## **1.** **Topological measures for identifying and predicting the spread of complex contagions**

识别和预测复杂传染传播的拓扑措施

### 1.1 基本信息

期刊：《Nature Communication》

发表时间：2021

作者：Douglas Guilbeault、Damon Centola

作者信息：

Haas School of Business, The University of California, Berkeley, Berkeley, CA, USA.

The Annenberg School for Communication, The University of Pennsylvania, Philadelphia, PA, USA.

School of Engineering, The University of Pennsylvania, Philadelphia, PA, USA.



### 1.2 文章主要研究内容

社会网络中距离的标准衡量标准--平均最短路径长度--假定了一个 "简单 "的传染模型，即人们只需要接触到一个同伴的影响就可以采用这种传染。然而，许多社会现象是 "复杂 "的传染，人们需要接触到多个同伴才能采用。在这里，文章表明，经典的路径长度测量方法无法定义复杂传染的网络连接性和节点中心性。基于经典的路径长度定义的中心性测量和播种策略经常误判对传播复杂传染病最有效的网络特征。为了解决这些问题，我们推导出复杂路径长度和复杂中心度的测量方法，这大大提高了识别最适合传播复杂传染病的网络结构和中心个体的能力。用43个印度农村的小额信贷项目传播的经验数据验证了我们的理论。

### 1.3 文章内容

**1.“平均最短路径”迷思**

当今最重要的网络测量方法之一是路径长度--定义为图上任何两个顶点之间的最短步骤数。这种测量方法被认为是网络中任何两个节点之间典型距离的稳健指标，因此，一个图的平均最短路径长度，也被称为其特征路径长度，被认为是所有网络的一般拓扑特性。然而，这种衡量路径长度的方法隐含地假设了一个依靠简单传染理论的网络穿越过程，在这个过程中，一个纽带就足以让传染病从一个节点传到另一个节点。但问题在于，许多社会传染是 "复杂的"，个人需要与多个有行为的同伴接触，然后自己才会被激活。根据路径长度的标准衡量标准，网络距离是穿越网络所需的步骤数，其中每一步都由一条纽带组成。然而，对于复杂的传染病，如新技术的传播、健康行为、语言习惯、互联网备忘录、社会运动和政治标签，社会网络中的每一步都需要多个纽带的同行强化。因此，衡量一个复杂的传染病的路径需要衡量网络中的每一步，而不是单一的联系，**而是加强的联系--通常被称为宽桥**。

鉴于社会传播中复杂传染的普遍性，我们认为经典的路径长度测量--以下简称 "简单路径长度"--并没有提供一个令人满意的方式来衡量社会网络中的连接性。简单路径长度的假设是，如果节点i和节点j之间存在有限的单一联系路径，那么传染就可以从节点i传播到节点j。然而，许多经验和正式的研究揭示了一个令人困惑的结果，即在由单一连接部分组成的社会网络中，复杂的传染病可能仍然不可能从一个节点传播到另一个节点。一些研究发现，在社会传染病的传播过程中，节点之间既有拓扑学上的联系，又有社会上的联系，这种经验性的频率表明，简单的路径长度并不能为社会网络中的社会距离和联系提供令人满意的衡量。

简单的路径长度无法正确衡量网络的连接性，这给长期以来解决识别社会网络中哪些个体（即种子）对传播新行为最有影响力的问题带来了新的挑战。在理论和应用社会网络中，成熟的节点中心性测量方法（如度数中心性、介数中心性、特征向量中心性、K核心中心性）已经成为描述对社会观念传播最有影响力的节点的常用工具。然而，一些关于社会传播的实证研究发现，这些节点中心性的测量方法错误地识别了最具影响力的行为者。例如，最近关于社会媒体的研究结果显示，具有最高的介数和度中心的人往往不是传播政治信息和有争议的新闻的最有影响力的节点，因为这类信息是复杂的传染病。

为了解决这个问题，以及它对网络结构和社会影响之间关系的临时概念所带来的挑战，我们推导出一种路径长度的测量方法，称为 "复杂路径长度"（PLC），它阐明了所有社会网络的连接性和中心性（"复杂中心性"）的新的、可推广的拓扑学属性。虽然我们的研究是由关于复杂传染病传播的经验发现所引起的理论挑战所激发的，但我们确定了在简单和复杂传染病中都具有普遍性的网络措施，包括传播可能需要的同伴强化的所有范围。

我们将我们的发现介绍如下。(1) 首先，为了清楚起见，我们定义了复杂传染的一般影响模型。(2) 其次，我们推导出一个一般的拓扑学措施，用于计算复杂传染病传播的图中节点之间的网络距离：即复杂路径长度。(3)第三，我们利用上述结果推导出一个通用的节点中心性度量：即复杂中心性。(4) 第四，我们提供了稳健性测试，证明复杂路径长度为估计模拟复杂图上的级联频率提供了很好的匹配。(5) 第五，我们提供的数字结果表明，在确定对复杂社会传染病的 "播种 "最有效的独特网络位置方面，复杂中心性优于最突出的节点中心性措施。。(6) 第六，我们使用复杂路径长度和复杂中心性的测量方法来预测对小额信贷项目的传播最有影响力的网络位置，正如43个印度村庄的社会传染的经验研究中所报告的那样。

**2.复杂路径长度**

首先，我们概述了支撑复杂传染的正式影响模型。首先，我们把G(V, E)定义为一个无权无向的图。我们把节点i的邻居表示为N[i]。在复杂的传染模型中，每个节点i∈V被分配一个采用阈值Ti，指定节点i要接触多少个激活的同伴才能采用这种传染。阈值可以是绝对的，即它们规定了触发节点i采用所需的激活邻居的原始数量，或者它们可以是分数的，在这种情况下，它们规定了必须激活i的邻居的一部分来触发i的采用。阈值可以同质分布（即对所有节点保持不变）或异质分布，即每个节点以某种概率收到不同的阈值（例如，每个节点从一个定义的区间均匀地随机分配一个阈值）。扩散在离散的时间步骤中展开：在步骤t，所有在t-1中活跃的节点保持活跃，我们激活任何有足够数量的激活邻居的节点j，以满足其阈值Tj。——类似LT模型

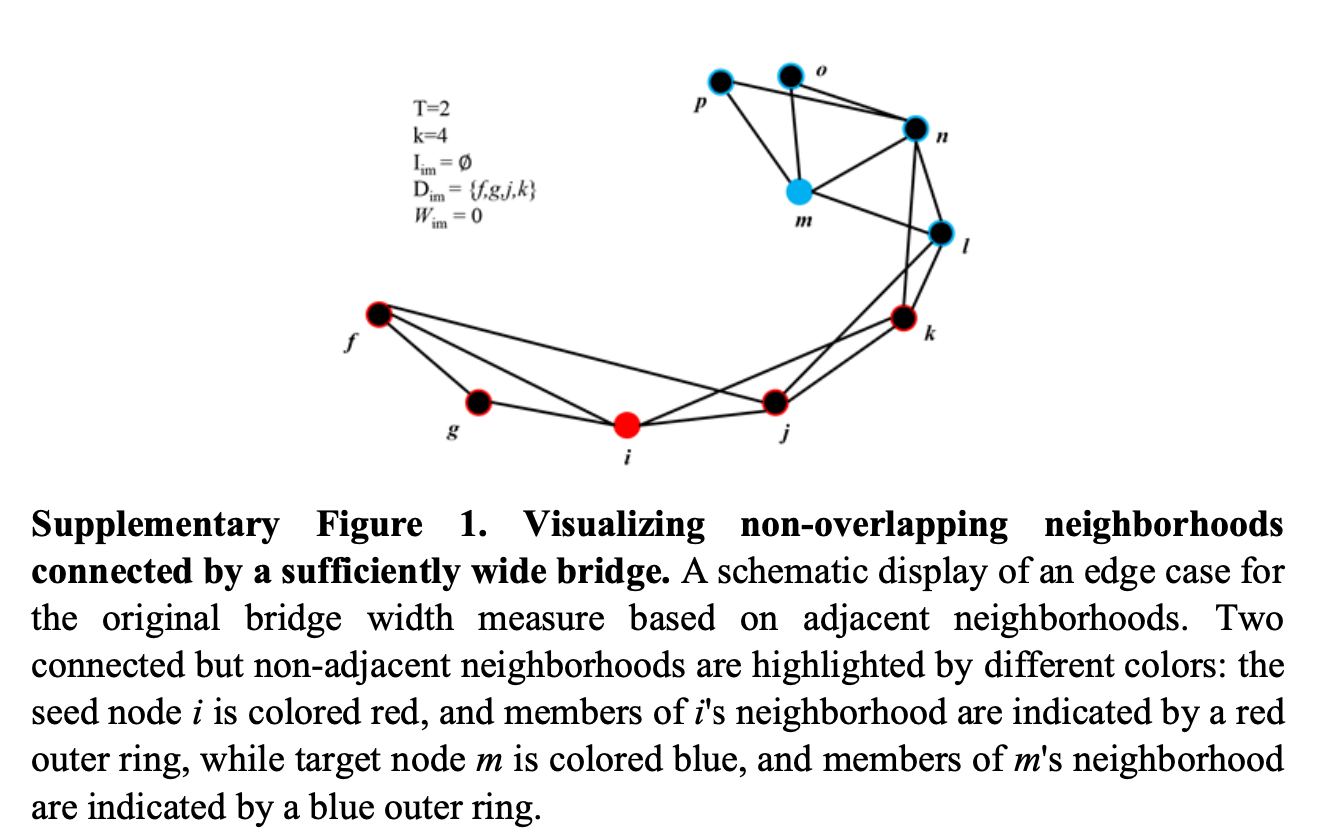
我们把节点i和节点j之间的复杂路径定义为传染必须穿越的邻域序列，从节点i的邻域N[i]到任何节点j。为了描述复杂路径的结构和扩散能力，我们提供了一个新的桥梁宽度的正式定义，以确定连接的邻里之间的加强联系的数量是否足以使复杂的传染病传播。节点i和节点j之间的复杂路径长度被定义为复杂传染病从**N[i]传播到节点j时穿越的足够宽的桥的数量**。我们用这种方法来确定网络中任何距离的节点之间的桥链，并确定这些桥链的长度和宽度（图1）。这种网络连接性的定义，PLC，促使我们对复杂传染病的节点中心性进行新的衡量，即复杂中心性（CC）。节点i的复杂中心性（CCi）是指从节点i的邻居N[i]延伸出来的复杂路径的平均长度，用PLCi表示

**3.方法**

**（1）桥梁宽度的局部测量**

**传统桥宽定义**

早期的工作以一维网格上两个相邻邻域之间的重叠纽带数量来定义桥梁的宽度。让N[i]指节点i的包容性（即封闭的）邻域，定义为G的子图，包括与i相邻的所有顶点以及i。邻域N[i]和N[j]之间的桥被定义为**邻域N[i]和N[j]的共同成员集（称为Iij）**与**包含不在N[j]中的N[i]成员**的不相交集（称为Dij）之间的联系。桥梁的宽度被定义为从Iij到Dij的连接集的大小。



这个定义的一个关键限制是，它不能超越相邻的邻里；也就是说，当Iij = ∅（当两个邻里没有重叠的成员）。因此，这个措施有一个问题，即即使两个邻里直接相连，它也将这两个邻里之间的桥梁大小确定为0，而一个复杂的传染病可以在它们之间蔓延。补充图1显示即使两个邻里之间没有重叠的成员，但复杂的传染病可以在它们之间传播。

出于这个原因，我们更新了桥宽的定义，以考虑到两个邻里相连但没有共享重叠成员的情况，使桥宽的测量能够超越格子的范围。我们开发了以下逻辑，以确定何时存在足够的桥梁，可以在两个邻里之间传播传染，**这两个邻里有共同的连接，但不一定相邻。**

给定无权无向图G。我们将节点i和节点j之间的复合路径定义为复合传染必须穿越的邻域序列，从i的邻域N[i]到任何节点j。我们使用邻域之间的桥梁概念衡量两个邻域之间的连通性。如果两个邻里之间的桥梁足够 "宽"，足以支持不同的融合，它被称为 "足够的桥梁"。对于具有采用阈值Tj的传染，当且仅当满足以下条件时，从节点i的邻域到节点j的邻域存在一个充分的桥梁。为了简单起见，我们假设Tj在所有节点上是同质的，但这些措施可以很容易地适应Tj的异质分布。

N[i]：节点i的封闭邻域，定义为G的一个子图，包括与i相邻的所有顶点，以及i。

E(N[i])：节点i的邻域的边集，包括N[i]内所有与i的联系。

Tj：节点j的采用阈值（即节点j需要遇到的激活对等体的数量或比例，才能激活）。

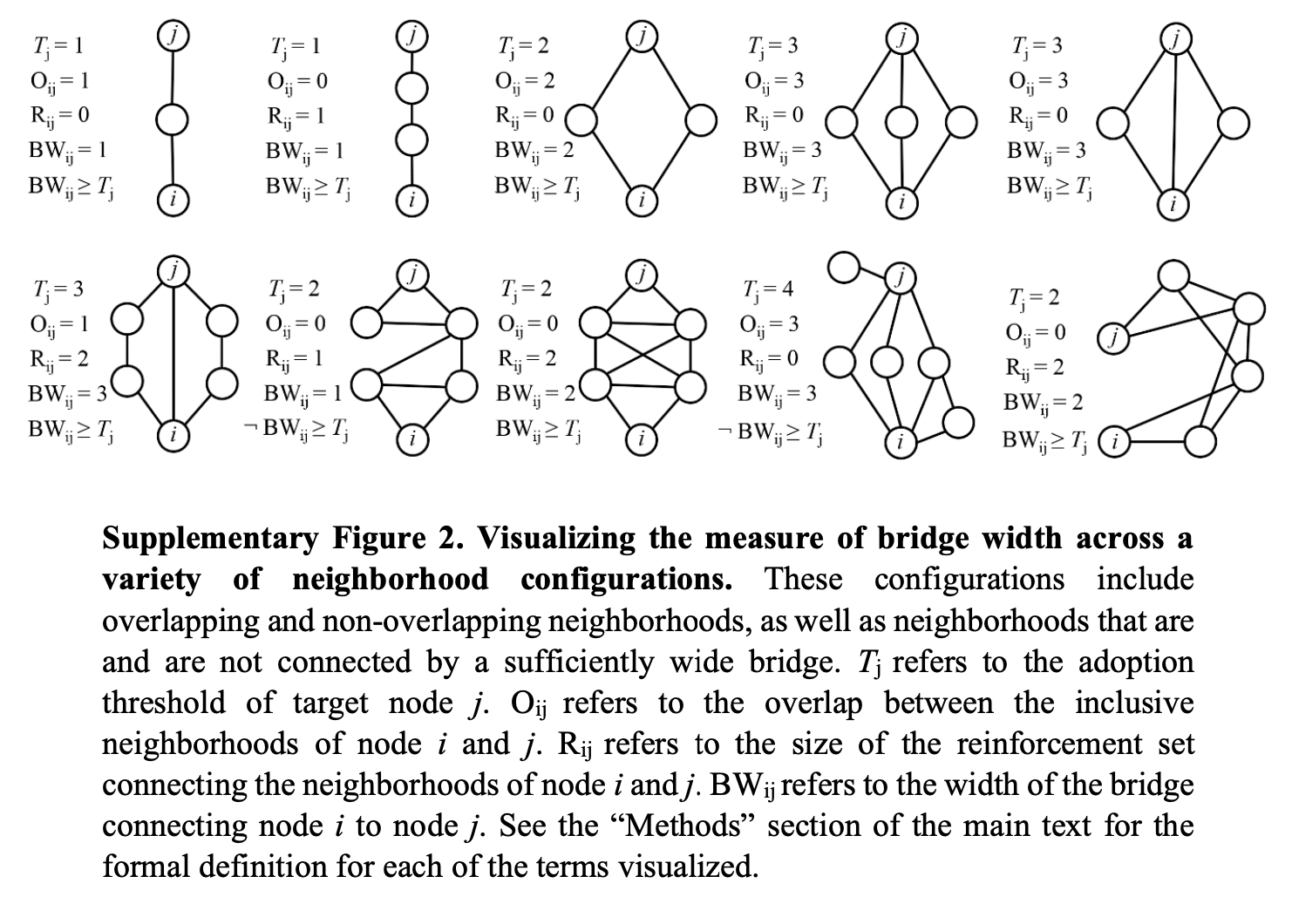
Oij：N[i]和N[j]的交集，即共同邻居。

Dij：N[j]中不在N[i]中的节点的不相交集合，即独属于j的邻居。

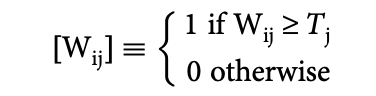
Rij：节点的 "强化 "集，它由Dij中与N[i]中的节点相连的节点组成，即j的朋友和i的朋友互相认识的部分。

桥：Oij与Rij的并集。

桥宽Wij：Wij≡|BWij|（节点i和j之间的桥的数量）。



同质激活条件：如果Wij≥Tj，N[i]和N[j]之间的桥可以支持传染的传播，即桥是局部充分的。给G中的每座桥赋予一个二进制值，表示该桥是否足够宽以实现扩散。

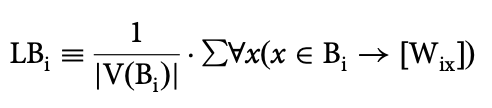


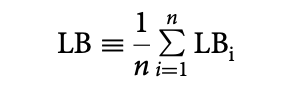
异质激活条件：可以通过要求Rij只由Dij中能被N[i]激活的节点组成，来适应门槛的异质性分布。这要求我们只在Oix≥Tx的情况下将每个节点x保留在Rij中--即如果有足够的来自N[i]的联系来满足Tx。（通过激活j足够多的邻居来激活j）。

**（2）基于桥梁宽度的网络联通性测量**

Bi：V中与N[i]局部共享一个桥的节点子集，∀v(v ∈ Bi →v∈V∧v≠i∧Wvi ≥1).。相当于去算所有与i构成桥关系的节点，计算其中能够激活i的桥比例。

基于上述定义，可以计算本地桥梁的比例，这些桥梁的宽度足以让一个节点将传染病传播到其邻域之外。——衡量网络的局部联通性





**（3）复杂路径长度**

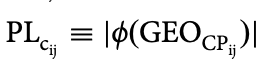
节点i和j之间的复杂路径长度（PLc）被定义为N[i]和节点j之间的复杂路径中所穿越的足够的桥梁数量。我们假设，如果传染病不能从N[i]传播到节点j，那么PLcij=0。 如图1所示，我们通过以下方式来定义

CPij：复杂传染病从N[i]传播到节点j时激活的节点子图，它包含N[i]和节点j之间可能的复杂路径集合。

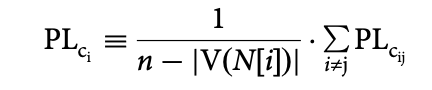
GEOCPij：指CPij内节点i和节点j之间的测地线，它确定了CPij内最短的复杂路径。

φ(GEOCPij)：GEOCPij中的顶点序列。

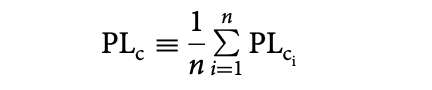
PLc：N[i]和节点j之间的复杂路径长度

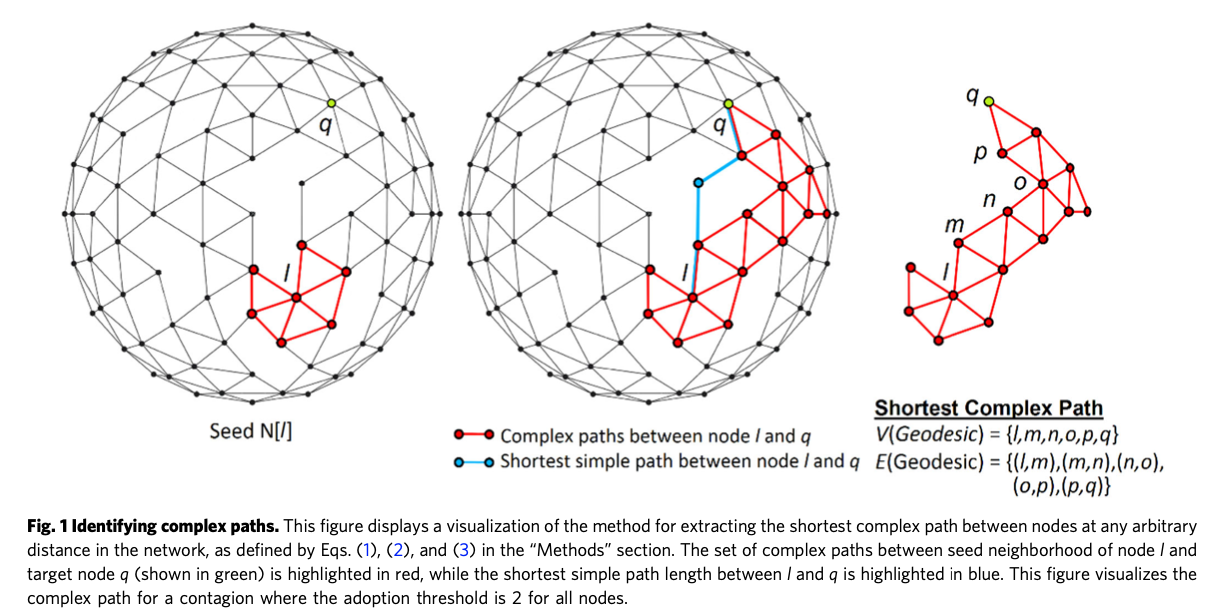


平均复杂路径长度：



全局复杂路径长度：

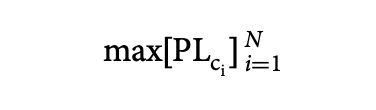




该图显示了提取网络中任意距离的节点间最短复杂路径的方法的可视化，如 "方法 "部分中的公式（1）、（2）和（3）所定义的。节点l的种子邻域和目标节点q（绿色显示）之间的复杂路径集合用红色突出显示，而l和q之间的最短简单路径长度用蓝色突出显示。该图直观地显示了所有节点的采用阈值为2的传染的复杂路径。

**（4）复杂中心性**

与度中心性的定义，一个节点i的复杂中心性（CCi）是PLCi，其中图中复杂中心性最高的节点是平均复杂路径长度最高的节点，形式上表示为。



通过用最高的平均复杂路径长度来定义一个节点的中心性，这种方法可以确定任何图中的节点，对于任何传染病，这些节点都有最长和最强大的足够的桥梁链，可以到达最大数量的目标节点。

**4.结果**

**（1）桥宽测量有效性验证**

研究假设：(i) 度均匀性，即每个节点在其邻域都有相同数量的联系人，以及(ii) 阈值类型，即每个节点都被分配了一个固定的绝对采用阈值，即一个人需要接触到的采用者数量。

首先研究了一个连续的k-规则图（同质网络）。以前的工作发现，增加图形的随机性会大大减少所有节点之间的平均最短简单路径长度，从而加速简单传染病的传播。然而，我们发现**复杂传染病的传播有相反的效果，网络结构中随机性的增加破坏了足够宽的桥梁的比例，导致复杂传染病传播能力的急剧下降。**在这里，我们分析了每个图中的每个邻里桥，并计算出足够宽的桥的比例，以实现扩散。我们把这个数量称为局部足够的桥的比例（LB）。我们用LB来估计一个图支持复杂传染的全球级联的能力，我们通过对试图使用每个节点及其邻居作为扩散过程的初始种子而产生的扩散结果进行平均计算。在证明了这些在规则图上的随机化结果后，我们又将这些发现推广到随机无标度图上。

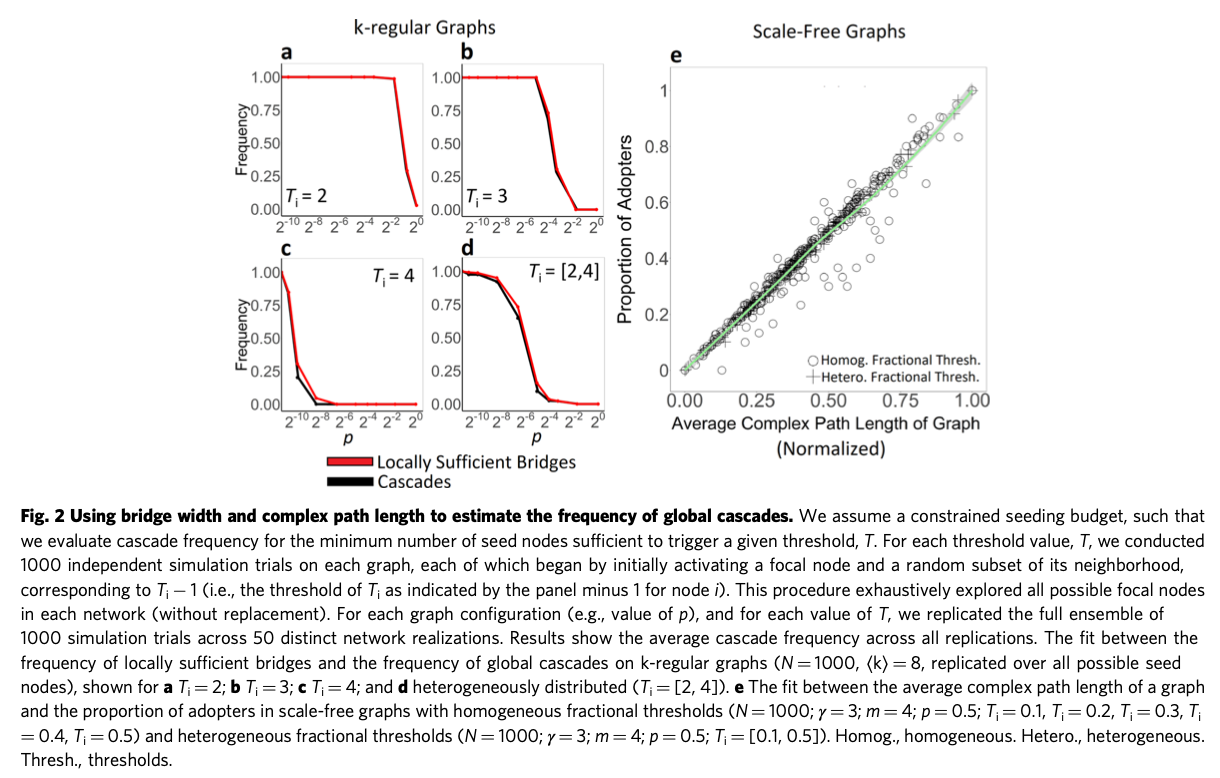


图2的a-d显示，随着图形变得越来越随机（激活阈值改变），全局级联的频率随着足够宽的桥梁的平均比例而急剧下降。局部足够宽的桥的频率准确地估计了具有不同程度随机性的k-规则图支持复杂传染的全球级联的能力--不管阈值是均匀分布还是异质分布。换句话说，图中局部足够的桥的频率提供了一个衡量图的连接性的有效方法--即图支持复杂传染病在群体中任何两个节点之间传播的能力。

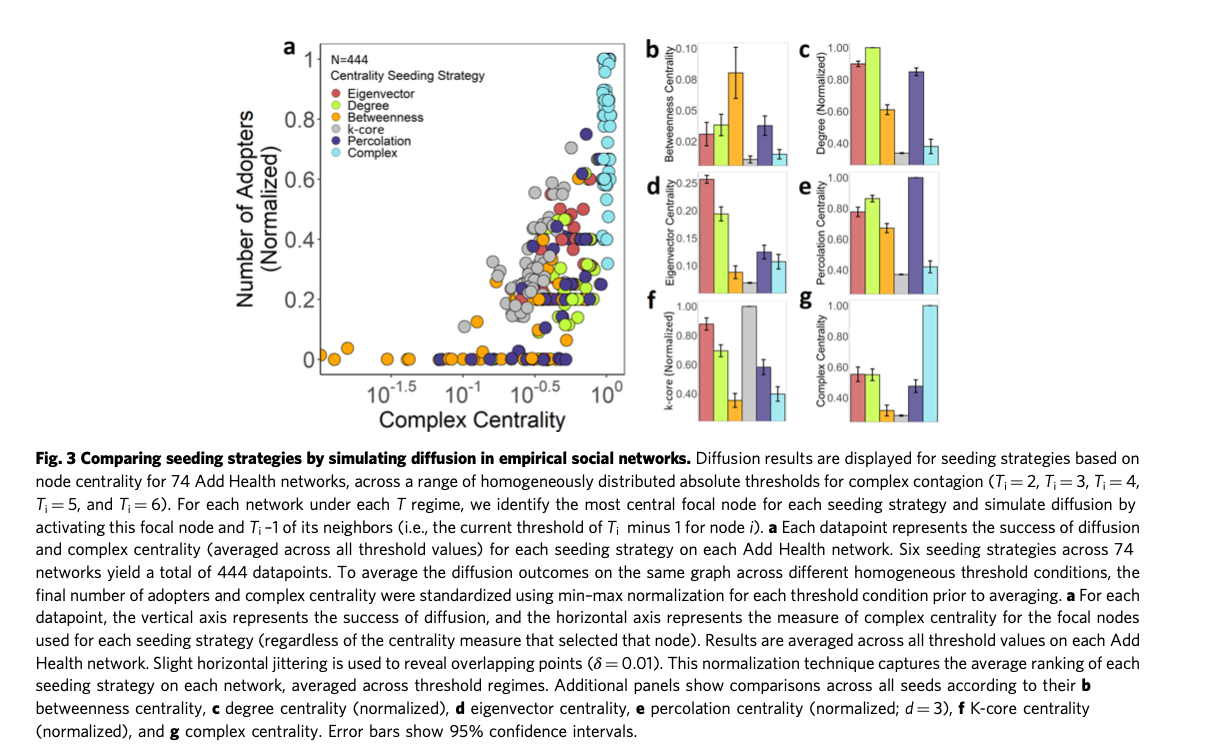
图2的e小组通过研究平均复杂路径长度将这些结果推广到随机无标度网络。由于复杂的路径是由足够宽的桥链形成的，因此，图的平均复杂路径长度应该有效地估计人口中可以被复杂传染病到达的节点的比例。面板e显示，复杂（无标度）图的平均复杂路径长度准确地估计了网络中所有可能的种子邻域中产生的级联的平均大小，使用同质分数阈值和异质分数阈值。这些发现对于绝对阈值的同质和异质分布，以及k-规则图都是稳健的。

**（2）复杂中心性验证**

本文评估了复杂中心度的性能，以确定在一组为研究公共卫生行为（例如，在COVID-19大流行期间戴口罩）的传播而收集的经验社会网络中最具影响力的种子节点。我们使用了74个经验社会网络，这些网络来自Add Health数据集--这是公开的最大的青少年社会网络集合，来自美国70多个不同的社区。在每个网络中，我们将具有最高复杂中心性的节点与理论上定义的衡量最 "中心 "网络位置的各节点--即通过**度中心性、介数中心性、特征向量中心性、K壳中心性和渗入中心性**--的有效性进行比较，并评估不同阈值的复杂社会传染病的传播。

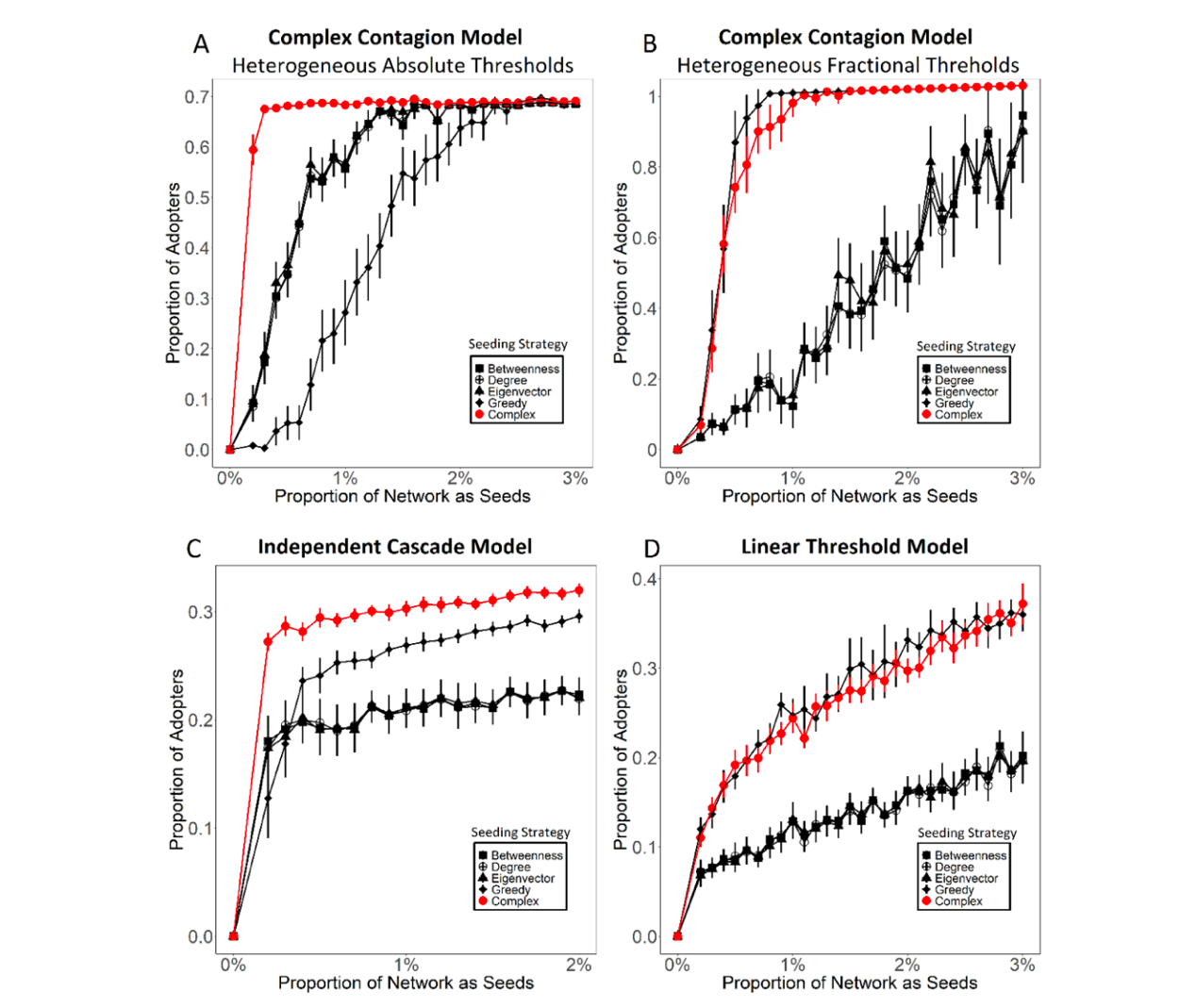
我们通过激活理论上确定的种子节点（根据每个节点中心性理论）和其邻居的随机子集来启动每个中心性措施的测试。

根据先前对复杂传染的研究，并为了阐述清楚，我们提出了我们对具有均匀分布的绝对阈值的群体的基本发现，范围从Ti=2到Ti=6。我们的完整结果（在SI中提供）表明，这些发现对同质和异质阈值分布以及使用绝对或分数阈值都是稳健的。在下面的结果中，每个网络对Ti的5个值（Ti=2、Ti=3、Ti=4、Ti=5、Ti=6）各产生6个观察值（每个播种策略一个）。这为每个网络产生了30个观测值，总共有2220个观测值。为了给每个策略提供一个总结性的结果，我们对每个网络上每个播种策略的所有阈值的扩散成功率进行了平均。



面板a中的横轴表明，无论选择该节点的特定播种策略如何，节点的复杂中心度都能有效地识别其在网络中的整体影响力。图3b-f显示，与基于简单路径长度的中心性测量所识别的节点相比，复杂中心性识别的有影响力的节点具有定性的不同拓扑位置。也许最令人惊讶的是，复杂中心性识别出的种子节点根据最流行的中心性测量方法具有低影响力。

在各种拓扑结构和影响模型中，复杂中心性也能比典型的贪婪算法更有效地确定复杂传染病传播的有影响力的种子，该算法模拟每个可能的种子的扩散，并根据它们的个别表现选择具有最大预期扩散的种子集。



为了检验我们的复杂路径长度和复杂中心性的网络衡量标准在预测网络结构对经验扩散过程的影响方面的有效性，我们利用一项关于印度农村小额信贷项目传播的经验网络研究来评估复杂路径长度和复杂中心性的理论预测。这些数据提供了异常全面的网络扩散记录，收集自43个不同的村庄，每个村庄都记录了完整的网络数据，以及一个新的小额信贷项目在每个村庄每个家庭的传播动态的完整记录。这些数据为我们的假设提供了一个非常有力的检验，即（i）复杂路径长度将预测小额信贷项目将在其中最有效地传播的村庄网络，以及（ii）复杂中心性将预示着每个村庄中对传播BSS项目最有影响力的特定家庭。

我们使用每个村庄的估计阈值分布的集合来计算每个村庄的平均复杂路径长度和每个家庭在每个村庄内的平均复杂中心度。与图3中采用的方法类似，我们评估了每个村庄中每个家庭的复杂中心度与由此产生的小额信贷项目从每个采用家庭向其网络联系人的传播之间的相关性。图4面板a显示了 "领导者 "家庭（即在最初的研究中专门挑选出来在每个村子里启动小额信贷项目的家庭）的结果，面板b显示了 "普通 "家庭（即所有43个村庄的所有家庭，其中许多是后来的项目采用者，他们帮助在各自的村庄里传播项目）的结果。图4中的a和b板块也报告了对上面讨论的所有中心性措施的相同分析：度、特征向量、间性、k-核心和渗入。此外，与上述图2采用的方法类似，我们还评估了每个村庄的复杂路径长度与每个村庄的总体采用率之间的相关性（如下图4的c组所示）。

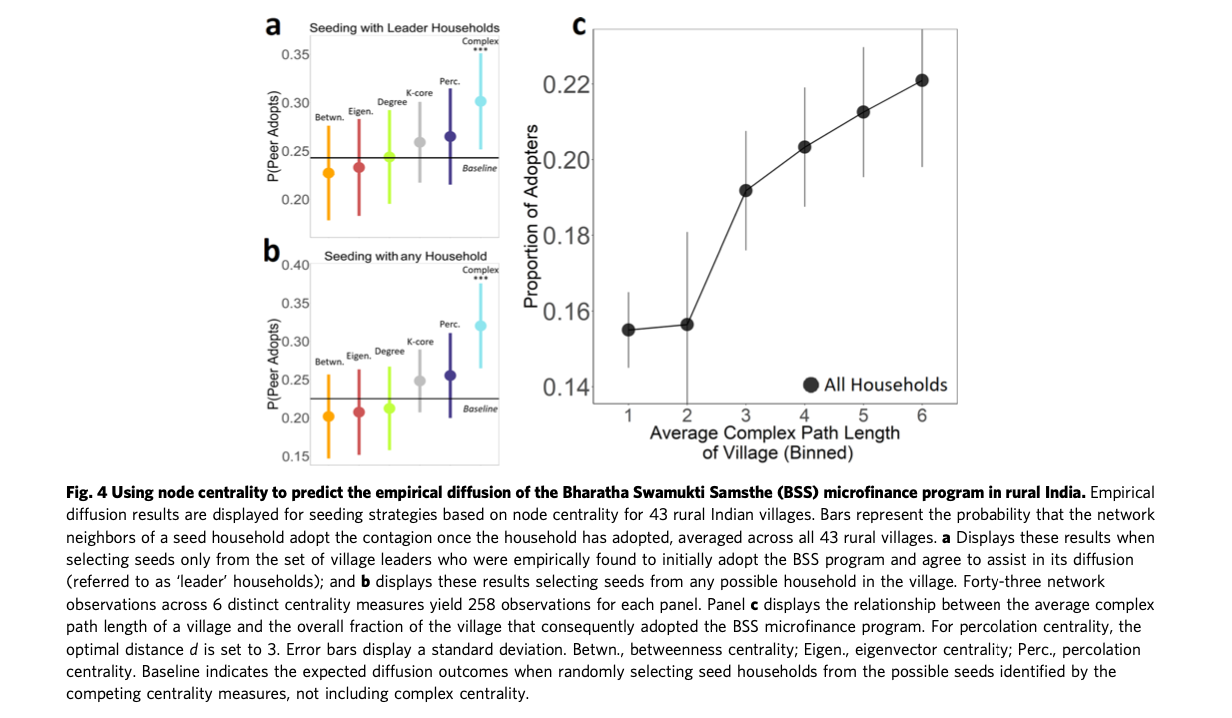


图4的面板a显示，具有最高复杂中心性的领导家庭与诱导小额信贷项目传播的概率明显高于基于现有节点中心性测量的最高中心性分数的领导家庭。值得注意的是，复杂中心性是唯一能够识别领导家庭的拓扑学措施。图4的b组复制了同样的结果，同时将种子的搜索范围扩大到Banerjee等人所选择的预先指定的领导者之外。在每个村庄的每个可能的家庭中，具有最高复杂中心性的家庭与诱导其网络邻居采用小额信贷项目的概率明显较高有关。图4的c组测试了村庄层面的假设，即整个村庄的平均复杂路径长度与小额信贷项目的总体采用率呈正相关。与图2中显示的理论预测一致，图4的c组显示，平均复杂路径长度较高的村庄表现出明显更多的项目采用。

**5.研究发现**

（1）网络扩散的理论和应用研究的一个共同假设是，拥有更多联系的人更有影响力。我们的研究结果不同意该文献中经常宣称的观点，即无论在什么情况下，度中心都是识别社会网络中最有影响力的个人的有效手段。

(2)在社会网络的组织研究中，一个常见的假设是，信息经纪人--即参与多个不同网络社区的人，在很大程度上是互不相干的--有超强的影响力，因为他们是社区间传染病流动的守门人。这一假设使得 "介数中心 "成为组织理论中最广泛使用的网络影响力衡量标准之一。

相比之下，我们的研究结果表明，低度中心性和低介数中心性的网络位置却可能是人口中最有影响力的地点。我们还发现，拥有最高程度中心性和介数中心性的个人通常在启动**复杂社会传染病**的传播方面占据无效的网络位置。我们预计，未来工作的一个重要方向是探索计算复杂路径长度和复杂中心性的理论属性的新算法，这可能会受益于最近的发展，改善新算法技术的可扩展性。未来研究的另一个有趣的方向是应用我们的拓扑学措施来确定特定的网络位置，这些位置可以用来有效地阻止现有的复杂传染病从网络的一个部分向整个人群的传播（类似于简单传染病的网络 "免疫 "问题）。