基于观测数据的社会关系网络测度

陈华珊

摘要:在社会网络分析方法中,用于判定社会关系网络的方式可以分为问卷测度法和观测法。其中基于互动或事件观测的数据绝大多数为双模数据结构,这类数据一直以来面对的是小规模群体,并依靠图形化或者描述性的数据分析方式来展示,缺少一种既能适用于大规模群体且基于统计概率估计的综合分析模型。本文介绍采用图形 lasso 模型来分析双模网络数据,并展示了其在大规模群体分析上的优越性和可扩展性。

关键词:社会网络分析　图形 lasso 模型　双模网络　在线社区

一、导言

在社会网络分析方法中,网络数据包含两种类型:针对某一特定总体的所有个体的关系联带,即全体网,以及针对与抽样个体相关联的一组关系联带,即个体网。网络数据可通过抽样调查及问卷访问、档案资料、观测数据、日记、实验记录等等方式获得。社会关系网络测量的关注点在于识别各种类型的社会网络,例如朋友网、照料网、谋职网、借贷网、讨论网等等;以及网络相关指标,包括网络规模、网络密度、网络异质性等等。从数值标度上看,绝大多数采用二分法,即用 0、1 值来表示某类关系联带是否存在;少部分研究采用等级量表以区分关系联带的强度。

在社会学研究中,对社会关系网络的测量最常用的方式为问卷测量法,即由被调查者自我报告有关联的网络对象,其中最为流行的方法为"提名法"(name generator)(Burt, 1984)。尽管问卷测量法通常能够很好的贯彻研究者意图,搜集到足够多的相关变量,但是问卷测量法往往需要调动较多的人力、物力和财力,特别是对全体网的研究,一旦所研究的全体网规模达到一定程度,研究者往往难以用问卷的方式逐个针对网络个体进行访谈。

问卷测量法涉及到的一个基本问题是实际存在的社会关系与行动者所感知的社会关系,即"认知"网络之间的划分(Marsden, 1990)。理查德斯(Richards, 1985)认为使用自我报告的方式来获得网络数据必然假定了某种解释性的或主观性的视角。贝纳德、科尔沃斯以及赛勒等人对此问题进行了长期的对照研究(Bernard and Killworth, 1977; Bernard and Killworth ,1982; Bernard and Killworth , 1984),他们通过问卷访问的方式,让被访者回忆在一个确定的时间段内与他人的交流情况,并与通过观测、日志等方式所得到的观测数据进行比较。他们通过一系列对照研究结果发现,这两种方法得到的结果存在较大的差异。他们甚至得出一个非常悲观的结论,认为"在可接受的精确度内,在任意时间段中,被访者并不十分清楚他们曾经和谁交谈过"(Bernard and Killworth , 1981)。其它类似的研究设计也得到与贝纳德等人基本一致的结果。米拉多对配偶的社会关系网络研究发现,通过问卷访问的提名法所获得的情感交换网络与通过日志法所记录的日常互动网络存在非常大的差异,平均来说,只有 25% 的人名同时出现在两个网络中(Milardo, 1989)。

除此之外,出于实际操作的需要,很多采用问卷测量法的研究对网络规模的上限做了一定的约束,诸如"请列举您最好的三个朋友"(Laumann, 1973),"除您家庭外,和您最亲近的六个人"(Wellman, 1979)之类的问卷问题设计非常流行。

与问卷测量法相比,基于日常互动或事件观测来测度社会关系网络的相关研究在社会学研究中则为数不多,所对应的分析方法也乏善可陈。实际上,社会人类学家是社会关系网络观测法的早期贡献者(Mitchell, 1969; Boissevain, 1974)。基于观测数据的社会关系网络从直觉上来说更为自

然,并且能够提供更佳的描述准确性。其缺点在于需要花费更多的观测时间,并且通常局限于规模相对较小的群体,而难以扩展到更大规模的群体上。

但是,相比于问卷测量法,观测法需要的时间更长,记录的数据更加庞大,因此缺乏有效的记录设备常常也令研究者感到头疼,不得不依赖被调查者的长期个人日志(Wheeler, 1977; Milardo, 1982; Conrath, 1983)。然而,随着现代通讯设备以及电子设备的发展,特别是随着手机通信和互联网社交的普及,研究者可以获得越来越多的互动观测数据,也迫切需要能够用于分析大规模群体的社会网络测度方法。

基于互动或事件观测的网络数据本质上属于双模(two mode)网络数据。双模网络数据在数学上被定义为一个三价体:(A, E, I)。其中一个模表示含有 n 个行动者的集合 $A(a_1, a_j \cdots\cdots, a_n)$,而另一个模表示含有 m 个社会事件(或其他行动者、组织)的集合,$E = (e_i, e_j \cdots\cdots, e_m)$。这两个集合通过一个成员关系连接起来,$I \in A \times E$。当行动者 a_i 是事件 e_j 的参与者时,那么 $(a_i, e_j) \in I$。双模网络数据可用矩阵 $n \times m$ 的二进制矩阵 P 来表示,其中 $p_{ij} = 1 \text{if}(a_i, e_j) \in I$,否则 $p_{ij} = 0$。当 E 表示某种社会组织时,又称为"隶属网"或"成员网"。

布里格(Breiger, 1974)认为,双模网络数据存在一个很重要的结构双重性特征,它既包含行动者之间的关系模式,又包含事件之间的关系模式。也就是说,除了判定网络成员之间是否存在网络连带关系之外,通过区分不同事件的类型,我们还有可能发现不同网络子群体及其对事件的参与偏好。

对于双模网络数据的分析,研究者们提出了各式各样的分析方法。弗里曼(Freeman, 2003)在一项元分析中收集了 21 种分析方法,其中绝大多数是描述法,基于概率统计模型的仅有一种。从近几年社会网络研究的相关文献上看,针对双模网络数据的分析方式以对应分析和主成分分析为主。这两种方法仍然是描述型的统计计算方法。对应分析和主成分分析的优点在于对小群体进行关系网络判定时具有一定的敏感性,并且绝大多数统计软件包提供了方便的二维图展示,有助于直观理解网络结构,但随着网络群体规模的扩大,它们不仅在计算上难以为继,而且难以根据二维结构图获得直观理解。

本文在此提出一种新的分析方法,即用基于统计概率模型的图形 lasso 方法(graphical least absolute shrinkage and selection operator)来解决双模网络数据的估计问题,与其它分析方法进行优劣对比,并将其应用到大规模社会网络数据上。

二、模型介绍

对于一个 $n \times p$ 矩阵 X,n 为观测数,p 为变量数,假设 X 为多元正态分布随机变量,$X = (X_1 \cdots\cdots X_p) \sim N(u, \Sigma)$ 假设其协方差矩阵 Σ 为正定矩阵,那么分布的条件依赖结构可用高斯图模型 $g = (\Gamma, E)$ 来表示,其中 $\Gamma = (1, \cdots\cdots p)$,表示节点集合;而 E 是一个 $\Gamma \times \Gamma$ 的边的集合。在高斯图模型中,节点表示变量,边表示一对变量的条件依赖关系。在控制所有其它变量的情况下,满足 $X_{\Gamma\setminus(a,b)} = (X_k : k \in \Gamma\setminus(a,b))$。两个节点的关系(a,b)出现在边集合 中,当且仅当 X_a 条件依赖于 X_b。对于没有包含在集合 E 中的其它成对变量,意味着在控制所有其它变量的情况下条件独立,在逆协方差矩阵中用 0 表示。在此,对矩阵 X 中节点的两两关系的估计,也称为"邻里选择"(neighborhood selection),其实质是协方差选择问题。邻里选择的目的是对于给定的 n 个 i. i. d 观测 X,分别估计每个变量(节点)的相邻变量。即对于集合 中的一个节点 a ,$a \in \Gamma$,它的邻里变量集合用 X_{nea} 表示,邻里选择的目标是让 X_{nea} 成为 $\Gamma\setminus(a)$ 的一个最小子集,使得对于给定的 X_{nea},X_a 条件独立于所有其它变量。从而邻里选择可以被转化为标准的回归问题并求解,一般采用最大似然法来估计"精度矩阵" Σ^{-1}。用 S 表示 X 的经验协方差矩阵,高斯对数似然的公式表达如下:

$$\log \det\theta - \text{trace}(s\theta) \tag{1}$$

其中 θ 表示逆协方差矩阵，即 $\theta = \Sigma^{-1}$。使公式一最大化可得最大似然估计 $\tilde{\theta} = S - 1$。

矩阵与偏相关系数有如下关系：

$$pij(i,j) = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$$

对于社会关系网络数据，当该偏相关系数矩阵的元素大于0，即表示所对应的两个网络节点之间存在联带关系；反之则不存在联带关系。从而，我们根据观测数据 X，得到对某群体的社会关系网络测度。但是对于社会网络数据，存在两个基本特征：(1)高维性。社会网络数据通常包含大量的节点(变量)，用矩阵表示的话就是变量数 p 大于观测数 n，在此情况下，经验协方差矩阵 S 为奇异矩阵，并不可逆，从而无法估计 θ 矩阵。即使 $p \approx n$，并且 S 不为奇异矩阵，θ 的最大似然估计也会由于过高的方差从而失去效力；(2)稀疏性。用图模型所表示的社会网络数据，存在大量的两两条件独立变量，即 θ 中存在很多0元素。而根据使公式一最大化所估计得到的 θ 一般来说不存在值为0的元素。基于这两种性质，传统的回归方法无法针对这类数据得到一致估计。

近几年来，统计学家针对高维稀疏数据提出了很多解决方案，其中蒂施莱尼所提出的罚似然回归法(Tibshirani, 1996)得到了较广泛应用，并被其它研究者进一步扩展和引进到图模型中(Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Peng and Wang , 2009)。其方法是在公式一中引入一个非负的优化参数 λ，通过对 θ 的加罚，当 λ 足够大时，θ 的一些元素的值将等于零，也就是说 λ 值越大，所估计的逆协方差矩阵越稀疏。并且，即使在 $p > n$ 的情形下，公式仍然能够求解，其表达式如下：

$$\text{maximize}_{\theta}(\log\det_{\theta} - \text{trace}(s\theta) - \lambda \| \theta \|_{i}) \quad (2)$$

其中，$\| \theta \|_{i}$ 为 l_1 罚，表示对矩阵 θ 的所有元素的绝对值求和。针对该最优化问题，研究者提出了不同的求解法。有的学者借用万德伯格等人所提出的“内点搜索法”(interior - point)(Vandenberghe and Boyd , 1998)进行求解(Yuan and Lin, 2007)。贝纳杰等人则提出用“分块坐标递降法”(blockwise coordinate descent approach)来求解(Banerjee and El Ghaoui , 2008)，弗里德曼和哈斯蒂等人在此基础上进一步提出用坐标递降法(coordinate descent procedure)来求解(Friedman and Hastie , 2008)。所有针对此问题的求解法都被称为图形罚极大似然法或图形 lasso 模型(以下简称 glasso)。glasso 模型近年来在基因研究、流行病学等领域有了很多应用研究，并且模型进一步扩展成动态图模型(Ahmed and Xing, 2009; Song and Kolar , 2009)以及分组图模型(Danaher and Wang , 2011; Guo and Levina , 2011)。但在社会学的社会网络研究领域则少见研究者使用。

本文接下来将 glasso 法应用于两个社会学案例中，对双模网络结构数据进行分析，并展示该方法的优点。

三、应用分析

(一) DGG 数据

DGG 数据来自一项在美国南方社区所进行的人类学研究(Davis and Gardner , 1941)。人类学家戴维斯和加纳等人使用访谈、观察记录、访客名单以及报纸来记载社区妇女是否参与以及同哪些成员共同参与社区活动的信息，数据中包括 18 名女性，14 次社会事件。研究者们用他们的人类学观察直觉及洞察力对这些妇女的社会生活进行了描述。他们认为这些妇女可分成两个组，并且在每个组中他们区分出核心成员、主要成员和边缘成员三个层次。他们把编号 1 至编号 8 的妇女分到第一组，其中编号 1、2、3、4 被称为核心成员；编号 5、6、7 为第二层次的成员；编号 8 为第三层次

成员。编号 10 到 18 被归为第二组,其中编号 13、14、15 是核心成员,11、12 为第二层级、编号 10、16、17、18 为第三层级。他们将编号 9 标识为同时属于两个组,且都为第三层次的成员。

表 1　　美国南方妇女社会活动日常参与记录数据

		参与者编码																
社会事件	日期	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15	p16	p17	p18
e1	6 月 27 日	×	×		×														
e2	3 月 2 日	×	×	×															
e3	4 月 12 日	×	×	×	×	×	×												
e4	9 月 26 日	×		×	× ×														
e5	2 月 25 日	×	×	×	×	×	×	×		×									
e6	5 月 19 日	×	×	×	×		×	×	×						×				
e7	3 月 15 日		×	×	×	×		×		×	×		×	×	×				
e8	9 月 16 日	×	×	×	×		×	×	×	×	×	×	×	×		×	×		
e9	4 月 8 日	×		×					×	×	×	×	×	×	×		×	×	×
e10	6 月 10 日										×	×	×	×	×				
e11	2 月 23 日													×	×		×	×	
e12	4 月 7 日								×	×	×	×	×	×					
e13	11 月 21 日									×	×	×							
e14	8 月 3 日									×	×	×							

数据来源:(Davis and Gardner, 1941)。

戴维斯和加纳的这个研究数据成为社会网络分析中的一个经典案例,简称"DGG"数据,并被 UCINet 等社会网络分析软件收为例子。弗里曼(Freeman, 2003)收集了 21 种针对双模数据的分析方法,并应用到 DGG 数据上进行了一项元分析。在 21 种分析方法中,有 6 种(BGR74、BCH78、DOR79、E&B93、FW193 和 FW293)使用代数方法,5 种(P&C72、FR193、FR293、BE197 和 BE297)采用最优分组算法,3 种(BBA75、BCH91 和 ROB00)使用不同方式的本征值解析法,还有两种依靠的是研究者的认知和直觉(DGG41 和 HOM50),只有一种采用的是统计模型(S&F99),即 P^* 模型。

弗里曼将 21 种分析方法所得到的分组结果汇总在下面的表中,其中行表示每个分析方法,用缩写 + 年份的方式表示,列表示每个妇女。每个单元格都用阿拉伯数字分别表示该分析方法判定出的分组类别。如果有的妇女同时归属于两个组别,则用一对数字表示。空白表示该节点无法被纳入分析模型而被删除。我们将 glasso 计算的结果按同样的方式添加在弗里曼的表格之后,用"glasso"表示(见表 2)。

从表 2 可以看出,首先从数据利用率上来说,glasso 法很好地利用了所有 18 个节点信息,无需排除部分节点。而在弗里曼所总结的 21 中分析方法中,有 9 种方法需要先排除部分节点才能进行计算,也就是说它们无法对这些排除的节点进行分组,其中 BCH78 法甚至从 18 个样本中排除了 6 个,导致了大量数据损失。

使用 glasso 法对这 18 位妇女的社会关系网络判定的结果可以看到,区分出三个群体:编号 1 至 7 以及编号 9 的妇女为一组,编号 8 至 16 为一组,编号 17 和 18 为第三组。编号 9 被判定为同时属于两个组,也就是说它承担了网桥的作用,连接两个群体。这两个组的判定结果与 21 种分析方法的绝大多数的判定结果是一致的。稍微有所不同的是,glasso 法单独将编号 17 和编号 18 的妇女判定为第三个组别。这是因为编号 17 和 18 这两位妇女与其他人的联接非常弱,从原始数据上可以看到,她们仅共同出席了两次社会活动。在弗里曼的分析中,BGR74 和 OSB00 这两个方法也都将他们判定为单独的组别,在戴维斯和加纳的人类学分析中,虽然她们与编号 10 - 16 被合为一组,但是她们被判定为边缘成员。由此可见,glasso 法对于小群体估计也是具有敏感性的。

表 2 针对 DGG 数据的 22 种分析方法结果

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	DGG41	1	1	11	1	1	1	1	12	2	2	2	2	2	2	2	2	2	
2	HOM50	1	1	1	1	1	1	1	12			2	2	2	2	2		2	2
3	P&C72	1	1	1	1	11	1	1	1	2	2	2	2	2	3	23	2	3	3
4	BGR74	1	1	1	1	1	1	1		1	2	2	2	2	2	2		2	2
5	BBA75	1	1	11	1	1	1	2	1	2	2	2	2	2	2	2	2	2	
6	BCH78	1	1	1	1	1	1				2	2	2	2	2	2			
7	DOR79	1	1	1	1	1	1	1		1	2	2	2	2	2	2			
8	BCH91	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
9	FRE92	1	1	1	1	1	1	1		1	2	2	2	2	2	2	2		
10	E&B93	1	1	1	1	1	1	1		1	2	2	2	2	2	2			
11	FR193	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
12	FR293	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2
13	FW193	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	12	2	2
14	FW293	1	1	1	1	1	1	1		1	2	2	2	2	2	2		2	2
15	BE197	1	1	1	1	1	1	1		1	2	2	2	2	2	2			
16	BE297	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
17	BE397	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
18	S&F99	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2		2	2
19	ROB00	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
20	OSB00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
21	NEW01	1	1	1	1	1	1	1	2	1	2	2	2	2	2	2	2	2	2
22	GLASSO	1	1	1	1	1	1	1	2	12	2	2	2	2	2	2	2	3	3

除此之外，glasso 法还具有另外一个优点，即通过概率估计，可以根据观测或记录到的频次值得到关系强度的估计，从而我们得到一个无向加权关系网（undirected valued network）。这是上述 21 种计算方法中，除戴维斯和加纳的直觉洞察力法（DGG41）之外所不具备的特性。感兴趣的读者可以根据下述 glasso 法计算的偏相关系数矩阵考察 DGG 数据之中的权力结构，并与戴维斯和加纳等人的核心－边缘划分结果进行比较。

表 3 glasso 法计算的偏相关系数矩阵

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	—	0.054	0.185	0.054	0.000	0.035	0.000	0.000	0.000	0.000	0.000	0.000	-0.083	-0.208	-0.090	0.000	0.000	0.000
2	0.054	—	0.072	0.165	0.000	0.089	0.091	0.000	0.000	0.000	0.000	-0.085	0.000	-0.074	0.000	0.000	0.000	0.000
3	0.185	0.072	—	0.066	0.052	0.033	0.033	0.000	0.061	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.054	0.165	0.066	—	0.113	0.088	0.091	0.000	0.000	0.000	0.000	-0.079	0.000	-0.074	0.000	0.000	0.000	0.000
5	0.000	0.000	0.052	0.113	—	0.000	0.000	0.000	0.000	0.000	0.000	-0.049	0.000	0.000	0.000	0.000	0.000	0.000
6	0.035	0.089	0.033	0.088	0.000	—	0.079	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.000	0.091	0.033	0.091	0.000	0.079	—	0.000	0.101	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	—	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.000	0.000
9	0.000	0.000	0.061	0.000	0.000	0.000	0.101	0.000	—	0.106	0.000	0.000	0.000	0.000	0.000	0.007	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.106	—	0.089	0.000	0.111	0.000	0.033	0.006	0.000	0.000
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.089	—	0.160	0.060	0.000	0.032	0.007	0.000	0.000
12	0.000	-0.085	0.000	-0.079	-0.049	0.000	0.000	0.000	0.000	0.000	0.160	—	0.300	0.000	0.000	0.000	0.000	0.000
13	0.083	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.111	0.061	0.300	—	0.085	0.001	0.000	0.000	0.000
14	0.208	-0.074	0.000	-0.074	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.085	—	0.000	0.000	0.000	0.000
15	0.090	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.032	0.000	0.001	0.000	—	0.000	0.000	0.000
16	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.050	0.007	0.006	0.007	0.000	0.000	0.000	0.000	—	0.000	0.000
17	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	—	0.100
18	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.100	—

(二)在线社区论坛数据

经典的社会网络全体网研究较少涉及较大规模的群体,网络规模通常在几百个节点以内。但随着近年来互联网的发展,针对互联网社会关系网络研究的需求越来越多。前述所例举的21种针对双模数据的分析方法,绝大多数只能适用于小群体的分析,对于大规模群体来说要么根本无能为力,要么需要大量计算资源,对计算机软硬件都提出非常高的要求,甚至超出现有的计算机硬件和软件配置能力。接下来我们将glasso法应用于互联网论坛分析,以展示glasso法在计算上的优越性。

近些年来,有许多研究开始用社会网络的分析方法来研究在线网络社区,比如高校BBS、论坛、博客等等。但在对社会关系网络的测度上,这些研究无一例外地大多延续了经典社会网络分析的频次加权方式:他们都事先设立了一个描述被研究群体关系的二分值矩阵,如果观测到成员x对成员y存在回复帖,就认为二者存在关系,并设值为1;否则设值为0,表示不存在关系,且回复次数越多表示关系权重越高(Goh and Eom , 2006; Gómez and Kaltenbrunner , 2008)。为了避免由此带来的网络密度过高的问题,有的研究者根据矩阵所记录的频次,选择一个阈限值将此矩阵二分化,最后再用社会网络分析方法进行分析(荣波等, 2009)。毫无疑问,这种社会网络测度方式尽管符合直觉,但是存在相当大不足。

首先,就一个网络社区来说,它是个开放和自由的讨论空间,因此从理论上说,一个社区成员可与任何一个成员建立联系。社区用户可以任意发布主题和回复主题,这就导致了社区成员之间相互回复关系的建立十分容易且频繁,因此只要给定相对长的观测期,就会得到大量观测值 。在此情形下,直接使用二分法测度将得到一个联带关系密度非常高的社会网络,给分析带来很大的混淆。而通过累计频次,然后根据某个阈限来确定关系的方式,存在非常大的主观随意性,并且难以比较不同阈限值的结果。为了规避此问题,甚至有的研究者人为选择及缩短观测期,以避免得到过于密集的社会网络。

其次,以频次加权为基础的测度方式隐含一个前提假设,即网络成员的总的行动次数不存在太大的差异。但在网络社区中,其成员的发帖量之间存在非常大的变差。举个例子,假设成员x对z有3次回复,而成员y对z则有5次回复,但x的总回复次数是3次,而y的总回复次数是100次。我们该如何判定或比较x-z和y-z这两对关系呢?显而易见,以频次加权为基础的测度法难以处理此类情形。

对于网络社区社会关系网络的测度,本文将采取一个新的视角。笔者认为,在网络社区中,某个时期内所观测到的发帖和回复是对该网络社区潜在的社会网络结构的反映。在一个主题中进行发言或回复的成员具有相同的兴趣关注点,或者具有一定的网络心理亲密度,甚至可能存在线下现实生活中的互动。也即是说,具有较高关系强度的社区成员,在相同主题中共同出现的概率会更高。因此,通过将网络社区的一个主题当作一次事件(E),每个社区成员的回复看作是对事件的回应或参与(I),我们可以将发帖和回复处理成标准的双模数据结构,并应用glasso法进行网络社区的社会关系网络测度。

本文的数据来自某个城市互联网社区论坛。该论坛具有大约2万名会员,笔者筛选了发帖量在30条以上的900名会员,以及该论坛成立以来的全部主题大约6万条,构成一个60000×900的二维矩阵数据。

通过对该网络数据的glasso建模,我们可以看到:该网络社区的关系网络呈现为一种较为松散的结构,总的联带关系为2719对,网络密度为3.02,这意味着平均每个id存在3个联带关系。进一步通过聚类法对该网络的群组结构进行分析,得到大约30个子群体,因此该网络社区成员存在非常明显的聚集性和区隔性,从而形成不同的讨论圈子。

四、讨论

本文通过一个社会网分析中的经典例子以及一个较大规模的网络社区数据展示了glasso法在

针对双模网络数据进行社会关系网络测度的优越性。首先，glasso 法不仅适用于小群体的网络数据，更适用于大规模的社会网络数据。因为 glasso 法其本质上是回归估计和模型变量选择，统计学家们通过模拟分析已经证明 lasso 模型具有非常好的稳健性，对于几千甚至上万的自变量选择具有一致性（Meinshausen and Bühlmann，2006）。其次，使用 glasso 法进行社会关系网络测度，可以根据无向无价的双模网络数据估计得到无向有价的关系网络矩阵，从而不仅可以进行有无社会关系的判定，还可以进行关系强度的比较，大大丰富了分析内容。

传统社会网络研究一般只分析单个性质的关系网络，比如朋友网、社会支持网等，当需要进行多个性质的关系网络对比研究时则往往缺少合适的分析工具。对高斯图模型的标准估计假设每个观测来自相同分布（i. i. d）。然而，我们有理由假设在很多数据中观测来自不同的分布，例如对某个群体的成员出席会议的观测与对其出席宴会的观测很可是两个独立分布，从而违背了经典假设。针对观测分布的 i. i. d 假设，将 glasso 模型进一步扩展为联合 glasso 模型，我们可以在一个统一的分析框架下考察在同一个群体中多个性质的网络关系如何叠加和扩展。限于篇幅，我们无法在此继续针对联合 glasso 模型及其其它扩展模型进行详细探讨，更多的文献请参阅（Danaher and Wang，2011；Guo and Levina，2011）。

最后，必须指出的是，高斯图模型的假设是数据为正态分布。但是在案例一中，数据来自二项分布，尽管我们直接应用了 glasso 模型并得到了很好的结果，但是针对二项分布数据的估计问题可进一步参考（Banerjee and El Ghaoui，2008）。

参考文献：

荣波、夏正友、朱永真、卜湛，2009，《BBS 在线复杂网络及其成员交互特性研究》，《复杂系统与复杂性科学》第 6 卷第 4 期。

Ahmed, A. and E. P. Xing 2009, "Recovering time - varying networks of dependencies in social and biological studies." *Proceedings of the National Academy of Sciences* 106 (29).

Banerjee, O. and L. El Ghaoui, et al. 2008, "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data." *J. Mach. Learn. Res.* 9.

Bernard, H. R. and P. D. Killworth 1977, "Informant accuracy in social network data II." *Human Communication Research* 4 (1).

Bernard, H. R. and P. D. Killworth, et al. (1981). "Summary of research on informant accuracy in network data, and on the reverse small world problem." *Connections* 4 (2).

Bernard, H. R. and P. D. Killworth, et al. 1982, "Informant accuracy in social - network data V. An experimental attempt to predict actual communication from recall data." *Social Science Research* 11 (1).

Bernard, H. R. and P. Killworth, et al. 1984, "The problem of informant accuracy: The validity of retrospective data." *Annual review of anthropology* 13.

Boissevain, J. F. 1974, *Friends of Friends.* Oxford, UK, Blackwell.

Breiger, R. L. 1974, "The duality of persons and groups." *Social forces* 53 (2).

Burt, R. S. 1984, "Network items and the General Social Survey." *Social Networks*(6).

Conrath, D. W. H. C. 1983, "A comparison of the reliability of questionnaire versus diary data." *Social Networks*(5).

Danaher, P. and P. Wang, et al. ,2011, "The joint graphical lasso for inverse covariance estimation across multiple classes." *Arxiv preprint arXiv*:1111.0324.

Davis, A. and B. B. Gardner, et al. 1941. *Deep south.* Chicago, Univ. of Chicago Press.

Freeman, L. C. 2003, "Finding social groups: A meta - analysis of the southern women data." *Dynamic social network modeling and analysis.*

Friedman, J. and T. Hastie, et al. 2008, "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9 (3).

Goh, K. I. and Y. H. Eom, et al. 2006, "Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions." *Physical Review E* 73 (6).

Gómez, V. and A. Kaltenbrunner, et al. 2008, Statistical analysis of the social network and discussion threads in

slashdot.

Guo, J. and E. Levina, et al. 2011, "Joint estimation of multiple graphical models." *Biometrika* 98 (1).

Laumann, E. O. 1973, Bonds of Pluralism: Form and Substance of Urban Social Networks, John Wiley & Sons Inc.

Marsden, P. V. 1990, "Network Data and Measurement." *Annual Review of Sociology* 16.

Meinshausen, N. and P. Bühlmann 2006, "High - dimensional graphs and variable selection with the lasso." *The Annals of Statistics* 34 (3).

Milardo, R. M. 1982, "Friendship networks in developing relationships: converging and diverging social environments." *Social Psychology Quarterly*(45).

Milardo, R. M. 1989, "Theoretical and methodological issues in the identification of the social networks of spouses." *Journal of Marriage and the Family*.

Mitchell, J. C. 1969, The concept and use of social networks. Social Networks in Urban Situations. J. C. Mitchel. Mancheste, UK, Manchester Univ. Press.

Peng , J. and P. Wang, et al. 2009, "Partial correlation estimation by joint sparse regression models." *Journal of the American Statistical Association* 104 (486).

Richards, W. D. 1985, Data, models, and assumptions in network analysis. Organizational Communication: Traditional Themes and New Directions. R. D. McPhee and P. K. Tompkin. Beverly Hills, Sage.

Song , L. and M. Kolar, et al. 2009, "Time - varying dynamic bayesian networks." *Advances in Neural Information Processing Systems* 22.

Tibshirani, R. 1996, "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B* (*Methodological*).

Vandenberghe, L. and S. Boyd, et al. 1998, "Determinant maximization with linear matrix inequality constraints." *Journal on Matrix Analysis and Applications*(19).

Wellman, B. 1979, "The community question: The intimate networks of East Yorkers." *American journal of Sociology*.

Wheeler, L. N. J. 1977, "Sex differences in social participatio." *Journal of Personality and Social Psychology*(10).

Yuan, M. and Y. Lin 2007, "Model selection and estimation in the Gaussian graphical model." *Biometrika* 94 (1).

作者单位:中国社会科学院社会学研究所、上海大学社会学系

责任编辑:赵联飞

Key Word: Jingpo Marriage Dispute Dispute Settlement Mechanism Intermediation Principle

Abstract: Does film influence people's political socialization? Or does film influence people's political support for specific political authority? As for films with leaders' theme, the empirical research and double variables analysis show that the "watching degree" does influence the samples' political support for specific political authority in reality. Further multivariable analysis indicates that the political support for the specific political authority is not only affected by the films, but also by certain control variables reflecting the samples' peculiarity, which enhance or weaken the political socialization effect of Chinese film medium.

Key Word: Film Political Socialization Films with the Theme of Leaders Empirical Research

Abstract: Among social network analysis methods, there are two methods of measuring social networks: questionnaire and observation. For observation method, the collected data is of "dual - mode" structure which is usually applied with descriptive or graphic analysis for small groups by sociological researchers. The demand for a synthetic probability based statistical method for dual - mode data analysis is urgent. We introduce the graphic model of lasso and apply it into large group social network analysis to demonstrate its advantages and extensibility.

Key Word: Social Network Analysis Graphic Model of Lasso Dual - mode Network Online Community

Abstract: Since the social transformation from the end of 1970's, the importance of emotional attachment among family members decreased sharply. Based on the property inheritance dispute cases, findings show that there are five main value orientations for the current emotional attachment among family members: personal royalty orientation, personal utility orientation, ordinary morality orientation, rights first orientation and value - balance orientation. Therefore, the significance of the emotional attachment among family members has appeared significant structural changes which manifests in three aspects: 1) the pursuit of values has changed from single to multiple significances; 2) the meaning for meeting the emotional needs has changed from being shading to being open; 3) the measurement on the value of emotional attachment among family members has changed from priceless to trade - off.

Key Word: Emotional Attachment among Family Members Significance Value Orientation Needs Meeting Pursuit of Values