# 1. Identification of influencers in complex networks by local information dimensionality（7.10 杨钧文）

通过局部信息维度识别复杂网络中的影响者

## 1.1 基本信息

期刊：《Information Sciences》

发表时间： 5 October 2019

作者： Tao Wen, Yong Deng

作者信息：

Institute of Fundamental and Frontier Science, University of Electronic Science and Technology of China, Chengdu, 610054, China

分区：



## 1.2 论文内容

### 一、摘要

本文提出了一种通过局部信息维度识别复杂网络中影响者的方法。所提出的方法考虑了中心节点周围的局部结构特性；因此，局部性的尺度只增加到离中心节点最短距离的最大值的一半。因此，所提出的方法考虑了准本地信息并降低了计算的复杂性。box里的信息（节点数）通过香农熵来描述，这更合理。当一个节点的局部信息维度较高时，它的影响力更大。为了证明所提出的方法的有效性，我们用现有的五种中心度量方法作为比较方法，对六个真实世界的复杂网络中的有影响力的节点进行排名。此外，还应用了易受感染（SI）模型和Kendall's tau系数来显示不同方法之间的关联性。实验结果显示了所提方法的优越性。

### 二、引言

复杂网络是一个热门话题，吸引了许多领域的研究人员的关注[14]，因为它们可以作为许多现实世界复杂系统的详细模型，如大脑网络[7]、信息网络[30,39]、人类生活[9]和社会系统[25,27]。复杂网络的许多结构特性都受到一些特殊节点的影响，例如，复杂网络的无标度[18]、自相似性[22]和分形[33]特性[23]。为了有效地测量网络的属性，许多研究已经进行了寻找这些具有特殊属性的节点，例如，寻找最相似的节点[34]，识别有影响力的节点[26,43]，以及预测潜在链接[36]。特别是，在复杂网络中具有高度影响力的节点逐渐引起了研究者的注意，因为它们比其他大多数节点对网络的属性和结构有更大的影响，如通过可见度图预测时间序列[11]、通过相似节点预测链接[2]、检测社会网络中的社区[37]、测量网络复杂性[50]和划分网络结构[41,42]。

一般来说，每个网络都有特定的节点重要性排名，不同的识别方法考虑到网络的不同结构属性，会给出不同的排名列表。目前已经提出了许多中心性措施来识别有影响力的节点，它们可以分为三类[13]：基于邻域的中心性、基于路径的中心性和迭代-细化的中心性。这些中心性有很多经典的测量方法，如度中心性（DC）[16]、间性中心性（BC）[16]、接近性中心性（CC）[5]和特征向量中心性（EC）[5]。此外，还有一些新的测量方法，如H指数中心性[44]、最优渗滤理论[3]和证据理论[15]中的测量方法、通过与理想解相似性进行排序偏好的技术（TOPSIS）[48]，以及其他测量方法[31,46,47]。这些中心性度量已被应用于各个领域，如博弈论[32]、人类合作[19]、进化游戏[20]、相关网站排名[49]和节点同步[4,38]。然而，这些经典的中心度量有其局限性。例如，DC只集中于局部信息，不考虑全球信息。BC和CC关注全球信息，但其高计算复杂性限制了它们在大规模复杂网络中的应用。EC不能用于不对称网络，这就减少了它的应用。最近，人们提出了一些新的中心度量。例如，Zareie等人利用熵对有影响力的节点进行排名[45]。Deng等人提出了一个局部维度来识别重要节点[21]。Makse等人追踪了社会网络中的真实信息流，以找到有影响力的传播者[29]。

熵是测量复杂网络信息的有用工具；因此，它被广泛用于许多应用中，例如，脆弱性的评价[35]、维度的呈现[17,24]、困境实验[6]、数据融合[8,28]、纠缠测量[40]和证据理论[10,12]。此外，复杂网络的结构，如节点和链接，可以被视为概率集。因此，结构属性可以通过熵来有效探索，这为解决网络中的问题提供了一种新的方法，包括重要节点的识别。

本文提出了一种新的中心性度量方法，在局部信息维度的基础上识别有影响力的节点。提出的方法通过香农熵来考虑box里的信息，这比经典的测量方法更合理。与以前的方法相比，所提出的方法的局部性尺度从最短距离最大值的1增加到一半，这可以考虑准局部信息并降低计算的复杂性。本地信息维度较高的节点在复杂网络中的影响更大，这与经典的测量方法相同。为了证明所提出的方法的有效性和合理性，我们考虑了6个真实世界的复杂网络，并将现有的5种中心性度量方法作为比较方法。此外，还使用了易受感染（SI）模型和Kendall's tau系数[13]来显示所提方法的优越性和不同方法之间的关系。

本文的其余部分组织如下。第2节介绍了一些现有的中心性测量方法和关于复杂网络的概念。第3节讨论了拟议的局部信息维度。第4节介绍了一些数值实验，以说明所提方法的有效性和合理性。第5节讨论了结论。

### 三、模型解析

#### 1、前置概念

##### 局部维度Local dimension (LD)

为了探索复杂网络的局部结构特性，Silva等人提出了复杂网络的LD。幂律分布已被证明存在于具有特殊属性的理论网络中，如小世界属性和许多现实世界的网络。

由于从每个中心节点的拓扑尺度不同，LD随着中心节点的选择而变化。Pu等人[21]修改了LD来识别复杂网络中的重要节点。对于半径为r的网络，人们发现，与中心节点最短距离小于r的节点Ni(r)的数量遵循幂律

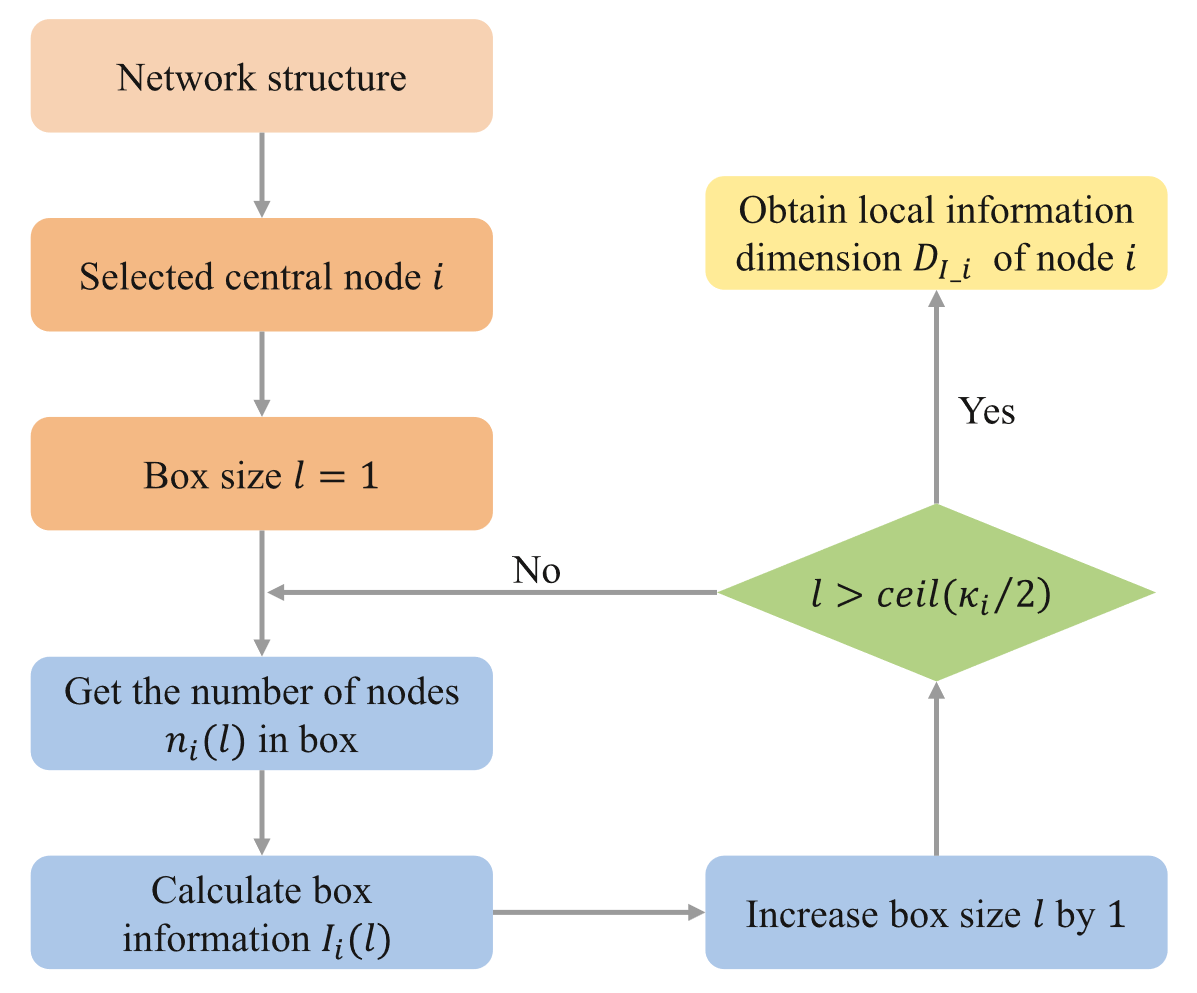
可以很容易地发现，节点i的LD 可以从对数图的斜率中得到，其表示方法如下：

其中d是导数的符号。半径r从1增加到与节点i的最短距离的最大值，由于复杂网络的离散特性[1]，公式（8）的导数表示如下：

其中是与中心节点最短距离等于r的节点数。当选择了中心节点后，可以确定中心节点的地域性规模，中心节点的LD可以通过对数图的斜率（ vs. ）得到。最后，一个节点的重要性可以通过LD的顺序来确定。与之前的方法相比，LD较低的节点在网络中更有影响力。

#### 2、模型解析

人们提出了许多中心度量方法来识别复杂网络中的有影响力的节点。不同的方法考虑了网络中不同的结构信息，并有各自的优势和局限。由于以前的方法大多集中在全局或局部结构上，所选节点周围的准局部结构不能被有效识别。本文提出了一种新的方法，根据复杂网络中每个节点的局部信息维度（LID）来识别重要节点。该方法考虑了每个节点周围的准本地信息，降低了计算的复杂性。第4节通过对一些真实世界的复杂网络进行实验，证明了所提方法的实用性和有效性。图1显示了获得一个选定节点的LID的流程图。



在本节中，提出了复杂网络的LID。在所提出的方法中，每个box中的节点数量是用香农熵来表示的，这更合理。与LD 类似，LID 也考虑了复杂网络中节点i周围的结构特性，其定义如下。

其中，是导数的符号，是box的大小，是box中的信息，其中心节点是节点, 大小为。与经典的LD不同，所选box中的信息是用香农熵来描述box中的节点数。此外，支配box大小增长的规则与经典的定义不同。box的大小从1增长到与中心节点最短距离的最大值的一半，即。box大小的变化意味着LID专注于中心节点周围的准局部结构，并降低了计算的复杂性。每个box里的信息是由box里的节点数决定的，使用香农熵来考虑选定节点的数量。因此，box里的信息可以更合理地表明节点的属性，它的定义如下:

其中是指在给定的box大小为时，中心节点为的box中包含信息的概率，它是box中的节点数与复杂网络中的总节点数的比率，通过以下方式获得:

因此，公式（11）中的LID可以被改写为：

从公式（14）中，围绕中心节点的box里的信息由box里的节点数通过香农熵得到。所选节点的LID 由 和 之间关系的拟合线的斜率得到。由于网络的离散性[1]，公式（14）中带导数的表达式可以改写为:

其中是与中心节点的最短距离等于box大小的节点数（），是与中心节点的最短距离小于box大小的节点数（）。

在所提出的方法中，定位尺度随中心节点的变化而变化，它被定义为与中心节点最短距离的最大值的一半，即。Box的大小 从1增加到地域性的尺度。每个box里的信息（box里的节点数）是用香农熵来表示。每个节点的LID可以从box的信息 的斜率和box大小 的对数得到。由于LID的特性，所提出的方法更合理地考虑了box中的信息，降低了计算的复杂性。

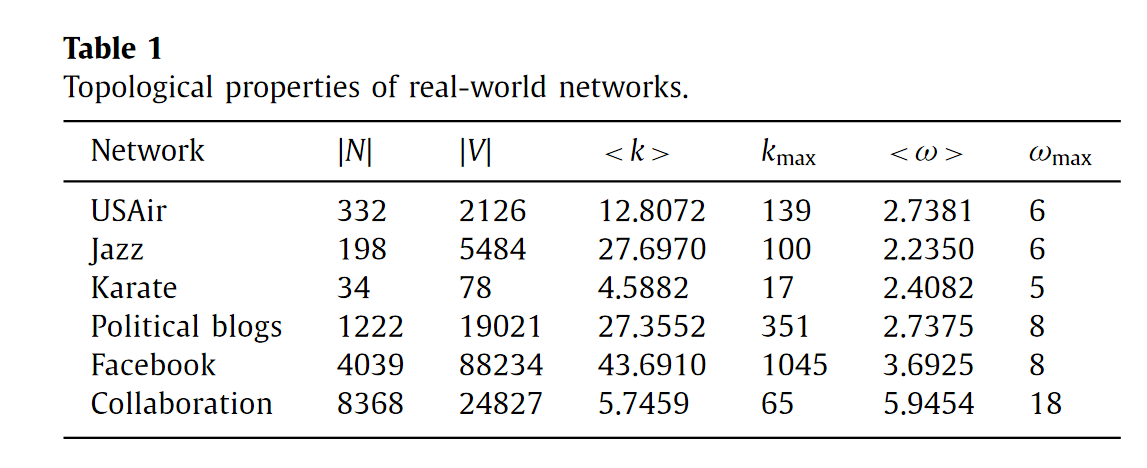
## 1.3 实证分析

### 零、数据来源

六个复杂网络：USAir网络、Jazz网络、Karate网络、Political网络、Facebook网络和来自arXiv的（High Energy Physics - Theory） collaboration 网络。可以从

http:// vlado.fmf.uni-lj.si/pub/networks/data/和<http://snap.stanford.edu/data/>

下载。此外，collaboration网络被选为原始网络数据中最大的连接子图。五种比较措施为BC、CC、DC、EC和LD。这六个网络的结构属性列于表1



实施了五个实验，包括：1、TOP10对比 2、SI模型感染能力3、节点分数关系图4、基于SI模型的Kendall's tau系数5、运行时间

### Top-10 nodes

现有的五种度量方法中的彩色节点与LID识别的前10名节点相同。



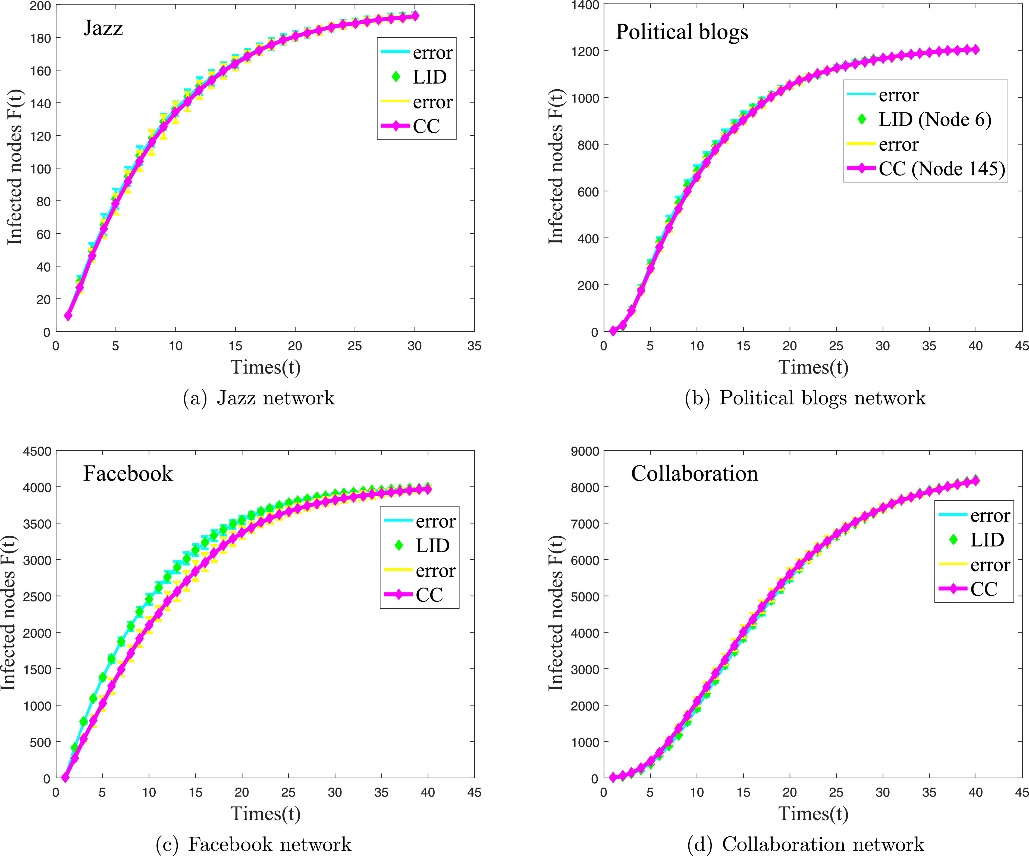
从表2来看，CC和DC是与LID相似的两种测量方法，因为它们可以获得比其他方法更接近的排名。BC和EC在这些网络中有不同的表现；有些网络的排名相似，有些则不同。EC和LID之间没有相同的前10名节点的原因是网络规模的影响。EC在大规模的网络中没有很好的表现，因为它的前10名节点名单与其他方法完全不同。

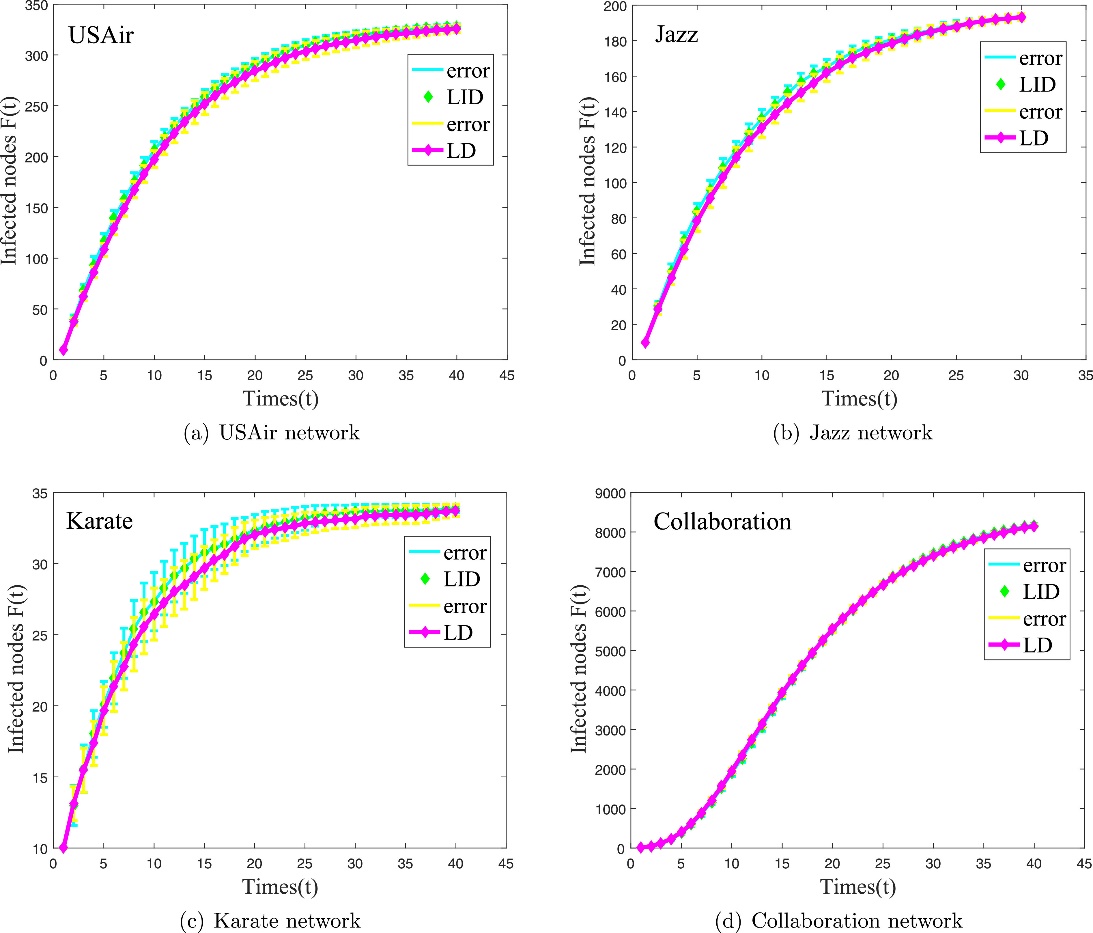
总之，LID表现出的性能更接近于现有的识别前10名节点的措施。下面将讨论与实验的更详细比较。由于所提出的方法是在LD的基础上修改的，而且它关注的是一个节点在不同距离上的影响，所以在下面的实验中选择LD（最相关的方法）和CC（相同的考虑因素）进行比较。

### SI model

受感染的节点有一个传播率，用于感染其邻近的易感节点，受感染和易感节点的总数等于复杂网络中的节点数。对于不同规模的网络有不同的设置。在时间t的感染后，感染能力较强的初始节点在网络中引起的感染节点数量较多，这可以说明这些节点的重要性。在某个特定的时间t，感染节点的数量被选为衡量初始感染节点的感染能力的指标。感染节点的数量越多，说明初始节点的感染能力越强，初始节点越重要。

选择LD和CC作为比较方法是因为它们对信息的考量。LID与CC和LD的比较是通过以下描述的SI模型进行的。首先，初始感染节点被选为不同方法得到的前10个节点，并列于表2。然后，感染过程持续时间为t，并记录感染节点的数量。最后，每个实验都是独立进行的，重复50次，。实验结果是50次实验的平均值，如图2和图3所示

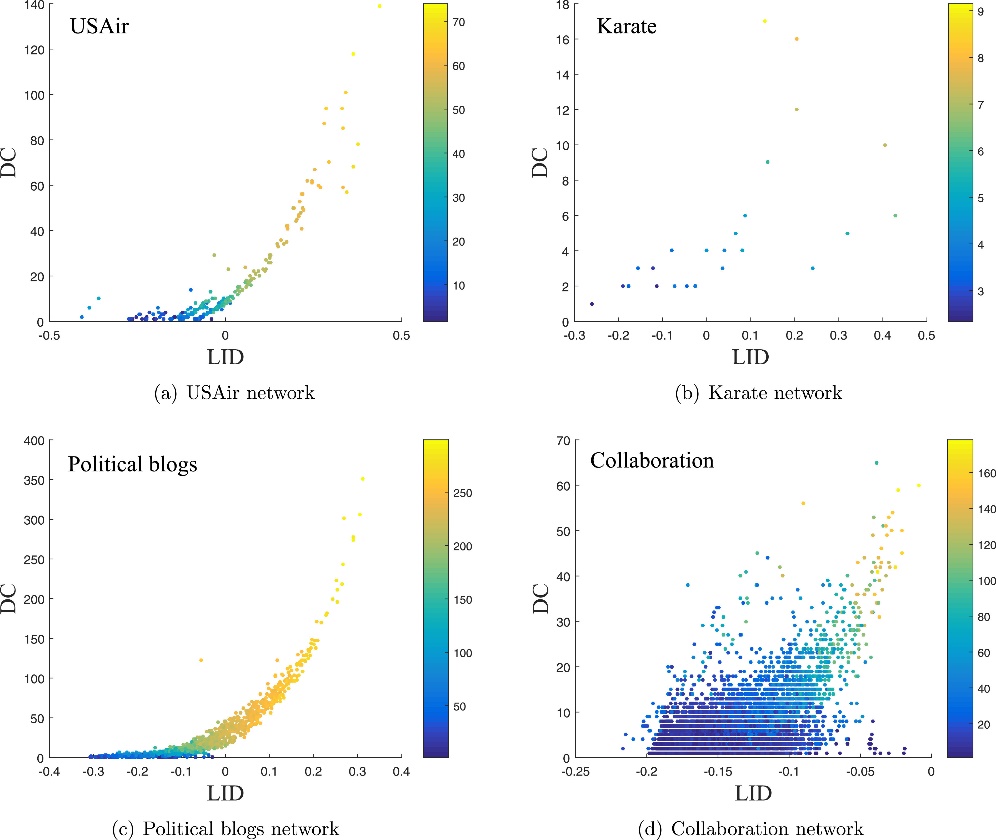


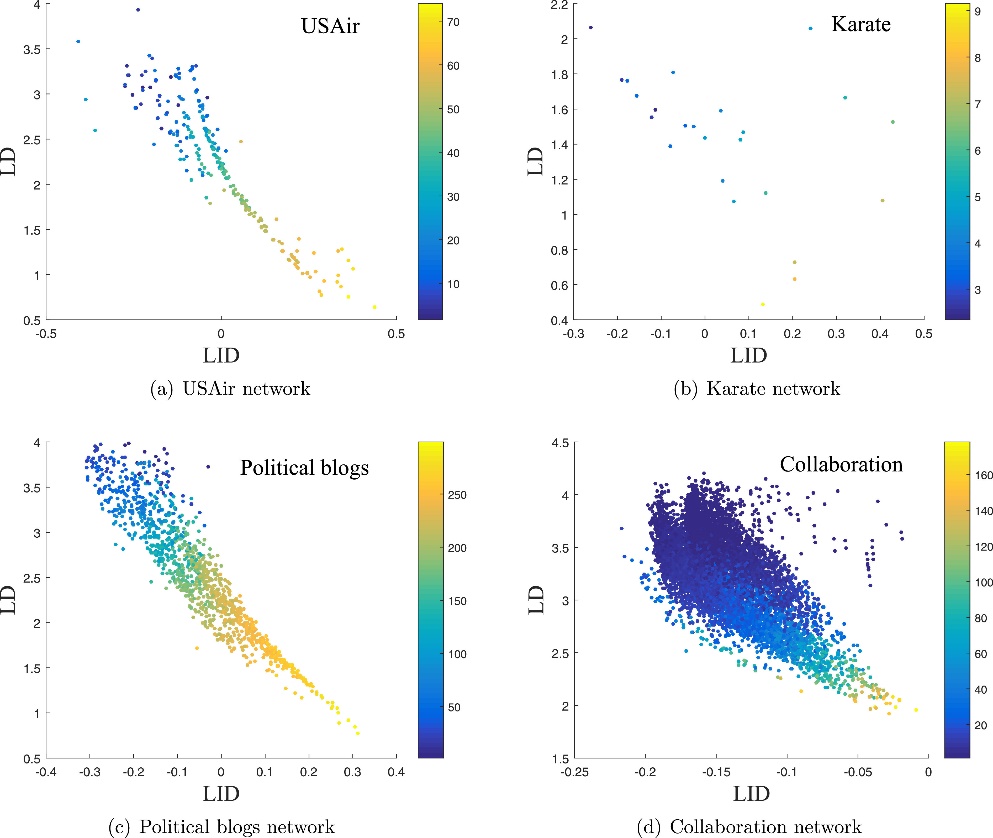


总的来说，在大多数SI实验中，从CC、LD和LID对不同网络的F(t)的观察中，LID的优越性是很明显的。在某些情况下，LID的性能与其他现有方法相似，只是稍有优势。

### Relationships between different measures

为了找到不同测量方法得到的数值之间的关系，提出了不同方法之间的关系图。选择DC和LD比较。在关系图中，每个点代表复杂网络中的一个节点。横轴和纵轴上的数值分别代表每个节点的LID分数和每个节点的DC或LD分数。SI模型中，点的颜色代表所选节点在10个时间步长内的感染能力（F(10)）（由100次独立实验获得）。当一个节点具有大的LID和大的DC或LD时，这两种方法具有正相关；当一个节点具有大的LID和小的DC或LD时，会得到负相关。详细结果见图4和图：

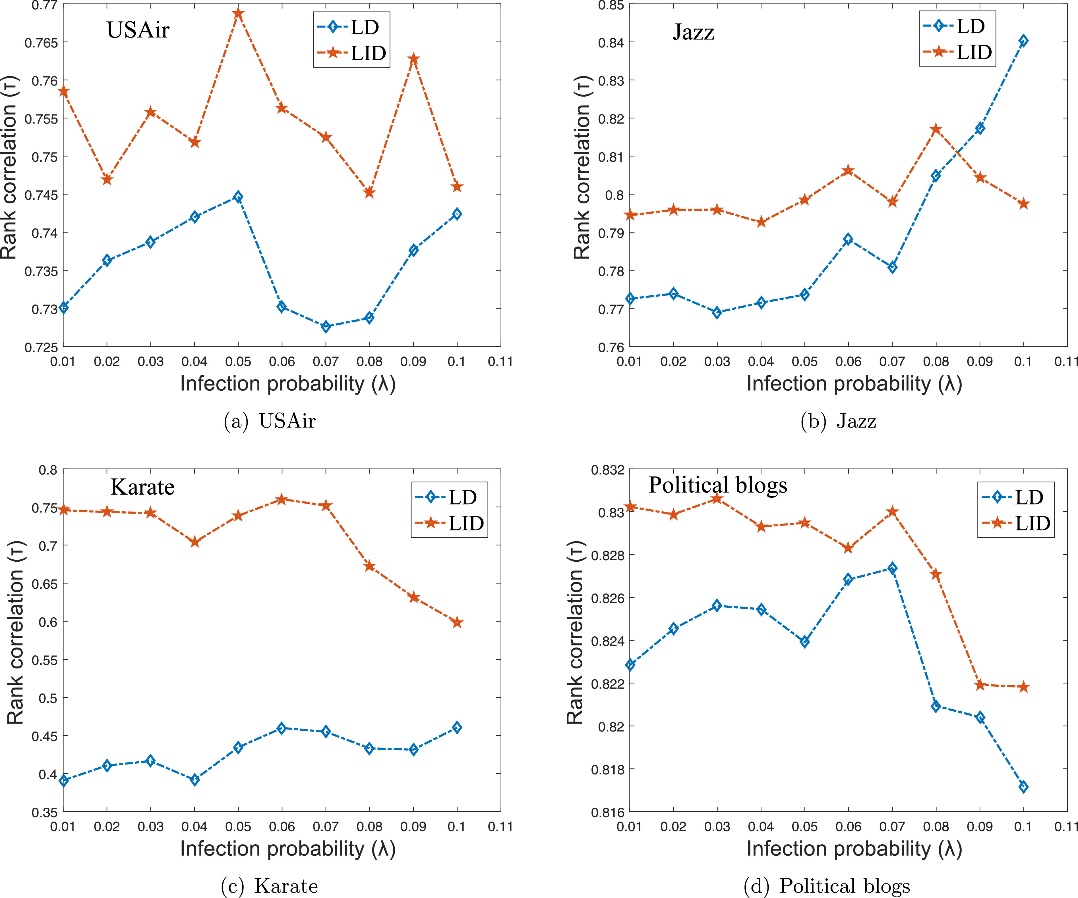




总之，所提出的方法与经典的度量方法相同，度量值越大，说明感染能力越强。在不同的现实世界的复杂网络中，LID可以与其他中心性度量取得稳定的相关性能。此外，LID可以更有效地识别小中心度的节点的重要性。

### Kendall’s tau coefficient

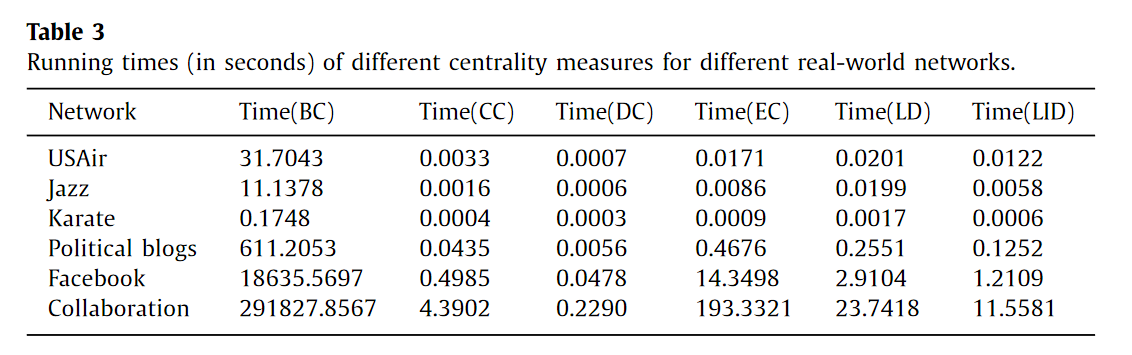
LID和LD之间的Kendall's tau系数和感染能力进行比较。每个节点的感染能力由SI模型的10步内的感染节点数表示。此外，在这个实验中还考虑了不同的情况。SI模型中的传播率从0.01到0.1变化，以考察。感染过程独立地重复100次，是通过平均得到的。较大的表明感染能力和中心性度量之间的关系更相关。图6显示了四个真实世界复杂网络的Kendall's tau系数τ的结果。



总之，在大多数情况下，LID与感染能力之间的要大于LD与感染能力之间的。这说明所提出的方法得到的结果与经典的感染能力更相关，而且在大多数现实世界的复杂网络中，LID比LD能保持相对稳定的相关性能。因此，从这个角度来看，所提出的方法对于识别影响者更加有效。

### Time consumption

介绍了不同网络的不同措施的时间消耗。所有的中心度量都是在配备了英特尔酷睿i7-5500U中央处理器（CPU）的个人电脑（PC）上用MATLAB 2016a计算的，CPU的运行频率为2.40GHz，随机存取内存（RAM）为8GB。运行时间较短的方法具有较低的计算复杂性。这些措施的运行时间列于表3。



总之，LID的运行时间比其他大多数方法要低，这意味着LID降低了计算复杂度。

## 1.4 结论

在本文中，复杂网络中的影响者是通过LID来识别的。覆盖中心节点的box的大小从1增长到，box内的节点数用香农熵表示，香农熵可以衡量box内的信息。

然后，中心节点的LID可以通过box信息和box大小之间的相关性得到。

最后，节点影响他人的能力可以根据LID的值来排序。

由于box大小的增加规则，所提出的方法考虑了中心节点周围的准本地信息并降低了计算的复杂性。与现实世界网络的现有措施相比，所提出的方法更加有效和合理，实验结果证明了其优越性。然而，它还可以改进。一个不可避免的问题是，当两个传播者的LIDs （或其他措施的结果）相等时，如何识别他们影响他人的能力。在今后的研究中，可以将框内的信息视为多尺度的，这样可以实现对信息的充分考虑，可以得到更好的结果。因此，基于维度的方法框架在识别复杂网络中的影响者方面会有明显的改进。