摘要

自然界存在很多复杂的动力学系统，这些动力学系统可被建模为一个复杂网络。通过对网络拓扑结构的深入研究，人们发现动力学系统产生的各类外部观测数据，与特定拓扑信息具有广泛的关联性，这使得过去许多只能从外部观测得到的数据在数理上得到了合理解释，并由此提出一个重要的科学问题：产生某种外部观测数据的网络拓扑结构应该是什么样的？复杂网络重构发展到现在, 已经取得了一些成就，如：动力学网络重构、多层网络的重构、时变网络的重构等。

从分析研究动力学数据出发，揭示网络结构对我们理解、预测和控制实际系统功能极其重要。基于完整数据的网络重构已有大量研究，然而现实生活中最常见的困难之一就是有效数据缺失，即获取的数据不完整，只能观测到边际数据。例如，由于银行、贸易系统具有一定的隐私性，无法得知各个银行之间具体的交易量或各个国家间的贸易信息，只能获取银行的总交易量及国家的总贸易额。为解决这类问题，我们考虑通过 Adaptive Signal Lasso 和 logistic 回归模型，在边际数据（不满足数据完整假设）情况下进行网络预测。此外，在连续时间步下观测到的数据，噪声影响远超于离散型数据，因此对于这类观测数据，我们希望通过离散化响应变量的方式，通过 logistic 回归模型进行网络重构。

本文的内容主要研究如下：1、研究 Adaptive Signal Lasso 和 logistic 回归模型在只有各个国家总进出口贸易额（即边际数据）情况下，重构世界贸易网络（WTW）的能力，并对两种方法的应用能力进行对比。2、研究 Adaptive Signal Lasso 和 logistic 回归模型在现实网络：社交和生物网络下的应用能力 3、通过 Kuromoto 同步动力学以及演化博弈生成动力学数据，研究当不满足数据完整假设时 Adaptive Signal Lasso 的重构能力以及对响应变量进行离散化处理后 logistic 回归模型的预测能力。研究发现在只有边际数据情况下，Adaptive Signal Lasso 和 logistic 回归模型重构世界贸易网 WTW 的精度都较高，且 Adaptive Signal Lasso 与 Lasso、Adaptive Lasso、Signal Lasso 三种 Lasso 型方法以及 CL、DBCM、DECM、

CREM 四种现有的基于概率的连边预测方法对比发现，Adaptive Signal Lasso 方法性能最好，应用能力最强。在 Kuromoto 同步动力学和演化博弈两个模拟实验下，Adaptive Signal Lasso 的重构效果仍然不错。对响应变量进行离散化处理后，logistic 回归模型重构网络精度较高。此外，研究发现，不管是世界贸易网络还是模拟实验和实证网络，在只有边际数据的情况下，Adaptive Signal Lasso 方法性能皆优于 logistic 回归模型。

关键字：网络重构；边际数据；动力学系统；Adaptive Signal Lasso；logistic 回归

Network Reconstruction and Edge Probability Prediction Based on Complex System Dynamics

Wang HanKun
Probability Theory and Mathematical Statistics
Directed by Shi Lei

There are many complex dynamic systems in nature, which can be modeled as a complex network. Through in-depth research on network topology, it has been found that various external observation data generated by dynamic systems have a wide correlation with specific topology information. This has provided a reasonable mathematical explanation for many data that could only be obtained from external observations in the past, and thus raises an important scientific question: what should be the network topology structure that produces a certain external observation data? The development of complex network reconstruction has achieved some achievements, such as dynamic network reconstruction, multi-layer network reconstruction, time-varying network reconstruction, and so on.

Starting from analyzing and studying dynamic data, revealing the network structure is extremely important for us to understand, predict, and control the actual system functions. There has been a lot of research on network reconstruction based on complete data, but one of the most common difficulties in real life is the lack of effective data, which means that the obtained data is incomplete and only marginal data can be observed. For example, due to the privacy of banks and trade systems, it is impossible to obtain specific transaction volumes between banks or trade information between countries. Only the total transaction volume of banks and the total trade volume of countries can be obtained.To solve such problems, we consider using Adaptive Signal Lasso and logistic regression models to make network predictions on marginal data (which does not satisfy the assumption of complete data). In addition, the impact of noise on data observed in continuous time steps far exceeds that of discrete data. Therefore, for such observation data, we hope to reconstruct the network through a logistic regression model by discretizing the response variables.

The main content of this article is as follows: 1. Study the ability of Adaptive Signal Lasso and logistic regression models to reconstruct the World Trade Network (WTW) with only the total import and export trade volume of each country (i.e. marginal data), and compare the application capabilities of the two methods. 2. Study the application ability of Adaptive Signal Lasso and logistic regression models in real networks: social and biological networks. 3. Generate dynamic data through Kuromoto synchronization dynamics and evolutionary games, and study the reconstruction ability of Adaptive Signal Lasso and the predictive ability of logistic regression models after discretizing response variables when the assumption of data completeness is not met.Research has found that in the case of only marginal data, Adaptive Signal Lasso and logistic regression models have higher accuracy in reconstructing the World Trade Network (WTW). Moreover, Adaptive Signal Lasso has the best performance and strongest application ability compared to three Lasso type methods: Lasso, Adaptive Lasso, and Signal Lasso, as well as four existing probability based edge prediction methods: CL, DBCM, DECM, and CREM. In two simulation experiments of Kuromoto synchronous dynamics and evolutionary game theory, the reconstruction effect of Adaptive Signal Lasso is still good, and after discretizing the response variables, the logistic regression model has a high accuracy in reconstructing the network. In addition, research has found that whether it is the world trade network or simulation experiments and empirical networks, the Adaptive Signal Lasso method performs better than the logistic regression model in the case of only marginal data.

Key words: Network reconstruction; Marginal data; Dynamic system；Adaptive Signal Lasso; logistic regression;

目录

第一章 引言 …… 1
第一节 研究背景 …… 1
第二节 国内外研究现状 …… 2
第三节 研究意义与创新点 …… 5
第四节 章节安排 …… 5
第二章 预备知识与准备工作 …… 7
第一节 复杂网络 …… 7
一、 复杂网络的统计特征 …… 7
（一） 度与度分布 …… 7
（二） 同质性和异质性 …… 8
（三） 中心性 …… 9
二、 复杂网络的基本网络模型 …… 10
三、 网络重构精度指标 …… 11
（一） TPR、TNR、FPR …… 12
（二） 准确率 ACC …… 12
（三） 精确率 PCR …… 12
（四） F1 值 …… 13
第二节 网络动力学 …… 13
一、 网络同步动力学 …… 13
二、 网络演化博弈动力学 …… 14
第三节 网络重构 …… 15
一、 线性重构方法 …… 15
（一） 线性重构原理与动力学关联 …… 15
（二） Lasso 族改进方法 …… 17
二、 其他基于概率的重构方法 …… 19
第三章 基于真实贸易网络的连边预测 …… 23
第一节 背景介绍 …… 23

第二节 基于 Adaptive Signal Lasso 的重构方法 …… 23
一、 Adaptive Signal Lasso 方法介绍 …… 23
二、 重构模型介绍 …… 26
三、 基于 Adaptive Signal Lasso 的重构结果 …… 26
（一） Adaptive Signal Lasso 与其他三种 Lasso 型方法的对比 …… 26
（二） Adaptive Signal Lasso 与其他基于概率的重构方法的对比 …… 27
第三节 基于 logistic 回归模型的连边推断方法 …… 28
一、 logistic 方法介绍 …… 28
二、 重构模型介绍 …… 29
三、 数据处理 …… 31
四、 基于 logistic 回归模型的重构结果 …… 33
第四节 Adaptive Signal Lasso 与 logistic 回归模型的对比 …… 34
第五节 实证网络分析 …… 35
一、 实证网络信息介绍 …… 35
二、 实证网络重构实验结果 …… 36
第四章 基于动力学数据的网络重构模拟实验 …… 38
第一节 Kuromoto 同步动力学 …… 38
一、 边际信息与数据生成 …… 38
二、 实验结果 …… 39
第二节 演化博弈动力学 …… 48
一、 边际信息与数据生成 …… 48
二、 实验结果 …… 50
第三节 稳健性分析 …… 60
第五章 总结 …… 62
参考文献 …… 69
附录 …… 70
致谢 …… 71
攻读硕士学位期间取得的学术成果 …… 72

第一章　引言

第一节　　研究背景

现实世界中存在大量相互关联的系统，例如人类社会中的人际关系和社会组织形式，又如工业领域，存在高度耦合的集成控制系统，再比如生物领域，宏观上存在依托生物链而存在的生态互惠系统，微观上存在功能彼此协调的生命功能系统。借助数字信息高度发达的时代，复杂系统产生了海量的动力学数据，为人类社会中各类现象的分析、解释、利用提供了丰富的数据资源。自 20 世纪 60 年代，科学家们基于复杂系统研究开始关注两个问题：（1）复杂系统如何运转；（2）复杂系统运转背后的数理机制是什么。解释这两个关键科学问题是困难的，最初人们往往需要付出高额成本描述一个复杂系统，因为一个高度关联的系统是一个黑箱，人们只能从外部观测中获取关于该系统的一切信息。具体的，如物理系统、工业系统、电器系统等可以通过各类传感器实现数据刻画。抽象的，如社交圈、社会舆论系统、媒体系统等可以通过社会调查和每日手工记录信息等方式实现刻画。此类描述系统的方式存在以下问题：（1）数据过度粗糙，无法提供解释系统所需要的信息。（2）由于主观偏见导致数据记录失准。因此急需一种可用于进行普适性的细粒度复杂系统建模的数理范式。

得益于复杂网络理论的发展，人们通过数学建模刻画动力学系统黑箱。通过对网络拓扑结构的深入研究，人们发现复杂系统产生的各类外部观测数据，与特定拓扑信息具有广泛的关联性，即网络拓扑性质能够模拟甚至准确描述某个特定系统的外部观测。这使得过去许多只能从外部观测得到的数据在数理上得到了合理解释。基于上述观点，学界提出一个重要的科学问题：产生某种外部观测数据的网络拓扑结构应该被如何描述？此类复杂网络重构问题，通常被视为是一种逆向识别工程。现实中拓扑信息的识别对理解、保护及合理利用网络功能具有重要意义。然而复杂网络重构是困难的。因为，（1）复杂网络的结构复杂多样：复杂网络往往具有非线性、非均匀的特点，其结构不仅包含丰富的拓扑特征，还可能存在动态变化。（2）信息非完整性：现实中的网络数据通常是通过有限的观测和采样得到的，存在不确定性和噪声，这给网络重构带来了挑战，此处引入“边际数据”的概念，边际数据（Marginal Data）是指由于数据获取受限或

者由于数据敏感性等原因，而导致数据获取不完整、局部或片段性的数据。（3）计算复杂性：实践中随着网络规模的增加，网络重构的计算复杂度呈指数增长，从而需要大量的计算成本和时间。现实生活中的银行和贸易系统，由于隐私性只能观测到各个银行的总交易量和各国家的总贸易额，因此边际数据在复杂系统研究中具有重要意义，我们有必要合理利用边际数据了解部分特征和行为，并通过合理方式降低噪声对网络重构带来的干扰。

学者为解决上述网络重构相关问题，提出了基于信号推断的线性统计模型。通常，线性模型将拓扑连边存在与否视为待估参数（信号），该参数估计是二元（binary）的。当系统中各节点的交互数据已知时，可以采用该线性模型进行连边推断。而当系统仅可观测边际数据时，线性模型重构精度不佳。为合理利用边际数据，学者提出基于概率预测的方法解决连边推断问题。因此，连边概率推断模型假设连边信号推断是定义在概率测度下，此方法对敏感数据或边际类数据已知时具有良好的解析作用。然而基于连边概率的推断模型间接导致模糊拓扑信号的产生，即概率处于 0.5 的邻域内时，无法判定此类信号的具体含义。

本文主要讨论边际数据已知情形下的复杂网络重构问题。本文的研究贡献主要是：（1）研究 Adaptive Signal Lasso 基于边际数据进行二元信号推断的能力；（2）通过应用 logistic 回归模型消除连续数据的噪声干扰并进行二元信号预测；（3）与当前各类基准方法进行对比实验，并分析 Adaptive Signal Lasso 和 logistic 回归模型之间的重构能力差异，并从网络性质、稳健性等角度分析了影响重构的因素。

第二节　国内外研究现状

近年来，随着复杂网络在数学、物理、经济等学科中的深入发展，基于网络功能和拓扑优良性质的丰富研究成果已经在交通电力(Pagani 和 Aiello, 2013; Zhao et al., 2005)、经济金融(Arthur, 2018; Battiston et al., 2016)、互联网安全系统(Artime et al., 2024; Wang et al., 2014)、生物医学(Deng et al., 2012; Schwikowski et al., 2000) 等诸多领域得到了广泛应用，甚至在人工智能领域，对神经网络算法的发展都产生了巨大的影响(Chen 和 Billings, 1992)。通过复杂网络提供的优良动力学特性，人们得以对抽象的动力学现象做出解释，并加以模拟和分析。然而我们仍对现实世界中各色动力学系统内部的运行机制和演化规律知之甚少，有

证据表明(Gómez-Gardenes et al., 2011; Pastor-Satorras 和 Vespignani, 2001; Santos 和 Pacheco, 2005)，网络拓扑的主要属性实际上对网络的动态设置和控制起着至关重要的作用。因此，揭示复杂网络内部各单元的作用机制和受影响的结构至关重要。

受限于动力学特征的不确定性和系统的高度非线性(Luenberger, 1979)，早期对复杂动力学系统产生的数据无法较好的进行建模。因此，以探究数据背后复杂系统为目的的众多研究涌现出来，主要基于不同动力学机制会产生不同的数据动态为指导思想(Brin 和 Stuck, 2002)，比如（1）基于数据推断的连续时间动力学系统(Goebel et al., 2009)；（2）基于数据驱动推断抽象的动力学系统(Runge et al., 2019)；（3）从数据随机性中获取非线性系统中的有用知识(Brückner et al., 2020)。尽管基于数据驱动的方式能够粗糙的概括某一系统的大概性质，但不能更准确的描述系统的内在交互机制，这对揭示复杂系统功能提出了挑战。20 世纪 60 年代，网络科学兴起，众多基于微分方程的潜在动力学系统(Hirsch et al., 2012; Smith 和 Thelen, 2003) 通过复杂网络获得了更为普适且具体的表达。同时，由于网络拓扑基于结构表示了内在交互机制，因此抽象的动力学系统可以得到基于数学形式的表达。在此基础上，动力学系统的推断步入建模时代。其中最简易的网络重构方法是以数据动态变化推理连边信息，根据数据随时间的动态变化规律进行信息总结，主要研究包括：（1）基于贝叶斯推断的复杂网络预测(Schäfer 和 Strimmer, 2005; Zhang et al., 2018)；（2）基于因果推断的复杂网络预测(Laghate 和 Cabric, 2017; Wu et al., 2012)。由于受先验信息或数据随机性的影响，基于贝叶斯统计或格兰杰因果推断的方法无法准确的预测网络的邻接矩阵，这导致预测精度低下、模型具有偏见、先验信息要求过高等难以接受的问题。在后续研究中也有学者发展出基于线性统计模型的方法，试图准确预测拓扑信息。然而，在社会、物理、生态等动力学系统中，复杂系统产生的连续动态数据不总能被时刻捕捉，因此观测数据本质上是离散的，比如 Kuromoto 模型(Acebrón et al., 2005)，演化动力学模型(Nowak 和 Sigmund, 2004)，以及 SIS 传染病模型(Allen, 1994)，这为建模提出了挑战。基于上述条件，网络重构问题可以转化为具有稀疏和高维性质的统计线性模型，因为表达网络结构的邻接矩阵往往具有高维且稀疏性(Boccaletti et al., 2006)。基于统计线性模型思想，有学者提出压缩感知 (Compressed sensing，CS)(Wang et al., 2011) 和基于 Lasso 的推断

方法(Han et al., 2015)，其中 Lasso 方法本质上是对凸优化问题的求解(Hastie et al., 2015; Tibshirani, 1996)。然而，通常情况下，邻接矩阵的维数远大于观测数据矩阵的维数，由于邻接矩阵中只有少数非零元素存在，问题表现出极强的稀疏性(Boccaletti et al., 2006; Friedman et al., 2010)，这使得 Lasso 方法能够将邻接矩阵中不重要的预测信号压缩到“0”，但不能将非零信号压缩到“1”。这意味着，实际不存在的拓扑连边可以被准确预测，但实际存在的连边无法被准确预测。有学者(Shi et al., 2020) 提出一种安全筛选降维方法，在基于 Lasso 方法的演化动力学系统重构问题中，一定程度上解决了高维稀疏预测中潜在的运算问题。同一时期，也有一些优秀的线性推断方法，通过改变惩罚函数使参数估计获得更优良的性质，如平滑剪切绝对偏差（SCAD）(Fan et al., 2009)、Adaptive Lasso(Zou, 2006)、Group Lasso(Bach, 2008)、Elastic Net(Zou 和 Hastie, 2005) 等，本质上关注于“0”拓扑信号，与 Lasso 一样无法准确推断实际存在的“1”拓扑信号。为解决非零信号预测不准问题，Shi 等人(Shi et al., 2021) 针对此类二元信号系统重构问题提出 Signal Lasso，通过在 Lasso 的惩罚函数中引入 L1 范数的控制项，使得真实参数估计可以被压缩至“1”。但受限于实际情况，一般无法在一组参数估计中同时使得拓扑信息被估计为“0”和“1”，因此，该模型通过引入调优参数来使模型自动决定偏向于哪种信号。在上述情况下，相比于前文的传统模型，Signal Lasso 能够较好的重构网络，但调优参数为实际应用带来困难。虽然此方法具有还原真实信号“1”的能力，但上述研究表明(Shi et al., 2021)，在压缩信号至“1”的过程中，会产生部分既不为“1”也不为“0”的模糊信号，尤其是位于 0.5 附近的信号，无法被归类为存在连边或不存在连边。为消除模糊信号，需要一种将二元拓扑信号完整压缩估计的方法，因此 Shi 等人(Shi et al., 2023) 进一步提出 Adaptive Signal Lasso（AS Lasso），该方法通过整合 Signal Lasso 的调优参数，并为惩罚函数赋予能够自主学习的权值。研究表明，AS Lasso 可在仅有少量观测的情况下完整且准确的估计拓扑二元信号，无模糊信号且对高斯噪声具有稳健性。同时，此方法仅需一个参数，减少了计算成本且便于应用。

然而现实场景中数据获取存在困难，例如，由于银行、贸易系统具有一定的隐私性，无法得知各个银行之间具体的交易量或各个国家间的贸易信息，只能获取银行的总交易量及国家的总贸易额，在该背景下只能获取到边际数据。为解决边际数据下连边信号的概率预测问题，有学者提出将连边拓扑信号视为概率