

**本 科 生 毕 业 设 计**

**开题报告**



**学生姓名: \_\_ \_周晓龙 \_\_\_\_\_\_**

**学生学号: 3071102314**

**指导教师: 何钦铭**

**年级与专业: 计算机科学与技术2007级**

**所在学院: 计算机科学与技术学院**

**一、题目： 基于通话记录的动态社会网络团体发现与分析算法研究**

**二、指导教师对开题报告、外文翻译和中期报告的具体要求：**

开题报告：

1. 介绍动态社会网络团体挖掘的背景。
2. 分析现有相关技术。
3. 明确研究的内容和任务。
4. 提出切实可行的研究方案和计划。

文献综述：

1. 阅读一定数量的文献。
2. 总结现有动态社会网络团体挖掘的常用算法。
3. 分析相关算法的优缺点。

外文翻译：

翻译一篇关于动态社会网络团体挖掘的论文的核心思想和算法，要求翻译准确，语句通顺，结构完整。

**指导教师（签名）**

**年 月 日**

**毕业设计开题报告、外文翻译和中期报告考核**

**导师对开题报告、外文翻译和中期报告评语及成绩评定：**

开题报告内容完整，对动态社会网络的团体挖掘的背景和关键技术介绍清晰，研究内容和任务较明确，研究方案和计划切实可行。在大量阅读相关文献的基础上，文献综述对动态社会网络团体挖掘的关键技术进行了分析，较完整地总结了现有的算法及其优缺点。外文翻译对一篇基于事件的挖掘动态图演化行为的论文进行了翻译，翻译准确到位，条理清晰，结构完整。总体来说，开题报告的三部分结构清晰，内容充实，书写格式一致。

|  |  |  |  |
| --- | --- | --- | --- |
| **成绩比例** | **开题报告**  **占（20%）** | **外文翻译**  **占（10%）** | **中期报告**  **占（10%）** |
| **分 值** | 95 | 90 | 90 |

**导师签名**

**年 月 日**

**答辩小组对开题报告、外文翻译和中期报告评语及成绩评定：**

|  |  |  |  |
| --- | --- | --- | --- |
| **成绩比例** | **开题报告**  **占（20%）** | **外文翻译**  **占（10%）** | **中期报告**  **占（10%）** |
| **分 值** |  |  |  |

**开题报告答辩小组负责人（签名）**

**年 月 日**

目录

[本科毕业设计开题报告 4](#_Toc187892542)

[1. 项目背景 4](#_Toc187892543)

[2. 目标和任务 4](#_Toc187892544)

[3. 可行性分析 4](#_Toc187892545)

[4. 初步技术方案和关键技术考虑 4](#_Toc187892546)

[5. 预期工作结果 4](#_Toc187892547)

[6. 进度计划 4](#_Toc187892548)

[本科毕业设计外文翻译 4](#_Toc187892549)

本科毕业设计开题报告

1. 项目背景

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（如共同的兴趣、共同参与某类事件）使相关人们形成一个团体。研究社会网络中的团体挖掘算法、团伙的进化规律对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。

通常来说，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团体的变化规律，有很好的应用前景，如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，帮助政府部门根据社会的发展制定最优决策；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，更好地服务参与者(如微薄用户、电子商务消费者等)，也能对商业决策提供有力的支持等。

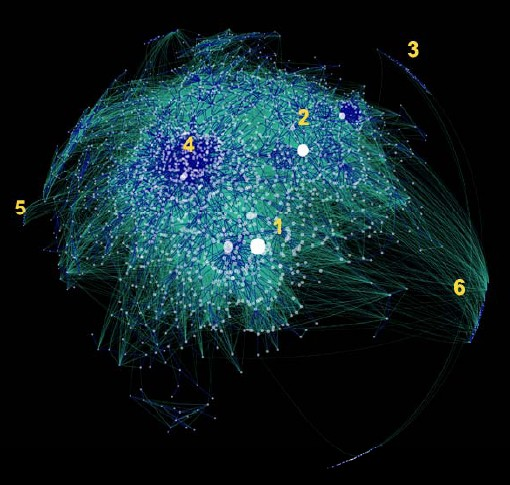


图 1：Blogosphere博客圈数据网络图。社会网络图的一个示例

将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。如图 1所示的就是一个博客社区的关系网络图，每个点代表一个博客，边代表博客文章的引用和评论关系，这个网络的结构随着博客社区的演化发生演化。

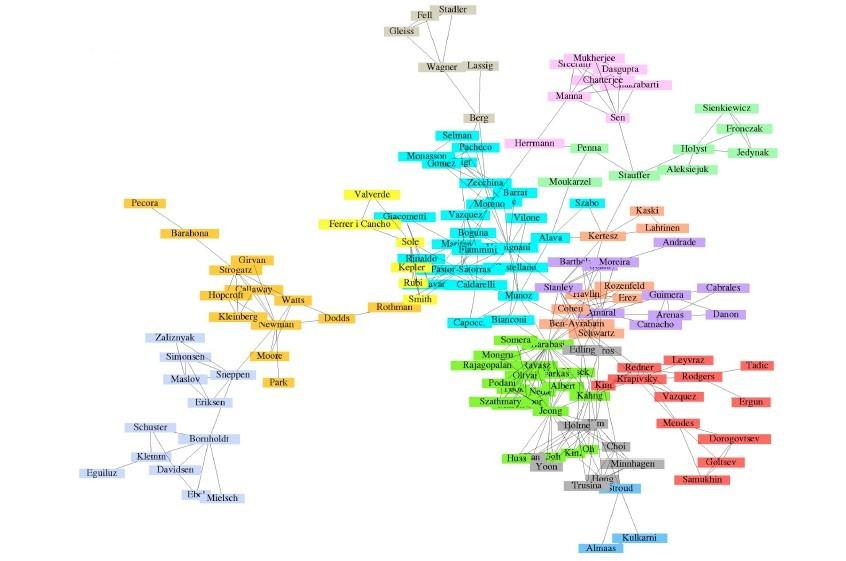


图 2：研究合作作者关系网络，社会网络的一个示例

研究基于这些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。如图 2所示的是一个研究者合作研究的社会网络，节点代表科学家，边代表他们合作参与共同研究的关系。通过一些社会网络挖掘的算法，可以将具有共同兴趣趋向的科学家聚集成团体，图中用不同的颜色代表不同的团体。

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[9]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社会网络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，将为相关领域的应用提供非常大的帮助。

对于动态网络的挖掘研究，近几年来才刚刚起步，但也已经受到相当大的关注。最初的研究来自Hopcroft等人[11]。他们的研究中定义了“自然团”（随时间推移比较稳定的团）的概念，并使用最佳匹配的方式追踪不同时间片的团来分析团的演化现象。Asur等人的研究中[3]，提出了事件（Event）的概念，定义了团伙的延续、合并、分裂、生成、分解事件，并提出了定量的数学模型。Palla提出[4]可以根据团伙在前面几个时间片上的行为预测在以后时间上的行为趋势。

我们的研究数据集对象——通话记录中，包含着非常丰富的人际关系信息。关系紧密（参与共同活动、社团关系等）的人之间有比较频繁的通话联系。对通话记录进行建模，以个人为节点，通话记录（包含频率和时间等）归一化为节点之间边的权值，可以将基于通话记录的人际关系网络抽象为一个社会网络图。随着时间的变化，这个网络图的结构也在发生着变化，这是一个动态网络图。利用这个人际关系网络图，我们可以为公安机关的团伙案件侦破提供有力的线索。我们将研究和应用动态网络的挖掘方法，对基于通话记录的人际关系社会网络进行挖掘，发现有用的团伙的信息和演化规律。

1. 目标和任务

在公安机关在侦破团伙犯罪案件时，往往是先已知一个或几个嫌疑人，然后顺藤摸瓜发现整个团体。利用六度空间理论（任何两个陌生人之间所间隔的人不会超过六个），我们从已知嫌疑人出发，在使用通话记录建模构建的人际关系社会网络上进行有限度数（2~3度比较适宜）的扩展，形成一张和犯罪嫌疑人有关人员的网络图，然后利用切边(edge removal)的聚类算法，排除一些无关人员，可以发现其他也有嫌疑的人员，为案件侦破提供线索。研究网络的动态进化行为，有利于掌握团伙的演化规律，挖掘出更加丰富准确的信息。

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，我们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

我们实现的算法和软件中，将能够：

1. 对通话记录进行有效建模，构建动态人际关系网络图
2. 对这个动态图进行准确的聚类分析，发现出其中的团伙结构
3. 侦测到团伙演化的事件和行为
4. 可视化地展示图的结构和演化行为
5. 可行性分析

本研究的可行性分析包括以下几个方面：

1. **通话数据的可建模性**

通话记录包含以下几个重要信息字段：主叫号码、被叫号码、通话时间（时刻）、通话时长。其中，主叫号码和被叫号码即可对应到个人，可以映射为网络中的一个节点。动态网络往往是根据时间片来描述的，通话记录中的通话时间字段可以用来分割时间片进行聚集。例如，我们在研究中定一个时间片的时间间隔是1个月，那个根据通话时间字段可以分出时间片(yyyyMM)：时间片“201101”、“201102”……在时间片内，两个人之间的通话记录条数即为通话次数，通话次数与通话时长合在一起，可以采用某种模型进行归一化，转化为两个人之间的联系强度，即为图中边的权重。至此，我们的动态网络图已经能成功构建。利用通话数据进行建模是可行的。

1. **动态网路的可挖掘性**

对于前一步已经建模构建的动态网络，在时间片内，我们首先利用六度空间理论从嫌疑点进行扩展，去掉扩展图之外的部分，不用对全网进行挖掘，减少挖掘的计算消耗。接下来，我们将采用Michelle Girvan 和Mark Newman提出的Betweenness切边算法[1]进行聚类分析，挖掘出其中的团伙。我们还将使用Modularity Qualify度量[1]进行聚类效果的评价并进一步实现自动地聚类（基于自动的结果可手动调节）。聚类完成后，我们将采用Asur等人提出的event的概念[3]设计我们的算法侦测团伙的演化事件。事件侦测成功后，我们还可以利用Palla提出的预测思路[4]设计算法，预测团伙的发展情况。因此，对动态网络进行挖掘的研究是可行的。

1. **可视化展现的可实现性**

考虑到个人的技术专长，我将采用Java平台进行算法实现和可视化软件设计。Java是一种很成熟的面向对象程序设计语言，可以无障碍地实现相关数据结构和算法。使用MySql（或其他RDBMS）可以非常方便的管理我们的通话记录和计算产出的数据。Java的JDBC连接可以非常方便地使用MySql数据库。利用丰富的Java Swing图形库，可以实现绘制网络图结构的功能。而JVM的便捷的可移植特性使得我们的程序在任何支持JVM和Java Swing图形扩展的环境中都可以运行。

1. **时间可行性**

我大四秋冬学期就已经和导师讨论和沟通，确定了研究的课题和方向，研究开始得比较早，有充裕的时间完成研究目标和软件设计目标。

1. 初步技术方案和关键技术考虑

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基线，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。



图 3：时间切片的概念，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=1,2. 这是t1时间片和t2时间片的关系叠加

所谓切片，并不是时刻点上的横断面，它是一个小时间区间Δt上的网络关系聚合。例如对于基于通话记录的网络图切片，切片t1是用户1~3月的通话记录建模构建的网络图，切片t2是用户4~6月的通话记录建模构建的网络图。t1与t2的叠加即为根据用户1~6月通话记录建模后构建的网络图。图 3所示就是3个切片的示意图。T1是在时刻t=1时的网络图结构;T2是t=2时刻的结构。T1,2是T1时间片和T2时间片的聚合。

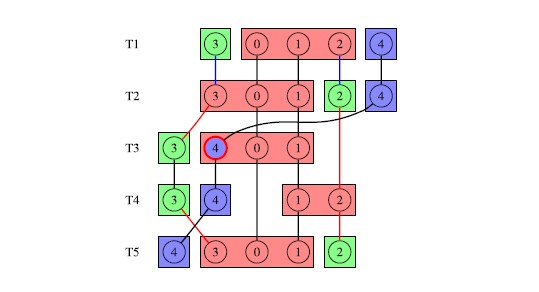


图 4：图中Ti是时间片序列。随着时间的变化，图的结构发生着变化。展示了5个时间片上不同的图结构，以及团伙的保持和进化

随着时间的推移，网络图时刻在发生着演化。如图 4所示，T1~T5五个时间片上，网络的结构各不相同。随着时间的推移，网络中的团伙也在发生演化，图中用相同颜色来标注相同的团伙。

### 单个时间片上的团伙聚集和分析

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[1]。算法基本思路比较简单。统计团伙中两两节点之间的最短路径，边上经过的最短路径条数称作betweenness。Betweenness越大的边上通过的最短路径条数越多。连接团伙间的边承载了团伙间的最短路径流，其betweenness值比较大，可以称为关键路径。通过去掉网络中的关键路径，即可以发现团伙。

在Betweenness切边算法中，其算法复杂度的关键在于发现关键路径的计算上。经典算法中根据betweenness的定义计算图中两两节点间的最短路径并统计每条边的betweenness值。通常，Betweenness计算的时间复杂度为O(n3)，空间复杂度为O(n2)，其中n为网络中节点的个数。我将采用Ulrik Brandes提出的一个优化的betweenness算法[8]，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。

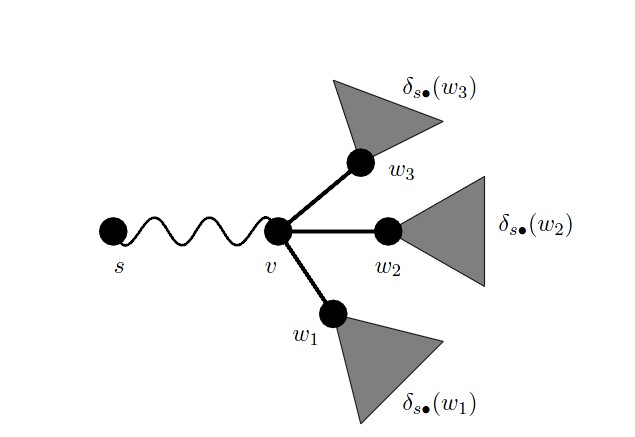


图 5：改进的betweenness算法示意图

在改进的betweenness算法中，Ulrik引入节点权重记录最短路径流，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。如图 5所示，三个阴影的三角区域是已经计算好betweenness的图区域，根据w1,w2,w3上承载的betweenness流量和v-wi边的betweenness可以递归地计算出sv上的最短路径。

### 聚类效果的度量和自动聚类分析

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[1]中就提出一个聚类效果的度量：Modularity Qualify。这个度量通过计算团内联系和团外联系的比值得出一个衡量团伙聚集度的值。这个值在0~1之间，值越大说明团伙的聚集度越高，即聚类的效果越好。通常，在实验中，Modularity的值一般在0.3~0.7之间，非常高的值在实际数据中是比较罕见的。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。

### 时间片之间的动态事件发现

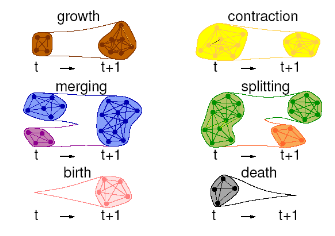


图 6：团伙在时间线上的基本事件

时间维度上的团体演化，可以分为以下几个基本动作，如图 6所示:1. 团伙成长；2. 团伙衰退；3. 团伙合并；4. 团伙分裂；5. 团伙出现；6. 团伙瓦解。对于个人来说，四个基本动作：1.出现;2.消失;1.进入团伙；2.离开团伙[3]

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[7]。简单解释就是在切片t上的团体A0，到了切片t+1上，我们要在聚类出的团体中找到和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[6]进行度量。度量公式为：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[10]来解决二部图匹配问题。

对于个人行为，出现和消失都是比较显而易见的现象，直接查询存储的数据即知。而在时间轴上完成了对团伙的追踪后，只需要检测个人所属的团伙id也可方便地知道个人的进入团伙和离开团伙行为。

研究团伙的活动，对于团伙的出现和团伙的瓦解两个动作也是比较简单的了。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），也是比较简单的。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k。这就有了如k-Merge、k-Split等一系列度量方法，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[3]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. 预期工作结果
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. 基于Java swing技术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。
5. 使用通话记录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, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
3. M.E.J. Newman, From the cover: Modularity and community structure in networks, PNAS 103 (2006) 8577-8582
4. S. Asur, S. Parthasarathy, D. Ucar, An event-based framework for characterizing the evolutionary behavior of interaction graphs. In Proc. of KDD'07, pages 913-921, 2007
5. G. Palla, A.-L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664-667
6. Nina Mishra, Robert Schreiber, Isabelle Stanton, and Robert E. Tarjan Clustering Social Networks, Computer Science, 4863 (2007), 56-67
7. Chayant Tantipathananandh, Tanya Berger-Wolf, David Kempe, A Framework For Community Identification in Dynamic Social Networks. In Proc. of KDD'07, pages 717-726, 2007
8. Tanya Y. Berger-Wolf, Jared Saia, A Framework for Analysis of Dynamic Social Networkss. In Proc. of KDD'06, pages523 – 528, 2006
9. U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001) 163-177
10. Santo Fortunato. Community detection in graphs. Physics Reports 486 (2010) 75-174
11. H. W. Kuhn. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2 (1955) 83–97

本科毕业设计文献综述

1. 摘要

通常来说，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。相对与已经比较成熟的静态网络挖掘研究，动态网络挖掘更加贴近与真实的社会网络，能够为实际应用（刑侦、决策等）提供更为准确的信息，有非常大的研究价值和实际意义。近年来，越来越多的研究者开始关注对动态网络的挖掘。本文将对其中几个重要的研究成果进行描述和分析。

1. 关键词

社会网络、数据挖掘、聚类算法

1. 文献综述

对于动态社会网络团体发现和团伙演化事件分析的算法研究，当前几乎所有的算法都采用先将动态网络按照时间线进行时间切片，再研究各时间片之间的演化过程。就像对影片的剪辑研究，需要先研究影片的各个帧。

将动态网络作时间切片之后，下一步的工作就是需要对各个切片上的团伙进行聚类。当前主流的研究在这点上可以分为两类：一是把每个切片上的网络看作静态的网络，进行聚类，挖掘出其中的团伙[3,4,11]；二是每个切片上对团伙聚类计算都需要考虑到前一切片，甚至前几个切片的状态。动态聚类的研究者认为，使用动态聚类的方法能够得出更为准确的聚类结果，另外利用前面切片的团伙情况可以减少后面切片的计算消耗。使用静态聚类的研究者认为，使用动态聚类并不能取得好的聚类效果。前面切片的聚类结果会影响后面切片的聚类结果，这使得每个切片的结果都引入了“脏”数据。特别是在动态网络发生重大事件的时候，如果受到前一稳定切片的不良影响，这个事件可能会被忽略，或者很难检测到。使用静态聚类的方式独立地分析各个切片是准确的。

### Hopcroft的动态算法

最初的研究来自Hopcroft等人[11]，他们研究了来自NEC CiteSeer数据库的论文引文数据，这个数据集包含1990年~2001年十余年之间的数据。他们使用层次聚类的方式来发现各个时间切片之间的团。在每个时间切片中，HopCroft等人定义了一些“自然团”。所谓自然团，就是在聚类树中，只轻微地受到图的小扰动影响的团。小扰动指在图中移去很小一部分节点。自然团的定义类似于相关研究中提出的稳定团。Hopcroft等人在研究使用最优匹配的方法来追踪不同时间片之间的团。使用这种匹配方式，也可以追踪某个团伙的整个历史演化过程。最优匹配方法也是我的研究中打算采用的方法。Hopcroft的研究中，最主要的缺点就来自使用了层次聚类作为聚类算法。使用层次聚类将产生一个庞大的层次聚类树，这样使得图有非常多的划分方式，这不利于找出最优的划分。

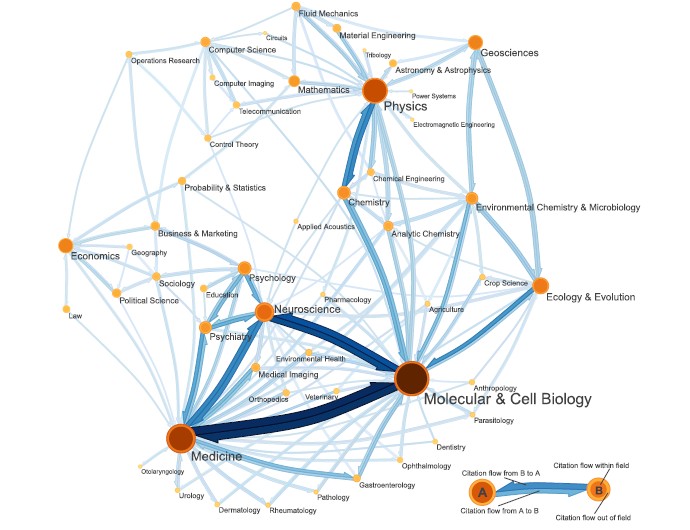


图 7：学科引文关系网络，表示不同学科之间的交叉引用关系。节点表示学科(或研究领域)，边代表在研究中的引用关系，边的粗细代表关联的紧密程度，越粗的关联度越大

图 7展示了各个学科(领域)之间的研究交叉关系网络。这张网络就是根据科学研究的论文引文记录建模构建的。图中节点代表学科(领域)，边代表学科(领域)之间的交叉关联，边越粗代表关联越紧密。

### Palla的动态网络分析和预测算法

Palla等人也研究了一个比较系统的分析动态网络的算法[4]。他们使用以下两个数据集作为研究对象：（1）某电话公司一年中的用户电话通信记录（2）来自Cornell大学图书馆的包含142个月的文章作品合作者数据。Palla等人的研究中采用了一种叫做CPM的方法(Clique Percolation Method)来进行聚类。CPM聚类方式同时考虑两个时间切片的情况进行聚类。研究另一个中心是时间片之间的团伙匹配。一个最直接的想法就是将两个时间片交叉重叠，重叠部分最多的一对团伙，就是两个时间片上最接近的团伙。但是这样做是有问题的，例如在时间片t上，团伙A1={1,2,3,4,5,6,7,8,9,10}，B1={11,12,13}，在时间片t+1上，A2={1,2,3,4,5,6,7 }，B2={8,9,10,11,12}，C2={13}。如果按照重叠最大匹配来匹配团伙，那么B2在时间片t1上重叠最大的是A1，这样子B2会被看做A1的延续，这样是不对的。所以他们将两个切片合并(merge)到一起，提出关联重叠度的定义。不单纯地考察重叠的面积。这个算法其实在早期的集合学中也被用作考量两个集合的相似度：Jaccard系数。如公式(1)所示

公式(1)

可以看出，计算这个度量的值非常快速，基本不需要消耗计算资源。准确、快速使得这个度量成为后来几乎所有研究中公认的一个度量两个切片间团伙相似度的算法。我的研究中也将使用这个度量衡量两个时间切片上的团伙相似度。

这个研究的另一个里程碑是提出预测网络在下一时间的状态的算法。Palla等人利用前几个时间片的演化趋势信息来预测下一时间片可能发生的演化。如图 8所示，Palla以一个节点对团伙内和团伙外的联系权值比来预测节点离开团体的可能性p1。随着节点对外通信比例的增加，这个节点离开团伙的可能性增加，相对的，稳定性就减小。

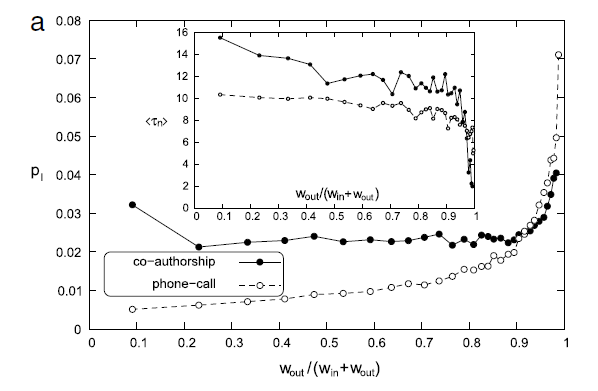


图 8：预测下一时间，某节点离开团伙的可能性指数

同样的，Palla也有研究一个团伙是否会发生分裂（或瓦解）的可能性。如图 9所示，随着一个团伙对外通信比例的上升，这个团伙发生分裂（或瓦解）的可能性上升，相对的，这个团伙的稳定性就下降。

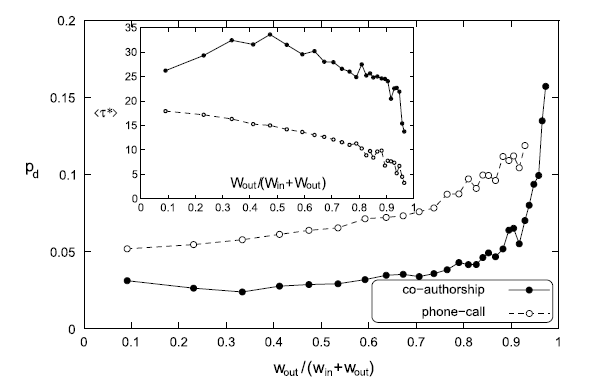


图 9：预测下一时间团伙发生分裂的可能性

### 基于事件(Event)的挖掘算法

能性上升，相对的，这个团伙的稳定性就下降。

Asur等人的研究[3]提出团伙的演化事件定量的定义：

1. 延续Continue：团伙保持其全部的成员，当边的结构可以有所改变
2. 合并k-Merge：时间片t+1中的某团伙保有前一时间片t上的两个团伙的k%的成员。显然，当k=100的时候，两个团完全合并成新团，没有成员逸散。

公式(2)

1. 分裂k-Split：时间片t中的某团伙在后一时间片t+1上分裂成两个团伙，两个团伙的成员总和要占原始团体的k%以上。显然，当k=100的时候，这个团体完全分裂成两个团体，没有成员逸散。

公式(3)

1. 生成Form：在某时间片t+1中出现了团体V，而团体V中的成员在前一时间片t上没有任何两个属于同一个团伙
2. 分解Dissolve：在某一时间片t上的某一团伙V，在下一时间片t+1分解，没有任何2个成员还属于同一时间片

Asur还提出节点在动态图上的活动：

1. 出现Appear：在时间片t上出现了前面时间片中没有的新节点
2. 消失Disappear：在时间片t+1上，曾经在之前的时间片t上有的某节点消失
3. 加入Join：以前不在团伙V中的某节点在这一时间片进入到这个团伙里
4. 离开Leave：以前在团伙V中的某节点在这一时间片离开这个团伙

### Betweenness算法与Modularity度量

对于动态网络挖掘的一个核心问题：聚类算法的选择，我们有必要比较一下各种聚类的思路。聚类算法从过程上看，大致可以分为两类：1)起初是把网络看做一个团，然后依据某些规则将这个起初的大团进行划分，如betweenness切边算法；2)起初把网络中的每个节点看做一个团，然后依据某些规则进行聚合,如层次聚类。两个思路各有优劣，但根据本次研究的应用背景起源来看，选择划分法是显然较优的。在挖掘犯罪团伙的社会关系的应用场景中，往往是先已知犯罪团伙中的一个或多个成员。对全网络的聚类计算复杂度太大。首先会根据六度空间理论以这些已知成员做一次k度的扩展(k<6, 通常取3即可)，对扩展好的网络再进行分析，去掉“无嫌疑”的节点。使用划分算法可以只着眼于包含嫌疑成员的团进行分析，从而大量减少计算复杂度。另外独立的划分算法容易开发成为并行化的计算程序，由于划分聚类在不同团中的计算是相互独立的，所以可以方便将分析放置在多个进程(多处理器或多机)上进行并行计算。

#### Betweenness算法

M. E. J. Newman and M. Girvan在其经典论文[1]中提出经典的betweenness算法。所谓betweenness，就是最短路径betweenness，这是边上的一个权重度量（当然后续的研究者也在节点上设置betweenness度量）。分析网络中所有点两两之间的最短路径，统计通过某边Ei的最短路径条数b，就称作这条边上的betweenness。而最短路径通量——betweenness最高的边就是网络中的关键路径，这些边是连接团伙间的桥梁。将这些关键路径”切”去，就可以发现网络中的团伙。

Betweenness聚类算法的流程如下：

1. 计算图中各边的betweenness
2. 找出betweenness最高的边(如果包含多条相等最大的边就随机选一条)，并在团伙中去掉这条边
3. 重新计算网络的betweenness
4. 从步骤2)循环计算

每次循环切去一条边，切去指定的边数，或者团伙情况达到预期结果就停止计算。

可以看出，本算法的关键在于betweenness的定义和计算上。下面，我们以树形的网络结构为例叙述betweenness的定义，如图 10a)所示。边上的数字就是从s点出发到各点的最短路径条数。对于一些复杂的图，两点间的最短路径可能不止一条，如图 10(b)所示。节点边的数字代表从s点出发到该点的最短路径条数，边上的数字代表该边上的betweenness。对于最短路径有k条的两节点，每条最短路径分流得到1/k的betweenness，叠加计算到通过的边上。

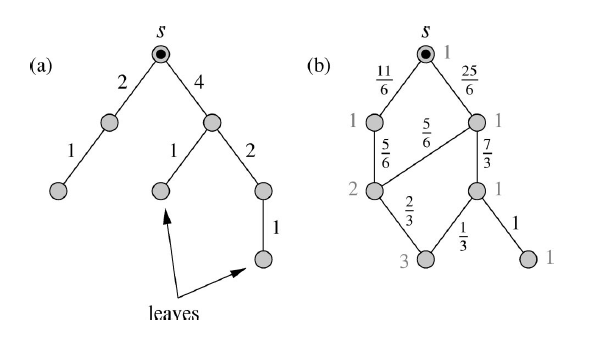


图 10：betwenness计算的例子，从s节点出发到各点的betweenness

假设一个图中有m条边，n个节点。那么从一个节点出发做一次宽度优先遍历发现最短路径的时间复杂性是O(m)。有n个节点就有O(n2)条最短路径。所以计算一次betwenness的时间复杂度是O(mn2)，这也是切去一条边的时间复杂度。

#### Modularity度量算法

本科毕业设计外文翻译