## 网络结构，元数据和对于缺失节点、注释的预测

社团检测方法的实验验证常常建立在关于节点的可用的注释，这种方法常被当做是公认的指标关于大尺度的网络结构。更一般的情况，这种注释作为拓扑结构本身的表述的合适性却并没有被评估，同时由于上面的原因，一方面这不可能从根本上区分社团检测算法真正的缺点，另一方面也不可能从根本上区分数据声明自身的不完整性，非精确性，或者自然的结构。在这次的工作中，我们提出了一个原则性的方法同时的去评估这两个方面。我们在数据和元数据之间建立了一个联合的产生模型，同时使用了一个无参的贝叶斯框架从带注释的数据集中推测出它的参数。我们评估元数据的好坏并不是依据跟它直接相关的网络社团，而是根据它预测网络中边位置的能力去评价。我们也展示出了当网络中节点丢失并且只存在元数据的情况下，我们的模型能够预测网络中丢失节点的连接（其实也就是边），即使丢失了元数据也没有关系。通过调查一系列范围很广的数据集，我们展示出虽然在被采用的元数据和被推测的数据群体之间存在很少的精确的约定，不过元数据对于网路的结构是非常的具有信息意义的，能够提高对于丢失节点的预测。这展示出了这种方法发现了一种有意义的方式在数据和元数据之间，而这并不需要或者是要求两者之间存在约定。

**介绍**

复杂网络的结构决定了它的功能，同时也充当了复杂网络演化的机理。然而，常常由于数据量非常大的属性并不能直接的去理解网络数据，这需要通过无参的方法间接的推断出。最显著的例子是在网络中检测社团或者模型的任务，这在最近的几年里有了大量的研究[1-3]。

尽管有这些努力，这任然是一个开放的问题对于如何的刻画大规模的网络结构和如何有效的检测他们在真实的系统中。为了去帮助消除这两者之间的隔阂，很多的研究这已经对比了这些特征，这些特征是从知道了元数据信息或者参考标准中提取出来的，由此推断出和大规模结构相关的指示器(翻译的有点别扭)。然而，这种推断在不严谨的地方往往会被接受，即使当元数据可能包含了一些需要考虑的噪音，它是不完备的和无关要紧的对于网络的结构。由于这些方面，这至今认为搞明白用社团检测的方法得到的元数据和获得的结果之间不符合是否是由于检测方法的无效所引起的，又或者是在元数据和真是的网络结构之间缺少联系所导致的的。

在这次的工作中，我们提出了一个原则性的方法去强调这个问题。我们采取的中心立场是在元数据和数据之间没有采取什么根本性的区别。我们构建了一个产生过程来同时的说明两者。通过推理元数据和数据的的联合模型，我们能够精确的量化和网络结构相关的数据注解的延伸，同时反之亦然。这就是方法之间的不同关于明确的假定元数据（或者是其中的一部分）或者是精确的，或者是粗略的估计对于最好的网络划分。在我们的方法中，如果元数据在网络结构中有信息意义，我们能够决定如何做；如果在两者之间并没有任何的联系存在，这将将会被发现。我们的方法和最近的由Newman和Clauset他们提出的方法是一致的[20]-他们建议使用可用的元数据去指导先验概率在网络的划分中----但是在这里，我们建议一个框架，这个框架更加更具有一般性在下面的三个方面：首先，我们不假定元数据以这样的方式呈现，即它仅仅和所划分出的节点相关。虽然后者可以直接的和常见的社团检测的方法的结果做对比，或者被用作先验在考虑经典的产生模型的时候，但是大部分的数据集都包含了比较丰富的元数据，在这种情况下节点被声明多次，多样化的声明频繁的出现，这样常常会出现很少的节点拥有准确的相同的声明。第二，我们新建了一个无参的贝叶斯推理方法，它并不需要先验信息或者特别的参数需要去指定，正如社团的数量这样的参数一样。第三，我们不仅能够获得建立在统计证据上面的结构和注解之间的相关性，我们能够去评估元数据在预测网络结构方面的能力，而不是评估在隐式划分中仅仅是它们的先关性。这之所以能够实现借力于元数据中可用信息去预测网络中丢失的节点，这或者在本质上是一种启发式的，或者是依赖于很具体的假设关于产生过程的数据[29,30].更进一步的，我们的方法也适合于聚类元数据本身，根据他们在网络中出现将他们划分成等价类。这种网络中元数据的聚类是伴随着网络中数据的聚类同时完成的。它们两者是相互辅助的，因此对于带注解的网络的社团检测的任务提供了一个完全一般化的方法。正如我们展示的那样，这两个特征允许我们去区分信息少的元数据和信息多的元数据，在网络结构方面适用，同时也在预测丢失声明方面适用。

图1:。图代表了数据和元数据联合的模型。数据层是由数据节点组成的，同时被一个邻接矩阵A描述。元数据层是由相同的数据节点和标记节点所组成，被一个二分矩阵T所描述。（最后一句没搞懂）

**2.数据和元数据的联合模型**我们的方法是建立在一个具有代表性的数据和元数据的网络的代表上的。我们假定更一般的情况元数据是离散的，和网络中的节点是任意的连接的。我们这样做是通过将数据和元数据描述在一个单一的图中，图中有2种节点和边类型（或层[31,32]）,正如图1中所展示的那样。第一层对应着网络自身（就是数据），在这张图中一条边连接着两个数据节点，用一个邻接矩阵A描述，当节点i和节点j之间存在边的时候Aij = 1,当两者之间不存在边的时候Aij = 0。该层将会涉及到所有的数据如果元数据被忽略的话。在第二层网络中，数据和元数据都存在，他们之间的联系是被一个二分临接矩阵T所描述，如果节点i被一个元数据j所注释则Tij = 1(以后被称作是一个**标记节点**)，如果一个节点i和元数据j并存在边的联系，则Tij=0.

因此，一个单一的节点有可能和0个，1个，或者多个标记相连接。使用这个一般性的描述我们能够说明一个很宽范围的离散的节点注解。特别的，随着下面一步步的变得透明，我们并不假设单一的元数据标记在实际中和指定的互斥节点团体相关联。

通过使用文献[34]中的度修正对于[8]中的边层的情况，我们使用归纳分层随机块模型（SBM）构建了一个对于矩阵A和矩阵T生成模型 。在这个模型中，节点被分成

Bd和Bt两个组，独自的。对于数据群体r和s之间边的数量使用参数ers表示的（对于r=s的情况将会是2倍的关系）。对于数据群体r和标记群体u将会使用mrs来表示。数据和标记节点都拥有确定的度顺序，{ki}和{di},对于数据和元数据层，分别的，和另外的一组参数集相关。通过在两个独立的层中随机的设置边便产生了一个图，同时有一个联合的似然函数：（公式）这里，b={bi},c={ci}是关于数据和标记节点的群体关系，特别的，θ =({e rs },{k i })和 γ = ({m ru },{d i })是两层之间剩余模型参数的速记。在每一层的内部，对数似然函数是这样的（公式）p(T|b,c,r)的定义也是类似的。因为在两层之间数据节点具有相同的团体关系，这在他们之间提供了一种耦合，同时我们也因此有了一个联合的模型对于数据和元数据之间。**这个模型是一般性的**，因为它能够同时的描述情况：1.在这种情况中在元数据和数据之间存在一个非常好的一致性。(例如：当Bd=Bt并且矩阵mru代表的是一个数据群体和唯一的元数据群体相连接)，当这种一致性不存在的时候，（矩阵T是完全随机的，同时Bt=1）,同时数据和元数据之间任何的详尽的描述都不存在。在原则上，我们可以最大化在Eq.1中的似然估计进而得到参数，然后匹配上述模型。在采用非常一般的假设的前提下，这样做能够发现数据和元数据之间非常精确的关系。然而，为了这个方法能够奏效，我们需要知道群体Bd和群体Bt先前的数量。这是因为似然估计Eq.1是参数化的，（也就是说，它依靠于对于bc, , θ和γ），同时模型的自由度将会随着Bt和Bd增加。随着自由度的增加，似然也会增加，模型的适合性的感知品质也会得到提升。如果我们盲目的遵从这个标准，我们将会把每个节点和元数据标记放在他们单独的群体中，同时我们的矩阵ers和mrs将会恰好邻接矩阵A和T对应，特别的。这是一种极端的过拟合，在这种情况中我们就不能够从真实的结构区别数据中随机的起伏，这种起伏应该是被模型刻画出的。使得模型无参数是处理这种情况的一种比较好的方法，通过包含无信息贝叶斯先验对于模型参数P(b),P(c),P(θ)和P(γ),就像参考文献Ref.[33,36](也可以看附录A).通过最大化无参的似然估计函数P(A,T,b,θ,c,γ) =

P(A,T|b,θ,c,γ)P(b)P(θ)P(c)P(γ)我们能够将节点和数据很好的划分成群体，并且包含群体的数量本身，不存在过拟合。这种情况的发生是因为，在这种设置中，模型的自由度是从一个分布中取样，这在本质上是由于越简单的模型拥有更高的概率，对于更加复杂的模型将会当做是一个有效的惩罚。一个等价的方法去证明这个是将会观测到联合似然将会表达成P(A,T,b,θ,c,γ) = 2 –Σ，这里Σ描述的是数据的描述长度，对应于用来编码模型中参数和参数本身的位的长度。因此，最大化联合贝叶斯似然是和最小化描述长度（MDL）标准是一样的[37,38],这是一种形式化的奥卡姆剃刀，这里根据统计的可用的证据，最简单的假设将会被选择去（其实也就是越简单越好）。

我们需要说明的是这里有一些需要注意的地方当选择上面的先验概率。在缺少先验知识a的时候，最简单的做法是选择扁平的先验去编码这个，这主要是因为对于所有可能的模型具有相同的概率[39].这种抉择，然而，引发了一些限制。尤其，它将会展示出在使用扁平的先验的前提下，使用随机块模型不可能推测出很多群体，展示出超过了一个临界值并伴随着一个比例关系Bmax ~√ N,N是网络中节点的数量[40].此外，扁平的先验（其实就是无变化的意思）对于对于真实的模型来说不可能是一个好的模型，因为它假设所有的参数的值都是等可能的。这是一种极端的随意的形式对于使未知的模型的参数最大化。然而，并不存在数据是从这样最大化随机分布中采用出来的。他们更可能是取样自一些不随机的分布，但是确实一种未知的形状。另外的一种替代的方法是对于先验不急着做决定，而是放在次要的位置上，知道我们观测到数据。通过超验获得先验。当然，这样做，我们就会面临一个相同的问题是选择超验。对于模型来说，我们开始于下面的形式：因为矩阵{ers}和{mrs}是多重图（有Bd,Bd+Bt节点）的临接矩阵，我们从其他的一系列的随机块模型中采样获得他们。根据一个嵌套模型（我感觉应该说的是递归模型），直到达到Bd=Bt=1的简单模型达到，正如文献[33]中所描述的那样。对于剩余的参数，我们从两层贝叶斯模型中选择，对于更高层次得到的参数将会有一个微小的提高。我们总结和回顾了先验概率在附录A中。在这种多层贝叶斯多重模型下，我们不仅能够显著的提高Bmax~N/lnN的精度限制，我们也能够在不同的尺度上提供对数据的一种描述。

这是非常重要的去强调说明我们并不严格的限制去纯粹的协调结构，正像大多数社团检测的文献一样，相反的我们提供了许多连同的方式，这种方式可以被SBM所刻画。正如在介绍中所提到的，我们的方法和最近由Newman和Clauset所提出的参数化的模型方法是不同的，因为在他们的模型中假设一个数据节点只能连接一个元数据标签，同时每个标签都是参数化的单独的个体。在我们的模型中，一个数据节点可以拥有0个，一个，或者更多的数据标签，同时标签能够被聚集成群体。因此我们的方法对于有很多的数据声明是很合适的，在这种情况中，全部的元数据节点的类能够被识别出来。更进一步的，因为他们的方法是参数化的，必须要事先知道群体的数量，而不是从数据中获得，这几乎是不可能的在实际的情况中。另外，当应用在文献[42]中所提出的快速的MCMC算法的时候，推理过程的规模达到了现行的O(N),(或者是对数现行O(Nlog2N)当获得所有的层次的时候)，这里N网络中节点的数量，和群体的数量无关，和文献[20]的置信传播的最大期望形成对比，它达到了O(B2N),这里的B是被推测出的群体的数量。因此，我们的方法不仅对于很大网络有良好的扩展性，同时对于任意的很多的社团也具有良好的扩展性。一个关于我们方法的完善可以当做图工具库可以自由的使用在<http://graph-tool.skewed.de>

这种联合数据和元数据建立模型的方法允许我们在细上明白网络结构和声明式相互关系的，这是在一种平等的方式上进行的。重要的是，我们并没有说明单独的节点作为一种标准标签在社团中，相反的，我们推测出从整个数据中推测出他们和数据群体之间的关系。因为元数据标签自身能够被聚集成群体，我们能够评估它们单独和集体的角色。比如，如果连个标签节点被分到了同一个组，这以为这他们有相似的信息在网络结构中，即使他们的目标节点是不同的，通过依据标签和节点群体的推测出来的概率，我们就获得了关于他们相关性的详细的图。这个可能在原则上（常常在实际中）是偏离普遍的假设即一对一的映射，但是包含它是作为一种特殊的情况。

在进入实验数据集的语意分析之前，我们用一个简单的例子说明一下这个方法的应用,涉及到的网路是美国大学的足球队,边说明的是在一个给定的赛季中两个队之间发生的一场比赛。这个数据当然也是可用的对于队伍属于那个联盟。因为预期的是队伍在相同联盟中相互比赛更多的次数，这被假定成网络结构一种指示器。如果我们匹配上述模型在这个数据集中，节点（队伍）和标记（联盟）被分成Bd = 10和Bt=10个团体，分别的（图2）。一些联盟精确的涉及到了推测出的队的群体，正如开始所期望的那样。尤其在独立性上，这就说明虽然他们是共同的信息性对于网络的结构，单个的他们并不能充当指示器对于整个的网络结构在某种方式上，这种方式指的是能够最终决定网路随机波动。

在图2中，我们使用了在文献[45]中所提出来的联盟的分配，这和在文献[44]中所提出的原始的分配是不同的。由于原来的出版中的一个错误，错误的赛季的信息被采用了而不是像文献[46]中的那样。我们使用这个作为一种机会去展示元数据中的错误和噪声是如何的被评估的在我们的方法中。与此同时我们强调一个重要的应用，即预测丢失的节点，我们以普通的属于描述它，然后回归到我们后来的说明。

1. **预测丢失的节点**

为了预测丢失的节点，我们必须要同时的计算出所有边的似然估计，也就是说。对于一个未被发现的节点I,它涉及到第i行的増广邻接矩阵，ai={A’ij},对于k!=i的时候，A’kj = Akj.如果我们知道了未被发现节点的群体的数量Bi(这里使用的是membership),另外的观测到了节点，丢失边的似然估计函数如下（公式）。这里ˆ θ 和ˆ θ’是对节点划分的兼容性参数。然而，我们不知道a的先验即丢失的节点属于哪个群体。如果我们仅仅有网络数据可用(并不是元数据)，我们唯一的选择就是依据于观测到的划分去决定概率：（公式）。这里（公式）。这就说明我们仅仅可以使用群体大小的分布去引导丢失节点的安置。然而，在实际的情景中，我们可能得到丢失节点的元数据。例如，在一个社交网络中，我们可能知道一个人的社会和地理背景（年龄，性别，国家，等等），这个认识我们想要预测的不知道的熟人。在我们的模型中，这被翻译成知道了在标记-节点图T中的边。在这种情况下，我们能够计算出在数据图中的丢失的边的似然估计如下（公式）。这里，节点的关系的分布通过完全的tag-node图中可用的信息被衡量，（公式），这里ˆ γ 和 ˆ γ ‘是选择性参数，兼容划分c和b.如果元数据和网络的结构紧密相关，上述的分布应该有更大的似然在将丢失的节点放到正确的组中。为了去量化相对预测的提高对于节点i的元数据的信息，我们计算出了预测似然比例λ i ∈ [0,1],（公式）。它应该取λ i > 1/2如果元数据提高了预测的任务，或者λ i < 1/2如果元数据降低了预测的任务。后者可能发生在如果元数据误导了节点的放置（我们下面讨论这正情况在何时发生）。

为了说明这种方法我们回到美国足球的数据，比较原来的和纠正过的联盟划分在预测丢失节点的能力。我们这样做是通过从网路中移除一个节点，推断模型依据更改后的数据，计算出它的似然根据公式5和公式7，我们用得到的数去计算网络中所有节点的预测似然比例（公式）。正如我们在图2c中所看到的，使用元数据显著的提高了预测，我们确实可以看到，通过对比原来不精确的元数据正确的元数据显著的提高了预测。总之，知道了一个足球队属于哪个联盟，确实增加了我们在预测那只球队将会和其比赛的机会，我们使用目前的联盟的划分将会得到一个更高的成功率，而不是使用前一年的数据。这是非常让人吃惊的在这说明性的文本中，但是这种情况将很快变得缺少直观性对于拥有成百上千个node和很多可比较的data tag的数据集的时候，对于这种情况只有自动化的方法比如我们的方法可以处理。

**3实验数据集**

我们进行了一项调查关于有元数据的网络数据集（在附录B中有详细的描述），这里我们移除了很多次被声明节点的一个很小的随机的部分（1%或者是100个节点，看哪一个更更小），同时计算了以上对于每个移除节点的似然比率λ i.对于每个数据集的平均值见图3.我们发现对于大部分的数据集元数据有能力提高对于丢失节点的预测，提高了相对分布的质量。虽然这意味着这里有积极和统计学意义上的关联在元数据和网络的结构方面，但是对于一些数据集这将会导致适度的预测性的提高。在另外的一个方面，这里有少数的情况，在这种情况中包含的元数据使得预测的任务变得更坏，导致了（公式）。在这种情况中，元数据似乎以网络原有连接方式的正交的方式区划分（没搞懂）。为了去说明这种情况，我们考虑了一些人为制造的数据集如下，在这之前回归到实验的数据集。

**A 数据和元数据之间的线性关系**

我们构建了一个含有N个节点的网络，在这个网络中节点被划分成Bd大小的群体，也就是说是极度的相对称的。一个群体的节点之只和相同群体的其他节点相连接。此外，群体中的边E随机的分布,以至于他们有个相同的平均的边的密度。这服从于一个简单的结构，这个结构由不想交的Bd所组成，拥有相似的密度的完全随机的网络。

在元数据层我们拥有相同数量的M=N个元数据标签，这些元数据标签本身也被分成相同的数量Bt=Bd=B的同等大小的群体。

为了去放置Em=E的边在数据和元数据之间，我们也考虑中一种可选择的划分{bi’}，将数据节点划分成B个群体，这和被用于构建网络的原始的划分{bi}不是等价的。一个元数据群体中的tag只能和一个指定的数据群体的节点进行随机的连接，同时反之亦然。也是就是说，这里有一对一的一映射在tag和data群体之间。

总之，我们考虑三个方面对于连接数据和元数据：

1. 和原始数据划分{bi}对其，也就是说，tag-node边连接着相同的数据群体，这个数据群体用来放置node-node的边。
2. 不和数据的划分一致，也就是说，tag-node的边连接的群体有可选择的数据划分{b’i};
3. 随机的：tag-node的边被完全随机的放置，也就是说，既不遵从节点也不遵从标记的划分。

我们强调的是2（不对齐）和3（随机）是不同的：前者涉及到结构化的元数据，和网络的结构并没与相互的关系，后者涉及到了无结构的元数据。换句话说，在不一致的情况下node-tag图并不是完全随机的，因为它连接的是指定的标记群体和制定的数据群体，然而在随机的情况下node-tag边是完全随机的。一个例子就是对于B=2时，每种类型的结构如图4中所展示的那样。

当在以这种方式在人造的网络上执行节点的预测,我们可以发现在有组织的一致的元数据上预测得到提高；然而对于不一致的元数据一个可测量的退化将会被看到。对于随机的元数据中立的值接近<λ> < 1/2,这在图4中可以看到。对于所观测到的退化关于不一致的元数据是由于将数据群体细分成了B更小的子群体，根据他们连接元数据标签的方式。这种子划分，然而，并不是一种有意义的方式关于刻画node-node连接的方式，因为所有的节点都属于同一群体这在统计上是不能够区分的。如果子群体非常的大，这将总是诱导不和谐的噪音进入模型，通过不同数量的发生在每一个子群体中的。因为这种不同的结果仅仅是来自统计的不同，它们是不好的预测因子对于没有被发现的数据，因此造成了预测质量的退化。我们需要说明的是，然而，在这种限制的情况下，在每个子划分的群体中噪音的数量将会变得非常的大，退化消失，因为这些统计的起伏将会变得毫不相关（看图4，曲线N/B=10^3）对于非常的不一致的元数据，推测出的数据群体的总数也会显著的增加Bd = Bd^0 \* Bt,这里，Bd^0被用来产生网络的数据群体的数量。因此，在现实的情节中，采用的结构化的元数据（也就是说不随机的）是和网络的结构非常的不相关的，这些元数据确实使得节点的预测退化。，正如在图3中所看到的少量的几个试验例子。

1. **单个的节点含有的信息量**

在上面平均的似然比率<λ>被用来衡量从网络总移除一个节点， 包含了注释被移除节点的所有的元数据节点的贡献。然而我们的模型也把元数据标签划分成类，这允许我们识别每个节点的预测力根据这个类的划分。通过这个，一个人能够分别有信息价值的节点和无信息价值的节点从一个单一的数据集合中。

我们再一次的量化一个元数据标签的预测力对于预测其他节点将会将会连接到声明该节点的标签上。根据我们的模型，某个数据节点i被标签t声明的概率被下面的公式给出（公式）。它所需要的条件是数据和元数据节点的群关系。类似的，某个数据节点i是被选择的数据节点j的邻居的概率给下面的公式给出（公式）。因此，i作为任意的一个节点j的一个邻居同时它又被标签t所声明的概率是下面的公式（公式）。为了去对比这个分布的预测的质量，我们需要将它和一个空（null）分布做对比，这里tag任意的和node做连接（公式）。

这里的Π(i) = d i /M ,其中(公式)，是节点i被任意标签随机声明的概率。这里在两个分布之间用注解获得的信息将会被KL评估（公式）。这个数量测得信息损失量当我们使用随机的分布Q而不是使用格式化的元数据Pt去刻画可能的邻居，因此我们获得了总量当我们采取相反的做法时候。这个是严格的正的数量，它可能取0到-lnQ\*中的任意可能的值，这里Q\*是最小的Q(i)的非0的值。如果我们用公式15替代公式11和公式12，我们注意到它仅仅依据群体的关系t,同时能够写成下面的形式（公式），这里（公式），是一个节点a属于一个群体u，同时被a的邻居标签所声明的概率（这个标签属于群体r）,对于结构化和随机这两种情况，（公式1）和公式（2），和（公式3）。因为这可以取0到-lnq\*的任何的值，这里q\*是q(u)的最小的非0的值。这在一般的情况下依赖于这个网络中有多少的边，假定q\*>=1/2E.对于不同尺寸的数据集之间的简明的对比，这通常是有用的即考虑一个相对的版本测量，这种测量不依赖于大小。虽然一个选择是通过最大的可能的值去标准化，但是这里我们采用的是q的熵。（公式）和得到预测性Ur关于一个标记群体r(公式)。这给了我们相对的提高关于被声明的格式化的预测。虽然这是不可能的对于Ur>1,这并不是典型的对于很高的信息量的节点，这将意味着一个特殊的不可能的邻居的集合将会变成特殊的可能的对于我们曾经考虑的声明。相反，一个更加典型的拥有很高的信息量的元数据注解会缩小预测对于一个取自q的样本群体的邻居。

使用上述的标准，我们现实的研究了图3中的数据集，量化了注解节点的预测性。正如图5中对于选定一个数据集所展示的那样。总体来说，我们发现数据集有非常大的不同，不仅在对于他们声明的整体的预测性上，而且在它们内在的结构上。典型的，在一个单一的数据集的内部，元数据的预测性是一个很宽的分布。一个很好的例子就是IMDB数据，它描述了演员和电影之间的关系，同时包含了电影的声明涉及到年份和出版的国家，制作者，出版的公司，流派，用户的评分，同时也有用户贡献的词汇（评语）。在图5中我们可以看到很大的一部分声明却拥有很低的预测性（它包含了大量的用户贡献的词汇和评分），然而，这里任然有很多的有意义的声明，他们很具有预测性。最具有预测性的类型的元数据是制片人和导演的联系（比如动漫产品），接下来的是指定的国家（比如新西兰，挪威）和电影的制作年份。