Toward link predictability in complex network

链路可预测性提出的背景

理解现实的网络组织能干什么？

终极目标：理解网络组织的规律性可以提供一个对未知网络同样适用的规律，例如生物学实验中我们观测到的网络只是真实网络很小的一部分，可以节省实验的时间，将精力放在网络物理意义的研究上，还可以研究社交网络的未来走势等等。

理解网络的组织在很多科学分支中是一个长期的挑战。尽管目前为止，一些原理被认为是网络组织形成的**基本**驱动力量：homophily，triadic closure，preferential attachment，reciprocity，social balance。

> 大都是在说现实网络的形成是基于节点之间的相似关系。这些原理是规则的，但只是一部分网络形成的原因

这些原理并不能对所有网络组织提出一个完整的解释。也就是说，链路构造是被规则的和不规则的因素驱动形成的，只有规则的网络才可以建立机械模型

> 对应前面的原理，规则的 == 有原理支撑的、有规律的、*通过部分网络概括整个网络特性的*；
>
> 不规则的==随机的
>
> a. 这就呈现给我们一个问题：如何估计网络哪些部分是可以被归类为规则的。又或者说链路构造在多大程度是可以**被解释的**（可以归纳为一个规律，如前面的homophily）。
>
> 怎么过渡的？（*规则的网络可以通过部分网络来推出整个网路的特性==通过已知网络预测未知网络*）
>
> b. 这个问题又可以想到链路预测问题中，网络中已观测到的链路可以被用来预测未观测到链路存在的可能性。
>
> 那么就可以说，链路构造的可解释程度就与我们预测缺失链路的能力相关，即链路的可预测性。（a等价于b）

一个有效的算法可以为网络组织的相关原理提供强大的依据，一个对网络组织好的猜想也可以转换为一个好的算法。这样看的话，将算法的precision作为网络中链路构造可以被解释的程度。然而这是不行的，同样的网络中，不同的算法所得到的precision是不一样，即特定的算法在同一个网络中会有不同的precision，因此算法的precision不能用来表征网络本身的特征。

> 因为网络本身的特征应该是不变的，不会受外界的影响，*自己理解*

link predictability要解决什么问题，有什么意义

链路构造（link formation）可以在多大程度上被解释，

1. 可以评估一个算法对一个网络效果

2. 估计一个网络可以被预测的范围（extent）

   > 1 2 是不是可以作为一个来说，因为在评估算法对网络的效果之前得知道这个网络在多大程度上可被预测

3. 监控网络的突发的改变
4. 判断一个算法是否已经达到预测的上界，是否还有提升的空间。

在本文之前对link predictability的定义

在本文之前，它通常被定义为一个预测算法可能的最大precision值`《limits of predictability in human mobility》`，作者认为，在这种定义下，real-world network的可预测程度都会达到1，因为他们的缺失链路（missing links）在未观测到的链路中都是可区分的。

> 此处有些不理解：作者给出几个例子
>
> 1. 如果未观测到的链路均相同（此处应该是指任意两个节点未相连的边数相等），因此可预测性为0
> 2. 点可迁图（自同构的，规则的，每个节点连接相同的边数）和（1）一样，未观测到的链路均相同，可预测性为0。基于这个定义，现实网络中都不是节点可达的网络，那么他必然有缺失链路可以被预测，因此可预测性为1。
> 3. 在此基础上，在ER图中完全是随机的，与自同构无关，那么它的可预测性也为1。这就没有意义了。
>
>   什么叫自同构，是一个性质还是每个图都可以求出来？（每个图都可以求出来）
>
>   为什么随机网络的可预测性为1？为什么要给出节点可达链路的例子？是想反证吗？
>
>   （*新的想法*：因为点可迁图是自同构的，因此它的缺失链路不能从未观测到的链路中区分出来。然而现实世界的网络有一部分是规律的一部分是不规律的`《链路预测》`，这样就导致基于这种定义下，将不规律部分变得规律的过程，是的可预测性为1。）
>
>   洛夫熵与Fano不等式的结合来进行刻画。从自同构群的角度来看，待预测的链路（真实存在但是不在观测集合中）与实际不存在的链路总是可以区分的，所以一个真实网络的链路可预测性，原则上应该都是1——这种上界是没有价值的。
>
>   什么叫自同构群？图的所有自同构组成的集合

怎么理解**规则网络**：在图论中又很严格的定义，指每个结点度都相同的网络。

在此基础上，当每个结点地位都相同时（自己理解为不考虑结点的属性），即对每个结点都有一个自同构映射，则该图称为**点可迁图**。`《链路预测》`

这种定义是没有参考价值的，因为所有现实网络的可预测性原则上都应该是1。

本文作者对link predictability的定义

因此从网络本身的结构出发，本文中作者给出的link predictability是表征预测的本身难度，而不是取决于某个算法。在这种定义下的可预测性可以作为网络内在的性质。

Hypothesis

在这个定义下，作者提出假设：如果缺失链路加入网络后对网络结构产生巨大的改变，那么该网络中的缺失链路很难被预测；相反的，如果缺失链路从网络中移去或者加入网络后，未对网络结构造成很大的改变，那么缺失链路是可以被预测的。

为什么要用这种方法？怎么过度过来的？

> *自己的理解*：出于对网络本身的定义，整个过程要在网络的自身结构上进行，没有对网络结构先验知识或者特定的结点（边）属性。既然要表征网络的可被预测的程度，就从预测的角度出发，同时还要度量网络的规律性，就想到从网络中取出一部分或者加入一部分缺失的链路来看这个网络结构是否和之前有差异，如何度量这种差异？

进一步的，提出**基于*一阶微扰*矩阵的"结构一致性"**指标。

> *自己的理解*：网络结构如果是规则的，前面也说到，点可迁图是规律的，那么在完整规则图上如果去除一组链路，或者在缺失链路的网路中加入缺失的链路均使得链路的结构。
>
> 问题：为什么基于一阶扰动的结构一致性指标可以表示结构性的变化？（特征向量可以很好的反映网络的结构特征，*Algebaric Graph Theory*，why?）

目的：**为了量化出假设中的变化，要计算出发生变化的数值，通过一阶微扰法计算。**

建立模型

给出无向图$G<V,E>$，其中V为节点集合，E为边的集合。选取E中占p^H的边组成ΔE，作为微扰集，剩下的边组成E^R，并求出其对应的邻接矩阵ΔA与A^R。其中A^R是实对称的，因此他可以对角化处理

$$A^R=\sum_{k=1}^{N}\lambda_k x_k x_k^T,$$

其中，λ为A^R的特征值，x为A^R的特征向量。

> 此处略去微扰的计算过程。（微扰理论：*从相关问题的确切解中找出问题的近似解的数学方法*。基于这个定义：**可以将作者给出的这种方法求得的矩阵$\tilde{A}$看作对矩阵A（矩阵A为网络的邻接矩阵）近似解。**）

通过一种*微扰*（**待解决**）的方法求的A^R的近似矩阵如下：

$$\tilde{A}=\sum_{k=1}^{N}(\lambda_k+\Delta\lambda_k)x_k x_k^T,$$

这个公式可以看做对邻接矩阵A做**线性近似**，作者认为：如果扰动没有明显改变网络G的结构特征，那么矩阵A^R（A^R为E^R的邻接矩阵）的特征向量与矩阵$A^R+\Delta A$的特征向量几乎一样。（对此处特征向量不变的理解：假设微扰没有改变结构特征，那么特征向量几乎不变，求得围绕后的特征值计算围绕矩阵，比较前后两邻接矩阵的差异。如果微扰改变了结构特征，那么最后得到的矩阵必然和原矩阵差异很大。那么如何很好的表示出这个差异？）

> 不能用$\tilde{A}-A$后的矩阵值来表示两者的差异性，最后的指标结果因该是一种数值型的表示而不是通过观察矩阵的差值来表征，应给出计算机可以理解的指标。

提出**"结构一致性"**指标来量化这个差异。

通过$\tilde{A}-A^R$可以得到在缺失链路作为微扰项后，除去原本已存在的链路，而对其他未观测到的链路（原来的缺失链路是不可知的）产生的影响，值越大，表示影响越大，越可以作为缺失链路。

结构一致性σ_c的定义如下：

$$\sigma_c=\frac{|\Delta E\cap E^L|}{|\Delta E|}$$

E^L为邻接矩阵$\tilde{A}-A^R$中前L（$L=|\Delta E|$）个链路的集合，ΔE为缺失链路的集合。

> 目前遇到的问题：
>
> 1. 如果将可预测性作为网络本身的性质，那么根据目前对结构一致性的计算发现，ΔE是否是重要的，不同的ΔE取法是否会得到不同的结果。（*自己理解*：如果网络本身是规则的，那么不论ΔE如何取最终的结果必然都是相近的。）
> 2. 如果网络有一部分是不规则的，而大部分是规则的，如果ΔE从规则的网络中取出来进行扰动，是否得到的结果会判断网络为规则的，又或者ΔE的扰动是全局性的，只要有部分是不规则的那么最终的σ_c会低一些。

σ_c的算法

1. 给定一个网络G，随机选择其中p^H的边作为扰动边集ΔE，剩下的边组成E^R，并求出其对应的邻接矩阵ΔA与A^R。
2. 计算A^R的特征值λ与特征向量x。
3. 使用如下公式求得$\Delta\lambda$

$$\Delta\lambda_k\approx\frac{x_k^T\Delta A x_k}{x_k^T x_k}.$$

4. 使用如下公式求得$\tilde{A}$

$$\tilde{A}=\sum_{k=1}^{N}(\lambda_k+\Delta\lambda_k)x_k x_k^T,$$

5. 根据给定矩阵$\tilde{A}$内的值对网络G中未观测到的链路进行排名。
6. 选择前L个未观测到的边，其中$L=|\Delta E|$，观察有多边是出现在扰动集ΔE中的，这些边在ΔE中所占比例即为σ_c的值。

> 至此，可以得到一种数值型的指标来表示扰动后的网络与扰动前的网路的差异性。但需要实验来验证。

SPM算法

上述给出的“结构一致性”思想还可以用作链路预测中

实验验证

对规则网络的验证

对随机网络的验证

####

算法比较

基于相似度的

CN

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|,$$

AA

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|},$$

RA

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|},$$

Katz

$$s_{xy}^{Katz} = \alpha A_{xy} + \alpha^2 A_{xy}^2 + \alpha^3 A_{xy}^3 + \cdots,$$

基于先验结构的

HSM

> 假设现实世界的网络是可以分层级的，因此网络中的结点可以被划分到多个群落中，在这个基础上进一步的划分到更小的子群中。

SBM

> 最常用的网络模型之一。结点划分到各个群落中，则两个结点相互连接的概率取决于他们所属的群落。