.1-5.20：**

完成毕业论文，并且给出根据实验结果而得到的解释，以及一些算法复杂度分析

1. 参考文献：
2. M. E. J. Newman and M. Girvan  *Finding and evaluating community structure in networks* (PHYSICAL REVIEW 2004)
3. M. E. J. Newman *Modularity and community structure in networks*  (PNAS 2006)
4. Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar *An Event-based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs* (KDD 2007)
5. Gergely Palla, Albert-László Barabási and Tamás Vicsek *Quantifying social group evolution* (Nature, 2007)
6. David Liben-Nowell, Jon Kleinberg *The Link Prediction Problem for Social Networks* (2004)
7. Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan *Clustering Social Networks* (2007)

本科毕业论文（设计）文献综述

动态社会网络团体发现与挖掘算法研究

AnApproach to Discover and AnalyzeCommunities in Dynamic Social Networks

本科毕业论文（设计）外文翻译

An Event-based Framework for Characterizing the Evolutionary Behavior ofInteraction Graphs

一种挖掘动态图演化行为的基于事件的算法框架

Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar

Department of Computer Science and Engineering

Ohio State University

fsrinig@cse.ohio-state.edu

（注解，Interaction Graph的本意是交流图，在文中的定义是基于时序发生变化的图，翻译为动态图更易于理解。Characterize的原意是刻画描写，此处按照数据挖掘的概念惯例可以将其翻译为“挖掘”更易于理解）

## 摘要

动态图在很多领域是非常常见的，例如：生物信息学、社会学、物理学等等。尽管已经有很多关于挖掘网络图的研究，但这些研究大多都基于对静态图的挖掘研究上。我们这个研究将基于动态图在时间线的进化行为，研究图中的个体、团伙以及它们之间的信息流基于时序的变化历程。在这个研究中，我们采用事件刻画的方式来研究随时序变化的动态图的临界状态模型。我们的研究采用这个动态图上的时间切片，切片上的团伙之间不重叠。我们还研究出一种算法框架来发现和获取动态图的事件。我们使用这些基本事件来描述这个动态图中个体和团伙间复杂的行为模式。我们还证明了关于行为模式在模型进化、链接预测、影响最大化模型中的应用。最后，我们将基于我们的算法框架展示一个进化网络的扩散模型。

## 介绍

很多社会网络和生物学系统都可以抽象成一个复杂动态图。个体就是这个动态图中的节点，而个体间的联系就是图的边。对这些图的研究横跨多个科学领域，如计算机科学、物理学、生物学、社会学等。另外，互联网上的社区，例如Flickr、MySpace、Orkut、Email网络、共同作者网络、万维网等都是动态图的很好的例子。我们的这个研究可以提供一个研究它们的结构、性质和行为上的理解模型。

在这个领域上早期的一些研究局限于研究静态的网络，而忽视了真实世界的网络图大多是动态的。事实上，很多这样的动态图都是基于时序变化的，时序变化也包含边和节点的随时间的增加和减少。最近，开始出现一些关于动态图的研究。这些研究有这些主要组成部分：定义动态图的变化行为，描述这些行为，根据前期的变化预测后续的行为，对通用的图进化模型的研究。而这些研究通常是比较有挑战性的。例如，要研究一个快速增长的网络社区，就需要分析海量的时序数据来描述社区的结构、动态活动和演化历程。

自然中的动态图一般来说是可组合的。存在于点与点之间的联系可以用于将点集聚类为团伙。例如，在社会网络中，团伙代表着一群具有某种相似的关联或共同爱好的人。在过去的十数年中，这个问题在静态图上被很多人研究过了。但是，在实际的动态图中，一个固有的特性就是这些团伙不是静态的。它们随着时序时刻都在变化着或者说进化着。我们相信，研究图中的团伙的形成、演化、分解，对研究动态图的演化非常有意义。

另一个方面，网络中节点的行为也是至关重要的。在一个出在演化中的动态图中，节点的行为反应了它所代表的实体的联系模式随时序的变化。节点的运动、行为以及它们对其他节点的影响等信息可以帮助我们预测团伙的演化行为。例如，在一个社会网络中，假如一个人在社交上非常的活跃，那么他/她和其他成员建立新的联系的可能性就很大，也就具有很大的可能性去加入新的团体。在一个描述研究合作者的网络中，如果一个人经常和不同的人合作研究，那么当建立起一个新的合作项目时，包含这个人的可能性就比较高。某个节点的影响力可以具体化为研究它对其他节点的影响。如果在某一个个体加入一个团体时，有很多其他个体也同时加入，那么这个个体就具有比较高的影响力。

对于一个在演化着的动态网络中的传播流或者信息流的研究，对社会学研究、营销网络应用以及传染病学等都有非常重要的意义。例如，流行病毒对社会造成重大的威胁，因为它们很可能会广泛地传播，造成严重而广泛的疾病和伤亡。在营销中，营销的目的是通过联系网推广一种产品或者思想或者技术。研究动态网络的进化以及挖掘出网络中具有强影响力的个体（或团队），在实际应用中有非常重大的意义。例如帮助政府制定有效的政策，帮助传染病学者和防疫部门建立起有效的防疫隔离机制，帮助营销和广告业制定有效的营销手段。

在这篇论文里，我们将介绍一种挖掘演化中的动态网络的框架。我们首先将动态网络按照时序切片成为静态网络。然后我们独立地对各个切片中的网络进行聚类。紧接着，我们采用临界事件描述这些团伙的演化行为。我们将指出一系列高效的算法包括位图矩阵的计算等。我们采用关键事件来研究并发现动态图中行为导向的度量的剧变，这将是一个对于描述动态图中的演化行为的非常新颖有趣的方法。我们使用两个具体的示例来演示我们的挖掘框架——DBLP共同作者网络，一个患者临床实验网络。在每个实例中，我们采用我们提出的算法框架进行挖掘得出的行为模式都帮助我们建立了有用的关于团伙进化以及演化语言的推论。例如在DBLP数据集中，我们使用的未来趋势预测模式就取得了非常成功的结果。在临床实验网络中，我们将展示，使用我们的行为模式可以帮助我们分析发现某一种特定药物引起的肝毒性影响。最后，我们使用行为度量去详细分析一个动态图中的扩散模型，并且证明了这些模型在影响最大化作业中的应用。

简单地说，我们的研究重点如下：

* 发现动态网络中的关键事件
* 用于发现关键事件的有效的增量算法
* 一个基于我们的算法框架的扩散模型
* 事件和行为度量在两个动态网络真实数据集挖掘上的应用，预测行为、趋势及影响最大化。

## 相关工作

关于动态图的兴趣模式挖掘（兴趣模式指社区发现和社区演化），学界早已表现出极大的兴趣。不过，这些相关的研究大多数是对静态图进行研究，挖掘出其中的团伙结构、模式以及突变信息。最近，关于动态图上的团伙的演化行为受到了几个研究团体的关注。Leskovec等人在进化的动态图上的研究，是基于图的拓扑结构的，例如度数(degree)的分布，团伙在整体网络上的性质等等。他们指出一个图的祖辈模型，称作forest fire（森林大火）模型，来描述他们在图的进化行为上的发现。Backstrom等人研究图的结构以及图的进化路径。为了判断个体加入团伙的可能性，他们应用了决策树技术来研究个体和团伙的性质。他们也使用决策树来研究团伙增长的可能性。

Chakrabarti等人对于动态图的研究基于两种应用广泛的聚类算法——K-means聚类和层次聚类法。他们使用有效的聚类方式进行团伙划分，并与前一时间戳的团伙情况进行比较。在获取一个时间片上的聚类的时候，他们也考虑了历史聚类信息，即前面一些切片上的聚类情况。Falkowski等人使用团伙的子团伙的情况来研究一个团伙是活跃的还是稳定的。尽管他们也是研究动态图的，但是他们的研究思路和我们的有很大的区别。他们检测重叠的动态图时间切片，然后使用统计上的度量来描述子团的延续性。我们所关注的重点在于发现动态图中可识别的重要事件和行为模式，建模以及预测未来的发展趋势。在这点上，我们还特别关注了图中的节点并分析他们的进化行为。

Samtaney等人所作的一篇有很大影响力的论文中，描述了在二维或三维的矢量或标量域中提取相关局部用于追踪发展趋势。为了研究这些域在时间线上的进化历程，他们展示了对象的确定性进化事件。基于事件的方法也被应用在坐标数据和聚类流数据上。虽然他们使用了事件分析的方法，不过他们研究的不是动态图。

本体上的语义相似性的应用研究在过去已经被研究过多次。它已经被应用在很多的分类法上，例如Word-Net Gene Ontology。Resnuk提出一种新方法评估基于信息实体概念的语义相似性。在我们的研究中，语义相似性的概念被用于度量个体(数据集中是作者)之间和团伙之间的相似度。

## 问题定义

在详细地介绍我们的基于事件的算法框架之前，请让我们先介绍一下本文中常用的一些标记。如我们之前提到的那样，我们所关注的主要是动态图的演化过程。特别是要研究动态图中的团伙和个体在时间线上的行为模式。为了更好地理解基于时间进化的图，很有必要先研究和描述一下图在时间线上的转换经历。在这点上，我们将使用动态图在不同时间上的时间切片来描述图的静态版本。

定义：当一个动态图的（内部）相互作用关系随时间发生了变化，我们就说这个动态图发生了演化。设G = （V，E）表示一个随时间变化动态图，其中，V为这个图中的点集（实体集），E为这个图的边集（实体间相互作用的集合）。设Gi = （Vi，Ei）为G的一个时间切片，设[Tsi，Tei]是这个时间切片的时间间隔。时间切片的点集和关系集合是图G在这个时间间隔上聚集。

随着图的演化，新的节点和边可能加入到图中。同样的，节点和边也有可能消失。这样的一个在时间维度上具有动态行为的图可以描述为一系列时间切片的集合，其中，每个时间切片具有S equal（间隔等量），不相交（时间片中的团伙不重叠）的性质。



Figure 6：时间切片，从左到右分别是：a)时间切片t=1；b)时间切片t=2；c)时间累积切片t=2

请注意，不同的时间切片是互斥的。它们不包含任何公有信息。这与之前的一些研究不同，那些研究考虑前一个或一些时间片对本时间片的影响。Figure 1(a,b)：使用时间片展示了一个动态图的进化过程。我们可以发现，在时间切片T1中存在与AC之间AD之间的联系在后一时间片中不存在了。Figure 3(c)：展示了两个时间切片的累积值。我们发现在这个图中，刚才发现AC边和AD边消失的信息丢失了，比起前两个切片的表达也没有反映出图的真实情况。为了防止这种信息丢失，我们尽量选取比较小的时间间隔来做时间切片，然后选取比较有代表性的时间切片来研究。时间T下的所有时间切片的集合为S = {S1,S2,…,ST}

要研究图的演化过程，我们还需要表示动态图在不同时间切片间的结构。为了达到这样的效果，我们需要先挖掘出每个时间切片中的团伙。设在时间切片Si中共有ki个团伙，这些团伙的集合为Ci = {Ci1,Ci2,…,Ciki}。对于时间切片Si中的第j个团伙，Cij也可以看做一个图，（Vij, Eij），其中Vij是团伙中的节点集合，Eij是团伙中成员的联系集合（边的集合）。于是，对于每一个Si = （Vi, Ei）来说，Vi1∪Vi2∪…∪Viki = Vi

为了选择一个有效的聚类算法，我们实验了大量的聚类算法，并使用很多的动态图来检测这些聚类算法在团伙聚类效果上的表现。（modularity是度量聚类效果的一种度量值，这里我都翻译成聚类质量）。我们发现MCL算法角力非常好的聚类效果，而且它在不同时间片之间也非常稳定，并且它具有很好的自动性。因此，我们觉得使用MCL算法来对不同的时间片进行聚类。MCL算法不需要指定需要聚类成的团伙数量。相对的，它使用一个颗粒度参数来确定划分的数量。因此，对于每个时间切片来说，这个切片上的聚类效果很大程度上依赖于节点在这个切片上的联系情况。对于我们用于实验的比较稀疏的图来说，我们采用颗粒度1.2来进行聚类。

Algorithm 1展示了我们的算法框架的基本思路。我们遵循增量的策略来挖掘团伙信息，找到动态图随时间演化过程中比较明显的变化(时间片中比较明显的变化)，并和关键事件建立起联系。这些关键事件在后续的研究中将被用于分析更加复杂的行为模式。在本文的第五部分，我们会叙述这些关键事件，以及我们怎样发现这些关键事件。在本文的第六部分，我们挖掘出这些关键事件，并用于分析复杂的行为模式。

|  |
| --- |
| Algorithm 1：Mine-Event(G, T) 算法1：事件发现算法 |
| Input(输入): 动态图G = (V, E)，时间T，时间间隔  首先将动态图G划归为时间切片的表示：  S = {S1,S2,…,ST}  **for** i = 1 to T **do**  Cluster Si  Ci = {Ci1,Ci2,…,Ciki}  **end for**  **for** I = 1 to T – 1 **do**  Events = find\_events(Si, Si+1) //Section 5  Mine Events for complex patterns //Section 6  **end for** |

Algorithm 1

## 数据集

在我们的研究中，我们采用了以下两个数据集：

### DBLP 合作作者网络

DBLP目录库（数据库）中保存了超过80万个计算机科学家的发表的论文。我们选取DBLP的数据中的一些重要会议和研究领域的论文数据来生成了一张合作作者的网络图。我们选取的研究领域主要是数据库技术、数据挖掘技术、人工只能技术领域。我们选取了10年间（1997~2006）横跨这三个领域的28个研究方向的所有论文。我们把这些论文生成了一张网络图，其中，每个作者就是这个图中的一个节点，而作者之间的合作关系就成为这个图中的边。于是，这个横跨十年的图包含了23136个节点，以及54989条边。我们选择时间切片的大小是一年，从而产生了10个时间切片。这些切片就被用作研究关键事件和模式。我们确信，研究DBLP合作作者网络的进化将为我们研究提供有利的信息，帮住我们研究自然状态下的合作关系演化，以及预测作者之间未来可能的合作。

### 临床实验数据

在临床实验中，制药商需要研究某种药物的效果以及药物的副作用，以确保这种药物是否能在可忍受的副作用范围能产生有效的医疗效果。