目 录

[1. The identification of crucial spreaders in complex networks by effective gravity model 2](#_Toc105695292)

[1.1 基本信息 2](#_Toc105695293)

[1.2 论文内容 2](#_Toc105695294)

[1.3 实证分析 7](#_Toc105695295)

[1.4 结论 23](#_Toc105695296)

## 1. The identification of crucial spreaders in complex networks by effective gravity model

通过有效引力模型识别复杂网络中的关键传播者

### 1.1 基本信息

期刊：*《Information Sciences》*

发表时间： 8 August 2021

作者： Shuyu Li, Fuyuan Xiao

作者信息：

School of Computer and Information Science, Southwest University, Chongqing 400715, China

School of Big Data and Software Engineering, Chongqing University, Chongqing 401331, China

分区：



### 1.2 论文内容

**一、摘要**

重力模型是识别复杂网络中影响者的一种特殊方法。然而，它涉及一个开放性的问题，即如何确定互动的范围。此外，在传统方法中，质量仅仅由节点的程度来表示。为了解决上述两个问题，本文提出了一个基于精确半径和数值信息的有效重力模型。粗略的截断半径被精确计算。而代表节点传播能力的价值信息则被修改为质量。简而言之，节点的影响范围和价值分数是根据每个节点的属性和网络中邻居节点的互动情况来计算的。与其他类似的方法和最先进的衡量标准相比，本文的方法的合理性和优越性通过在11个真实世界网络上的6个实验得到了证明。

**二、引言**

一般来说，识别重要传播者的措施大致分为不同类别，其中包括邻域中心性、路径中心性、特征向量中心性和分形维度。度数中心性（简称DC）、H指数和K壳分解是基于邻域中心性的特色方法。两个有代表性的基于路径的方法是CC和BC。EC和PC是基于特征向量中心度的代表性方法。基于分形维度的典型测量方法是基于信息熵的最新局部信息维度（简称LID）和基于模糊集的模糊局部维度（简称FLD）。尽管如此，上述指标也有一定的局限性。DC只考虑一个节点的局部信息。虽然k-shell算法将网络中的节点划分为不同的层次，但这种方法的缺点是不能对同一层次的节点进行排序。BC和CC侧重于全局结构，这导致了模型的高计算复杂性。由于EC仅仅可以用于对称网络，所以它有较大的局限性。虽然LID和FLD区分了与中心节点不同距离的节点的贡献度，但有时测量的准确性并不高。近来，在这一领域采用了一些先进的技术方法。例如，Tang等人提出了一种离散洗牌的蛙跳方法，该方法结合了确定性和随机搜索策略来实现影响力最大化。作为一种基于种群的优化方法，Zareie等人提出了灰狼算法来研究最大影响的问题。Pelusi等人提出了一种基于双曲函数的引力搜索算法，以寻找探索和开发之间的最佳平衡。Zareie等人提出了一种有影响力的营销者用户检测算法，以识别社交网络中最具影响力的用户，该算法考虑了用户对信息的兴趣。

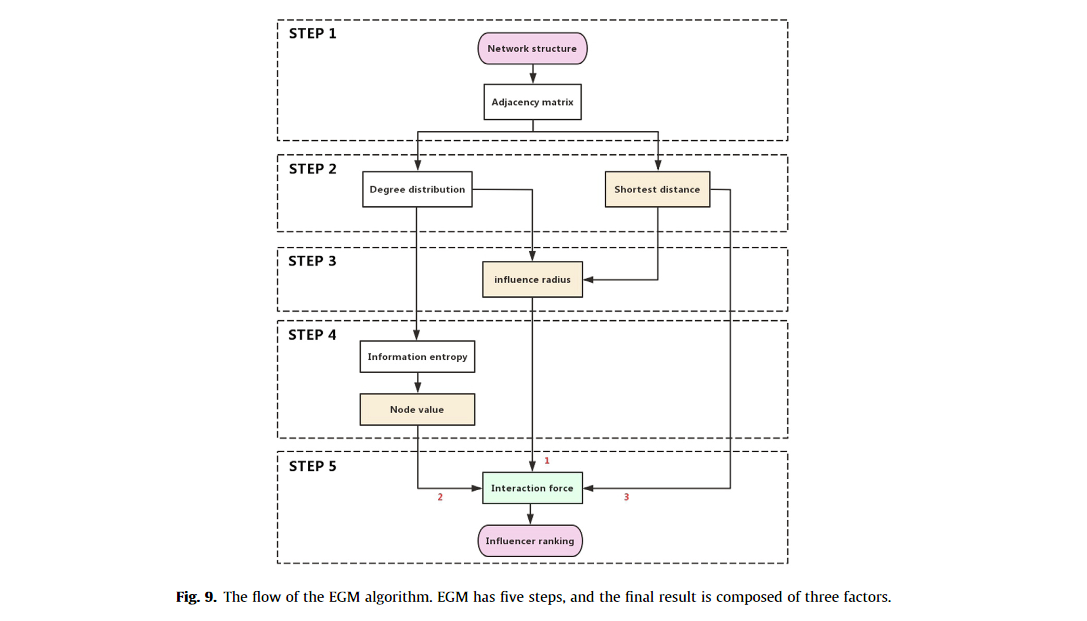
最近，Li等人提出了一种使用重力模型来发现复杂网络中的重要节点的新方法，重力中心性模型（简称GC）的灵感来自于万有引力定律。节点的度数被视为质量，而节点之间的平均路径被视为距离。Liu等人[31]做了进一步的改进，为每个节点增加了一定的权重。此外，Bi等人[32]提出了一个考虑使用多个全局信息的时间引力模型。在这个模型中捕捉到了网络的时间演变。Li等人提出的最新的广义引力模型[33]（简称GGC）准确地使用节点的传播能力作为质量。尽管如此，要正确确定这一措施的参数是很复杂的。

值得注意的是，在上述基于重力的度量中仍然存在两个共同的问题。一方面，节点的截断半径被设定为网络平均距离的一半。这意味着对于网络中的每一个节点，它们的影响范围都是一样的，这不是一个合理而精确的考虑。另一方面，一些传统的方法只是简单地将节点的度数视为质量，这将忽略一个节点的邻居信息。本文提出了一种新的方法，即有效引力模型（简称EGM），它基于精确的半径和数值信息来解决上述的两个问题。本文的亮点在于可以精确计算网络中每个节点的模糊影响范围。因为我们需要得到一个节点在网络中最远的影响力，所以用节点和它最远的节点之间的关系来得到它。此外，本文的另一个贡献是，创新地定义了作为节点质量的价值信息。具体来说，一个节点的度和它的邻居在其本地网络中的度分布信息都被考虑在内。总而言之，本文对影响半径和节点质量进行了修正，可以很好地识别复杂网络中的关键节点，这一点通过在11个真实世界网络中的6个不同实验得到了证明。

本文其余各章的结构如下。第2节简要介绍了复杂网络的初步知识，包含了经典的代表措施和最先进的方法。第3节介绍了一个有效的引力模型，它基于精确的半径和价值信息。在第4节中，与其他类似的方法相比，EGM的可靠性以及与最先进的方法相比的优越性在11个真实网络的6个不同实验中得到了证明。最后，第5节讨论了关于EGM措施的结论和反思。

**三、模型解析**

为了修正近似的截断半径，本文根据节点本身和其最远节点之间的关系来定义影响范围。更重要的是，为了选择合适的节点质量，信息熵被用来评估本地网络的程度分布。基于上述分析，本文提出了一种叫做有效引力模型EGM的方法，它是基于精确的半径和价值信息来识别复杂网络中的影响者。EGM由以下五个过程组成，其流程图如图9所示。



**第1步：建立网络**

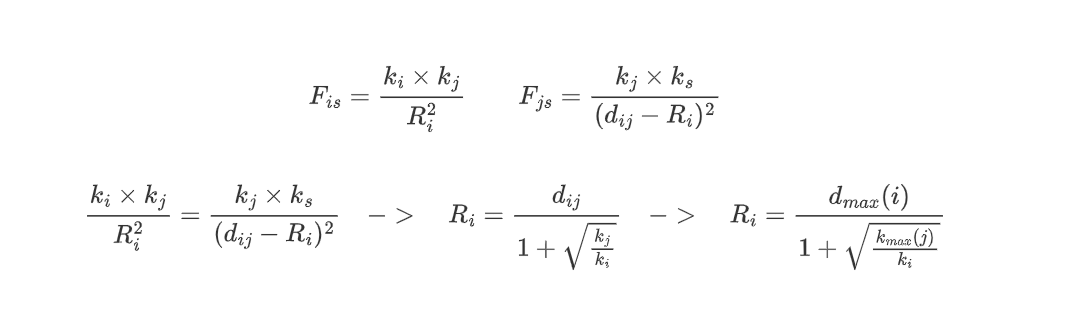
**第2步：获得节点的度分布和网络的距离矩阵**

源点和终点之间的最短路径是包含最少边数的路径。在本文中，如果两个节点之间没有路径，它们的最短距离就被分配为无穷大。而节点到自身的距离被设定为零。

**第3步：计算出准确的影响半径**

在这一部分提出了一种新的方法来计算每个节点的影响范围。本文认为，事实上每个节点的影响半径是不同的。由于EGM需要探索一个节点的最大影响范围，所以要考虑一个节点和它的最远点之间的关系。

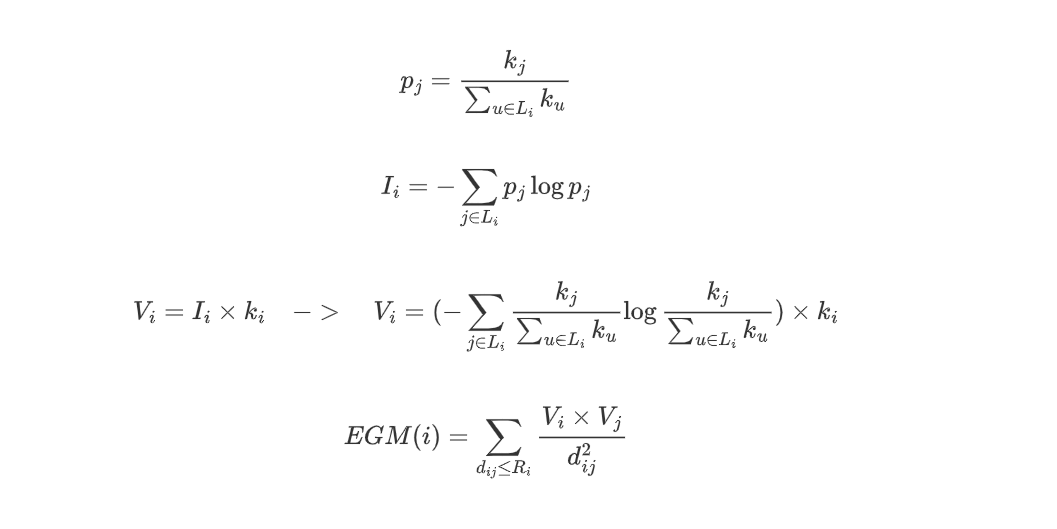
对于节点i来说，我们假设在节点i和离i最远的节点j之间的路径上存在一个虚拟节点s，其中来自节点i和j的影响力是相等的。这是因为当某个节点x位于节点i和节点s之间时，即x∈(i, s)， x只受节点i的影响；同样，当节点x位于节点s和节点j之间时，即x∈(s, j)， x只受节点j影响。受万有引力公式的启发，一个节点对周围节点的影响与自身的信息成正比，与距离的平方成反比。因此，给定一个网络，截断半径的计算过程如下：



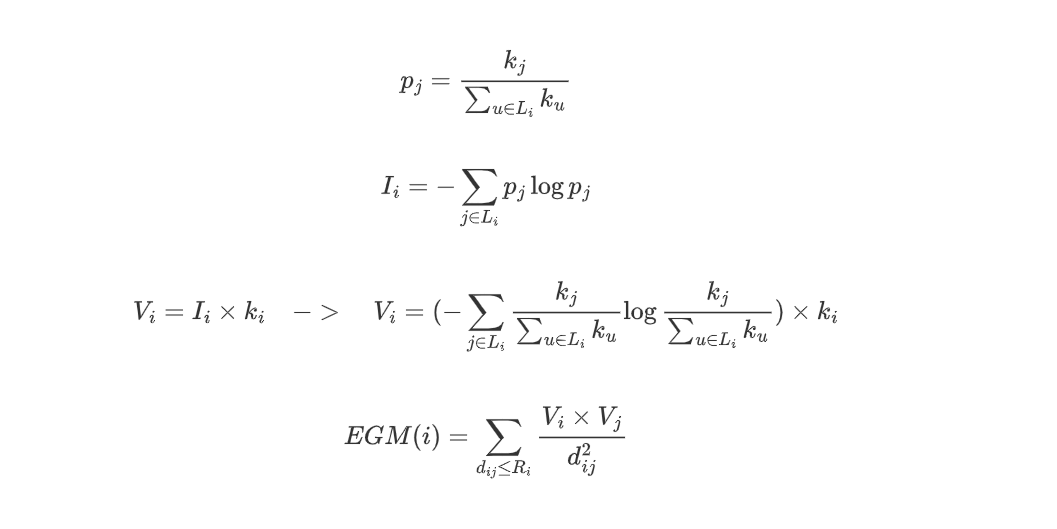
其中Fis是节点i对分界点s的影响力，Fjs是节点j对分界点s的影响力，各自用节点i和节点j的度以及两点间的最短距离表示，联立两个方程，可以解出节点i的影响半径Ri的表达式。又因为j是离i最远的节点，可以改写成最后所示的形式。特别是，如果有多个节点与i的最远距离相同，这些节点构成最远节点域Fi =（k1，k2，...， ks）。最远节点域中节点的平均度数表示为. kmax(i)将被kmean(i)取代。

**第4步：计算节点的价值**

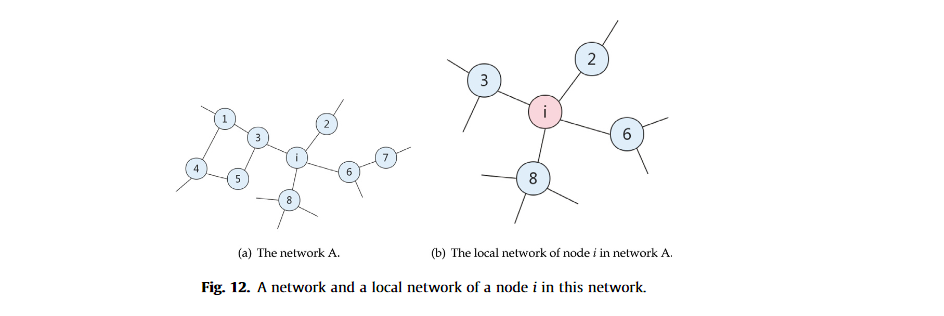
节点的价值，评估了节点的本地信息，在这部分提出。在我们看来，一个节点的价值取决于影响整个网络的能力。从局部来看，一个具有较高影响力的节点一定是具有较高程度的节点，也就是有更多的邻居。而从影响力传播的角度来看，如果一个节点更重要，它就有更大的概率向外扩散到多个潜在方向。换句话说，如果一个节点对传播方向的选择有限，那么这个过程就更有可能受到阻碍。对于一个具有一定本地影响力的节点，即在本地网络中具有一定的度数，本地网络的度数分布对节点的重要性起着重要作用。如果邻居的度数分布比较均匀，信息传播的不确定性就会增加。基于上述分析，信息熵可以用来从信息传播的角度衡量信息的价值，可以很好地描述社会影响的不确定性。因此，本文将同时考虑自身和邻里信息的节点的价值定义如下。



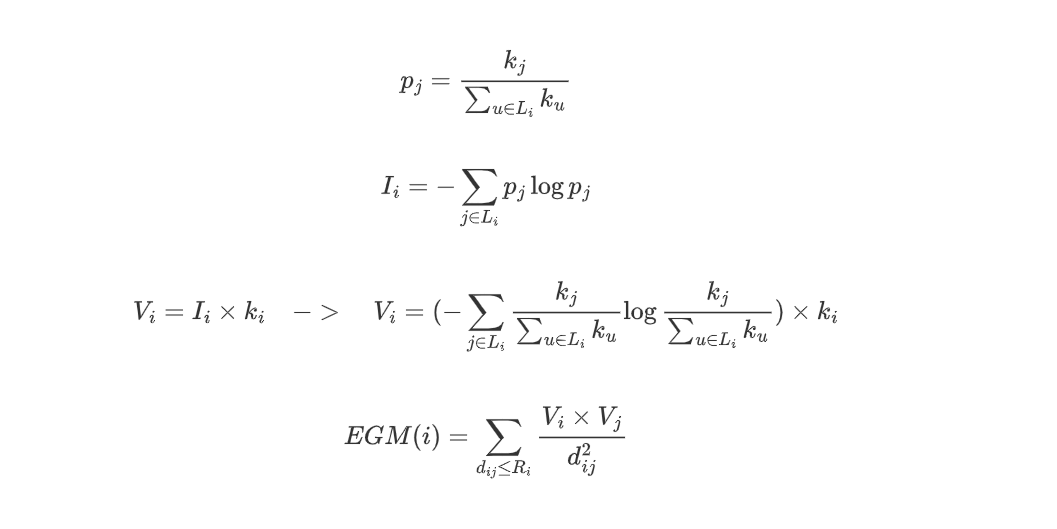
其中，Vi定义为节点i的值, Ii表示节点i的信息熵，ki表示节点i的度。每个节点的信息熵包含两个内容，分别是节点本身的特性和邻居节点的属性，比较合理和全面。节点i的信息熵的定义如下。



其中pj表示信息包含在节点i的局域网络中的概率，它是节点j的度数与它的局域网络中的度数之和的比率。节点i的局域网络用Li表示，包括节点i、其邻居的节点和它们的度。图12显示了Li的一个直观的例子。

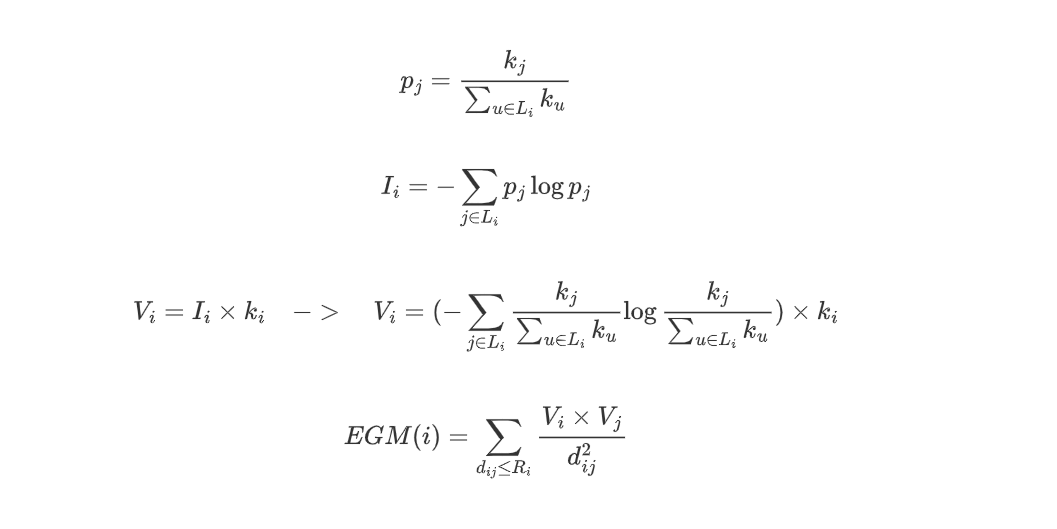


因此，pj可以通过以下方式获得。



其中kj表示节点j的度数，而表示涉及节点i的局域网络中所有节点的度数之和。

总之，节点i的价值可以写成以下形式。



其中包括属于节点本身和邻居节点的两部分信息。节点的信息由度反映，而邻居的信息则由信息熵代表的邻居节点的度分布来考虑。

**第5步：获得影响节点的排名**

受万有引力定律的启发，基于修正的精确半径和数值信息的EGM定义如下:



其中，Vi和Vj分别代表节点i和节点j的价值。dij表示i和j之间的最短距离，Ri是节点i的影响半径。

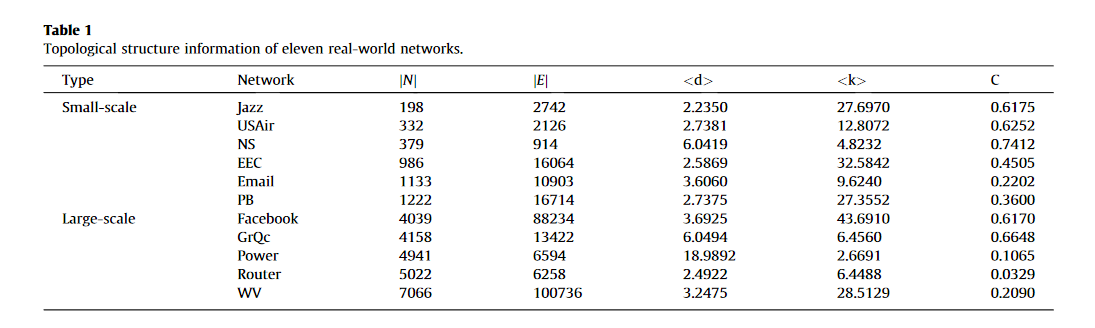
总之，在以前的方法中，模糊的影响半径被设定为平均路径长度的一半，在本文中被精确的影响半径所取代。同时，节点的价值信息被视为质量，而不仅仅是节点的度。更重要的是，复杂网络的局部信息和全局信息在EGM中得到了充分的考虑和结合。具体来说，对于局部信息的考虑，EGM通过充分利用节点在局部网络中的度分布信息来衡量节点的价值，并通过信息熵来评价节点的传播能力。对于全局信息的组合，EGM通过考虑节点与其最远节点之间的交互作用来定义半径范围，并通过引入每个节点之间的最短距离来评价其重要性。总之，修改后的精确影响范围和考虑到环绕度分布的节点的价值信息构成了本文提出的EGM。

### 1.3 实证分析

1. **数据来源**

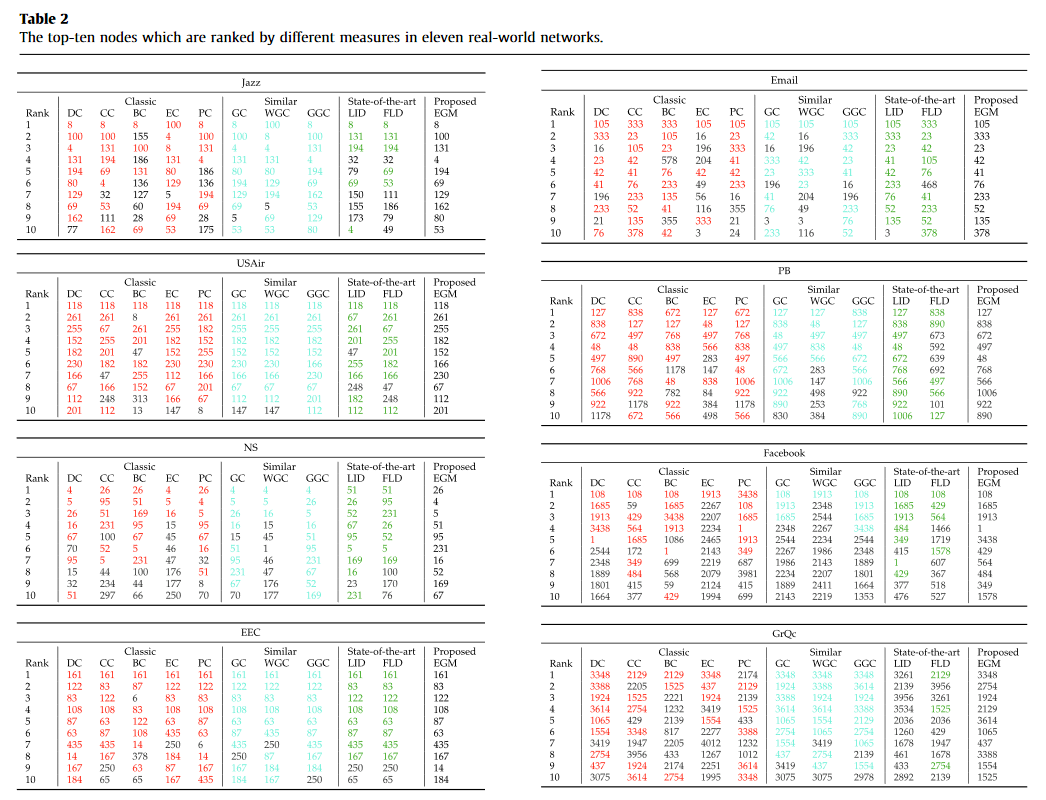
本文使用的数据集分为六类，包括社交网络、协作网络、通信网络、交通网络、基础设施网络和技术网络。

表1显示了包含11个不同类型网络的详细信息。Jazz、USAir、NS、EEC、Email和PB被划分为小规模网络。Facebook、GrQc、Power、Router和WV被划分为大规模网络。



1. **相似性实验：TOP-10**

本实验的目的是验证所提出的方法EGM，包含DC、CC、BC、EC、PC的经典方法，包含GC、WGC、GGC的类似方法，以及包含LID和FLD的最先进措施之间的相似性。表2显示了上述方法在11个网络上进行的真实实验所产生的前十名节点的排名表，它反映了不同度量之间的相似性。用区别性的颜色来描述由不同测量方法获得的相同的前十名节点。例如，如果EGM排名表和另一种方法产生的排名表中有相同的节点，那么这些相同的节点将被涂上颜色。相反，黑色表示由EGM和另一种方法获得的不相似的节点。





如表2所示，DC、EC、GC、WGC和GGC获得的前十名节点与EGM在Jazz中获得的节点大致相同，所有的方法都与EGM有较高的匹配度，包括在USAir中EGM与它们的相似度较大。在NS中，除了EC和WGC之外，每个措施之间的相似度都比较高。对于EEC，EGM得到的top10节点完全等同于DC所拥有的节点，这说明它们之间的匹配度比较高。CC和EGM产生的结果在Email中是完全相同的。在PB中，LID和EGM的topten节点是相等的。在Facebook中，EC和WGC与EGM的匹配度最低。在GrQc中，LID和EGM之间的差异最大。对于Power，各种方法之间的相似性并不明显。在Router中，CC、GC、GGC和FLD的结果与EGM大致相似。值得注意的是，DC的结果与EGM在WV中的结果完全相同。

综上所述，除Power外，每种方法在每个网络上的前十名节点的排名与EGM的排名大致相同，这表明本文的方法与大多数现有的测量方法相似，并证明了其可靠性。

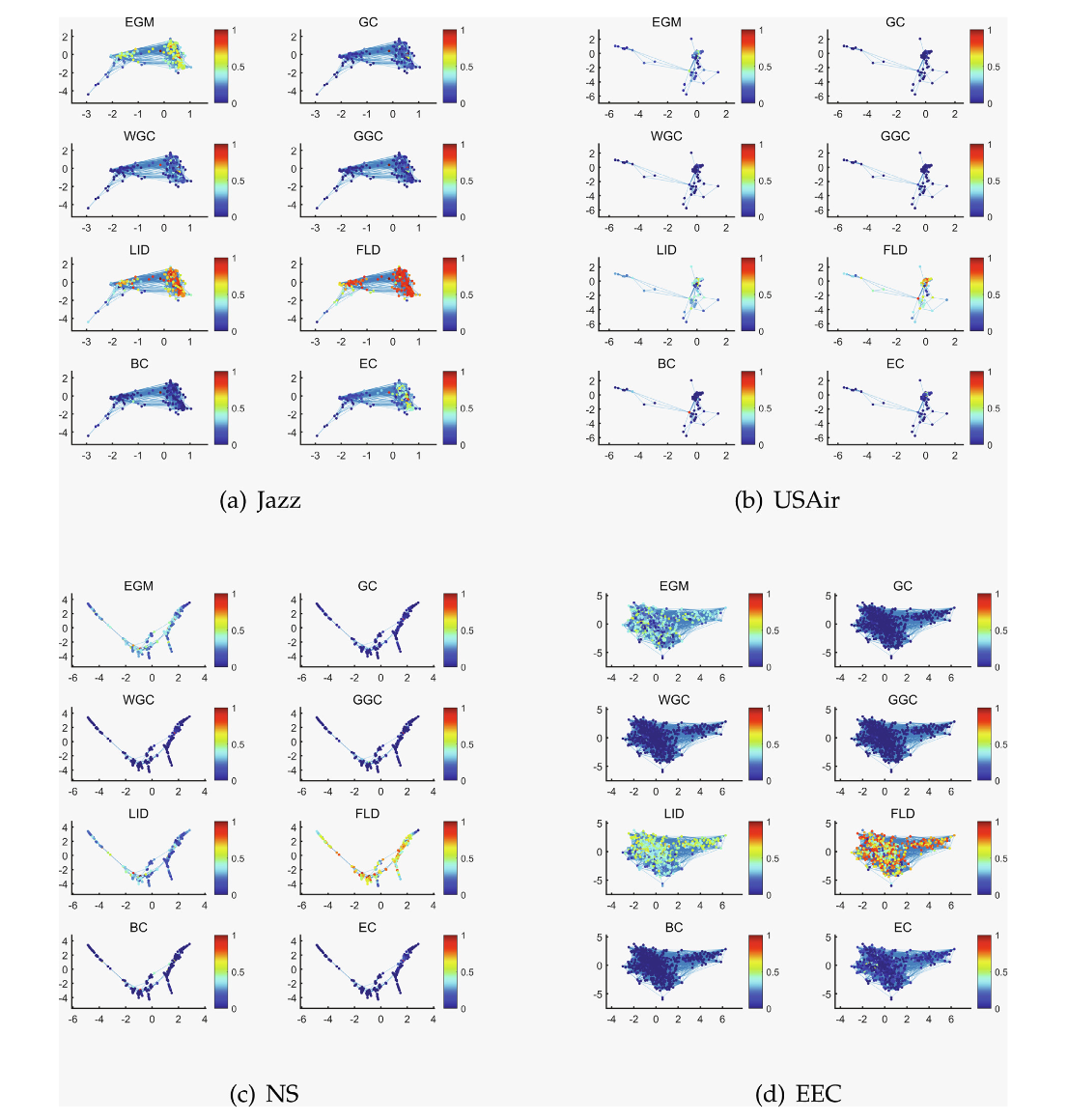
1. **有效性实验**

有效性实验分为四个章节，分别是中心性得分、分离度、SI模型、Kendall系数。

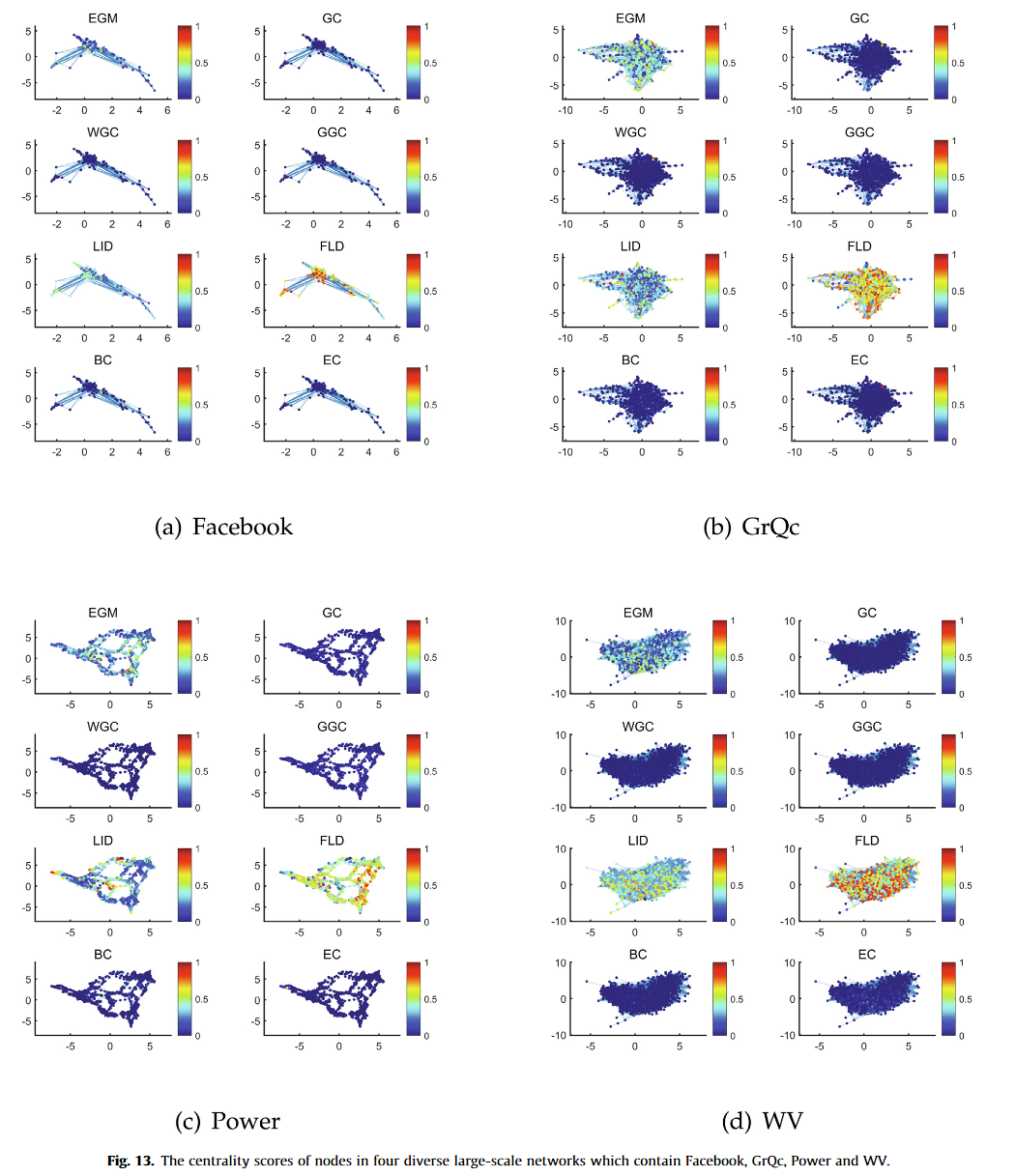
1. **中心性得分**

本实验的目的是计算节点的标准化中心性得分，结果以热图的方式显示。在本实验中，类似的GC、WGC、GGC、最先进的LID、FLD和传统的BC、EC测量方法被用来进行比较。图10-13显示了每种方法的相对重要性分布，其中颜色代表节点的重要性。如果一个节点的颜色较深，说明该节点的地位较高。如图10-13所示，所有措施得到的节点的相对重要性分布几乎相同。

图10和图11显示了在六个小规模网络上的结果。在Jazz中，FLD获得的中心性分数太大。EGM和LID都能更好地区分节点。在USAir中，FLD和LID的表现更好。对于NS，EGM和LID有类似的结果。而EGM、LID和FLD可以很好地区分节点。EGM、LID和FLD在EEC中具有很强的区分能力。在Email中，EGM的效果仅逊于FLD。在PB中，FLD计算的分数过大，EGM和LID的表现更好。

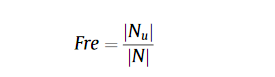


四个大型网络的实验结果如图13所示。对于GrQc，EGM的性能只比FLD差一些。EGM、LID和FLD可以区分Power和WV中节点的重要性。总之，EGM在所有网络上的完成度都优于GC、WGC和GGC等类似方法，这表明我们的方法对重力模型的改进对识别影响者是有效的。同时，EGM、LID和FLD的区分能力也超过了其他方法。

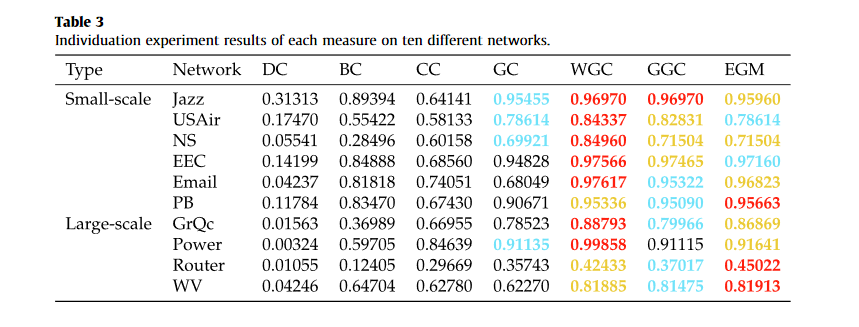


1. **分离度**

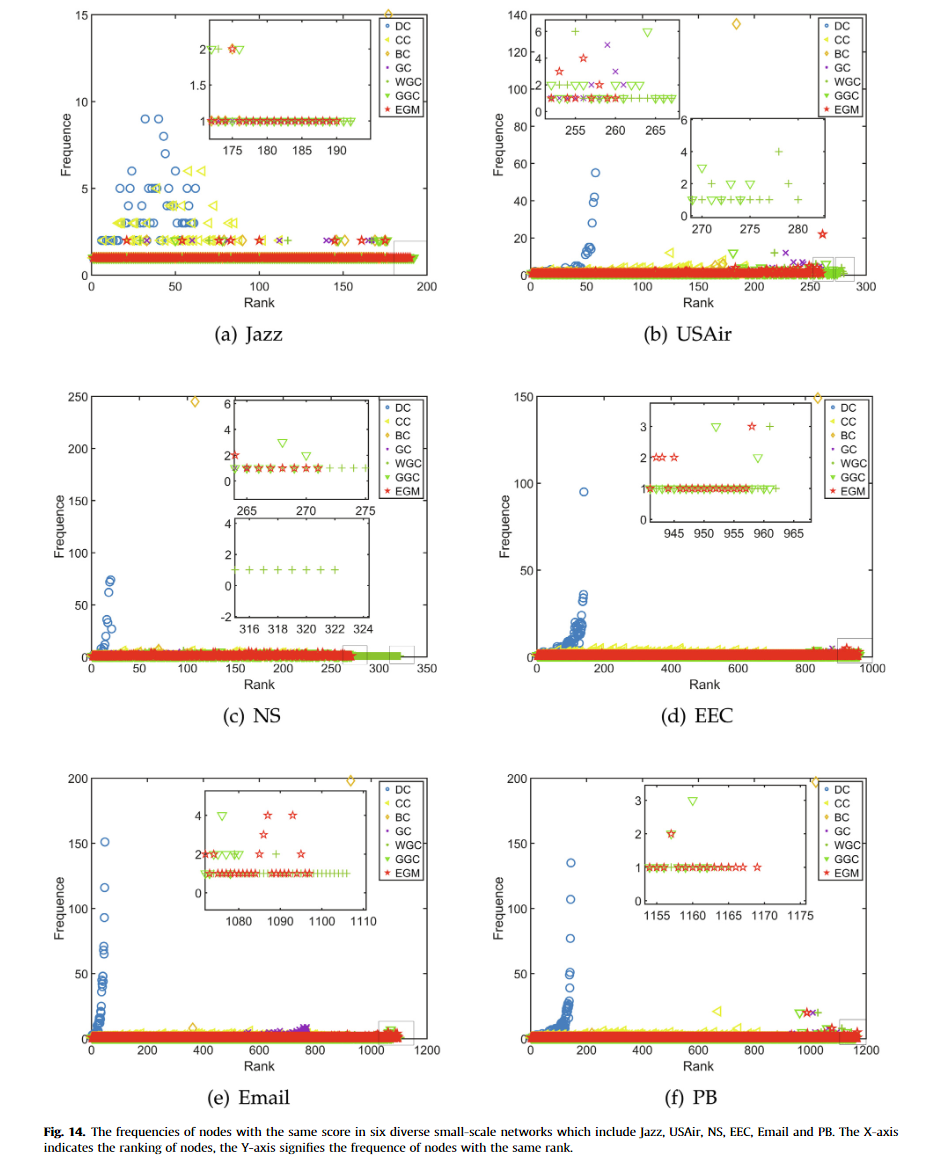
为了比较不同测量方法对节点的区分能力，采用了分离度实验，计算每种方法在10个网络上得到的具有相同分数的节点的频率。分离度被定义为以下公式：



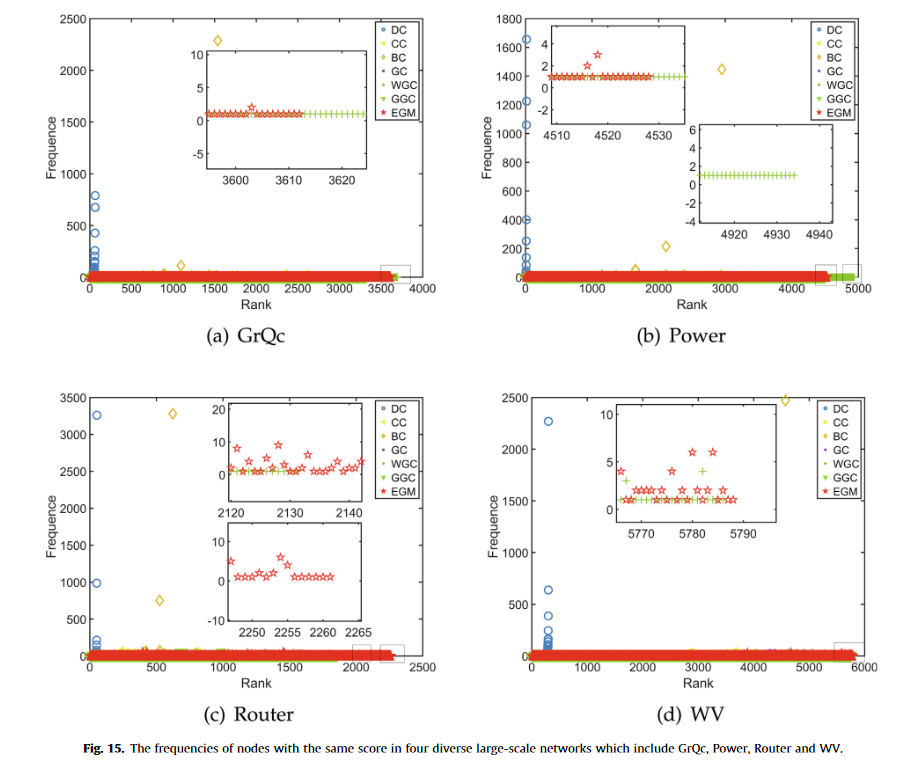
其中|Nu|表示具有唯一分数的节点数，|N|是整个网络中的节点数。在这个实验中，如果一种方法获得了更多的排名，那么它的效果更好。此外，如果一种方法有更高的分离度，也证明这种方法有更强的区分能力。通过对10个网络的区分实验，图14和图15显示了通过不同措施获得的节点频率。同时，表3中详细记录了不同方法的区分度，其中前三名的区分度值分别用红色、黄色和蓝色表示。



从图14中可以看到，对于Jazz来说，WGC、GGC和EGM在同一排名中的节点较少。对于USAir，虽然EGM的表现不是最好的，但它击败了DC、CC和BC。在NS和Email，EGM的表现只比WGC差。对于EEC，EGM的区分能力排名第三。在PB中，EGM获得的每个排名中的节点频率是最低的，排名区间是最大的。



如图15所示，EGM在GrQc和Power中打败了除WGC以外的所有方法。在包括Router和WV在内的两个大规模网络中，EGM获得了每个排名的最低频率和最大跨度。此外，从表3可以看出，EGM的区分能力在所有网络中都排名前三。尤其是EGM在PB、Router和WV中获得的频率最低，排名最高，能够很好地分辨出节点的感染能力。



总的来说，通过对图14和图15的分析，我们可以得出这样的结论：EGM的性能只比WGC差，但对于大规模网络，EGM的性能相对更好。

1. **SI模型**

为了验证所提出的方法在传播能力方面的优越性，在十个真实世界的网络上进行了基于SI模型的实验。

在本实验中，选择前面每种方法选取的前10个节点作为初始感染节点，其余的节点被视为易感节点。这些感染性节点将以一定的概率β感染周围的易感节点，该概率被设定为0.1。传播时间用t表示，感染节点的数量F(t)作为一个指标来确定最初选定的节点在某一时间t的感染能力。每个实验独立进行一百次以得到客观结果，得到的平均结果在图16和17中描述。在所有的措施中，如果被某种方法初步选定的前十个节点具有更大的影响力，那么这些节点的感染能力将更强，最终的实验结果将是更多的被感染节点。

图16描述了SI模型在六个小规模网络中的实验结果。在Jazz中，当t≥40时，EGM达到了最佳效果。对于USAir，EGM采购的感染节点数量每时每刻都超过其他方法，这证明了EGM的卓越能力。在NS中，当t = 20时，EGM具有最强的传输能力，这在很大程度上显示了EGM在几乎整个时间段的优势。对于EEC，当25 ≤ t ≤ 40和42 ≤ t ≤ 46时，EGM获得最大的感染节点数。在Email中，当t≥2时，EGM显示出击败其他方法的卓越能力。当3 ≤ t ≤ 25和t ≥ 28时，EGM在PB中具有最强的传播能力。

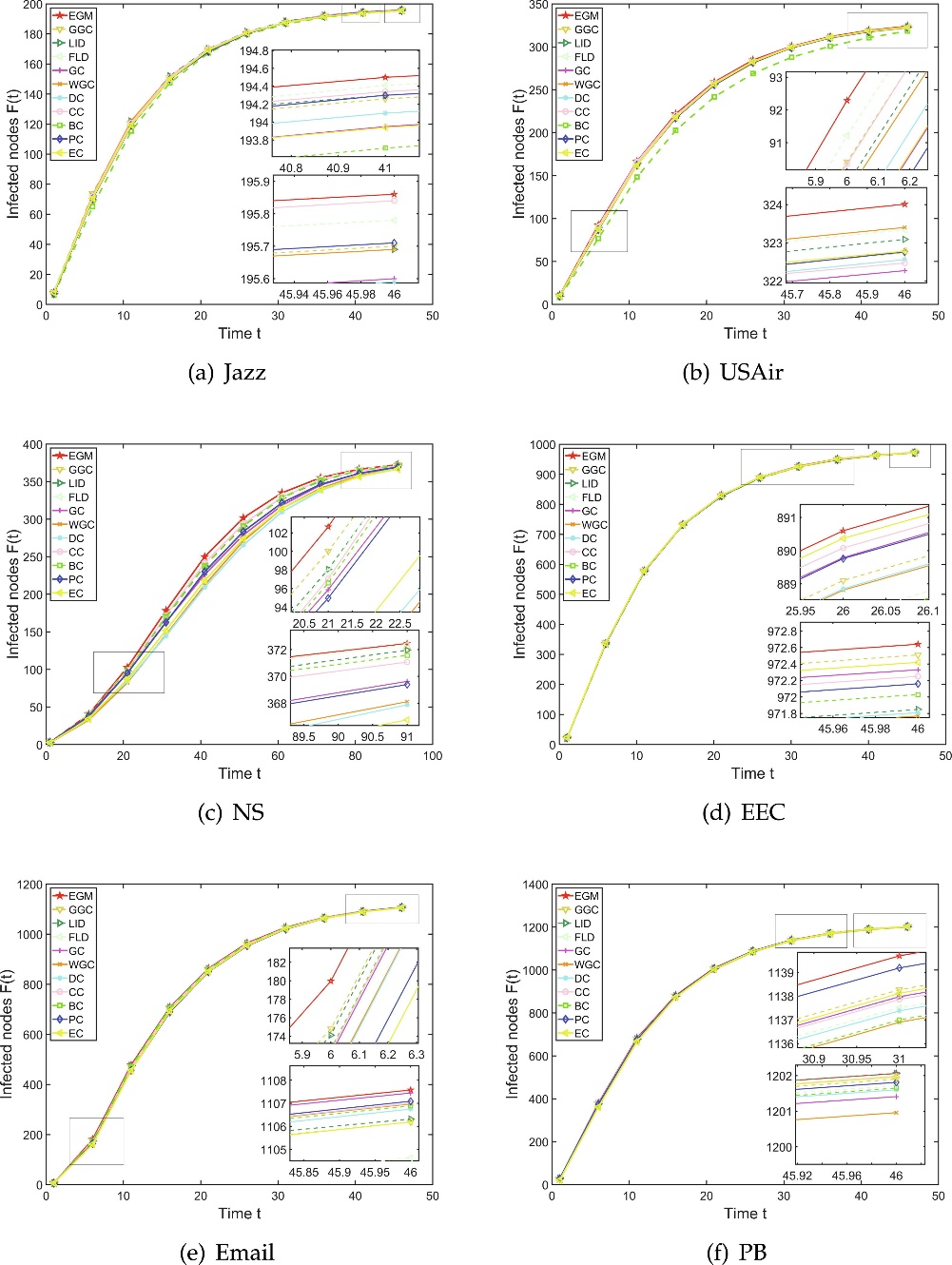
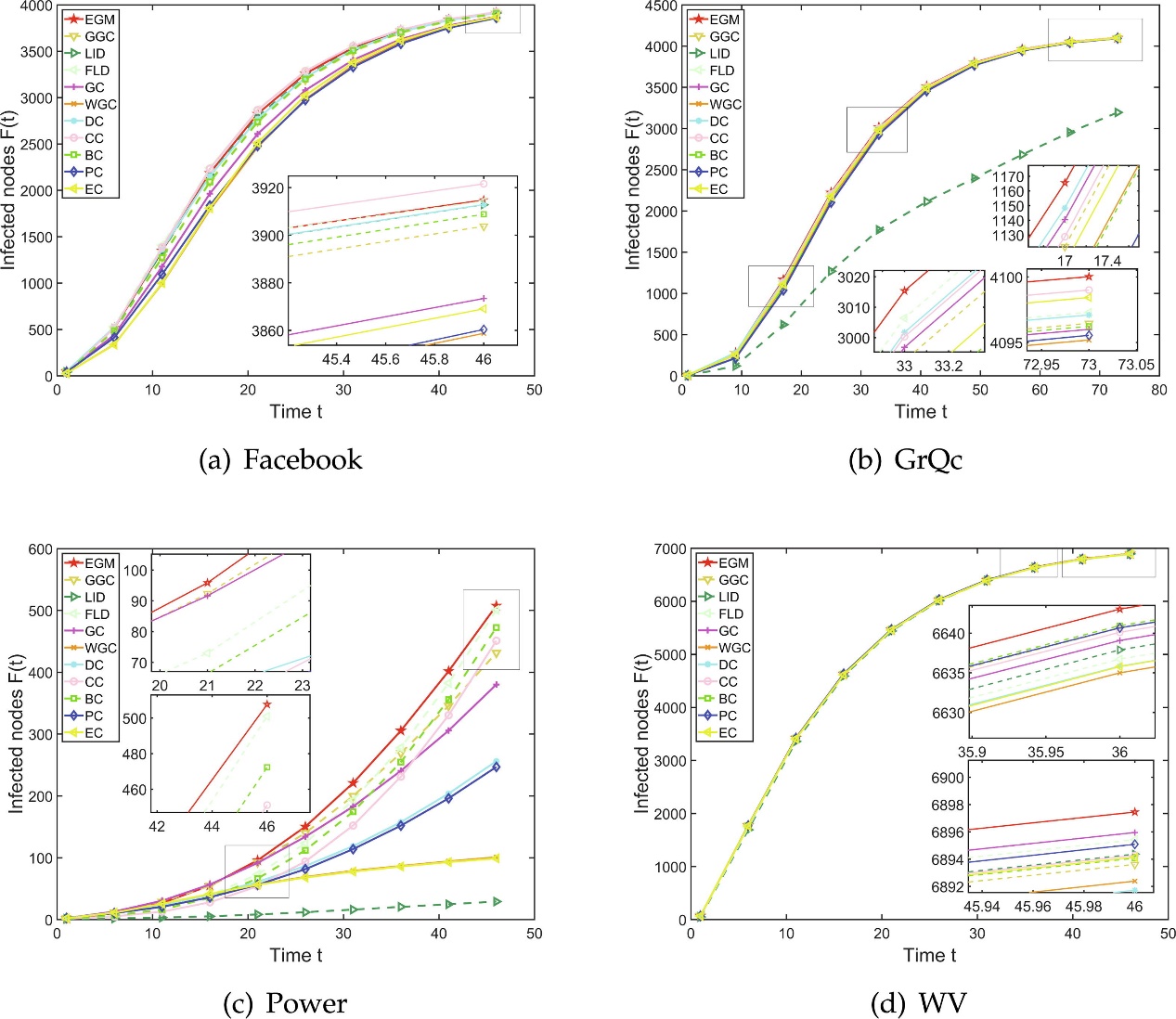
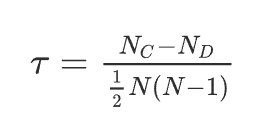


图17描述了在四个大规模网络中的传播能力测试。虽然在Facebook中感染最多的节点不是EGM所有，但它仅次于CC。对于GrQc，当t ≥ 11时，EGM的传播能力优于其他方法。在Power中，当t ≥ 18时，EGM拥有的感染节点数量明显高于其他方法，这表明EGM的优势最强。对于WV，当t ≥ 33时，EGM显示了最好的成绩。可以得出结论，在实验网络中，除了Facebook，EGM在一定时间内拥有最强的传播能力。更重要的是，EGM的表现远远超过了GC和WGC等类似的方法，这表明EGM是迄今为止基于引力模型的最佳测量方法。



1. **Kendall系数**

为了测试所提出的方法的可靠性，用Kendall系数[50]来评估各种中心性方法和SI模型之间的相关性，SI模型是衡量感染能力的标准。Kendall系数s是一个用于衡量两个随机序列之间相关性的统计值，被定义为一致对和不一致对之间的减法与总对数的比率。Kendall系数τ用以下形式表示：



其中NC用来表示一致序列对的数量，不一致序列对的数量用ND表示，序列的长度用N表示。

在本实验中，SI模型中用十步感染的节点数(F(10))来表示每个节点的感染能力。为了使结果更全面，改变SI模型的传播概率β来测试不同的情况。每个实验都独立进行一百次，以获得客观的结果。实验结果的平均情况如图18和19所示。如果一种方法最终拥有较高的τ值，表明这种方法与SI模型的标准度量更为相似，这也证明这种方法在准确性方面有更好的表现

六个小规模网络的实验结果显示在图18中。当感染率β = 0.18,0.19和0.20时，EGM的τ值在Jazz中是最高的。在美国航空，虽然EGM在感染率b 6 0:37时排名第五，但它对类似的算法如WGC有很大的改善。对于NS，当0.11≤b≤0.13时，EGM的成绩是最好的。在EEC中，EGM的τ值只比WGC和EC低。在Email中，当感染率β= 0.45,0.46,0.47和0.51时，EGM拥有仅次于CC的τ值。对于PB，EGM的τ值超过了所有的方法，这表明EGM具有最好的性能。

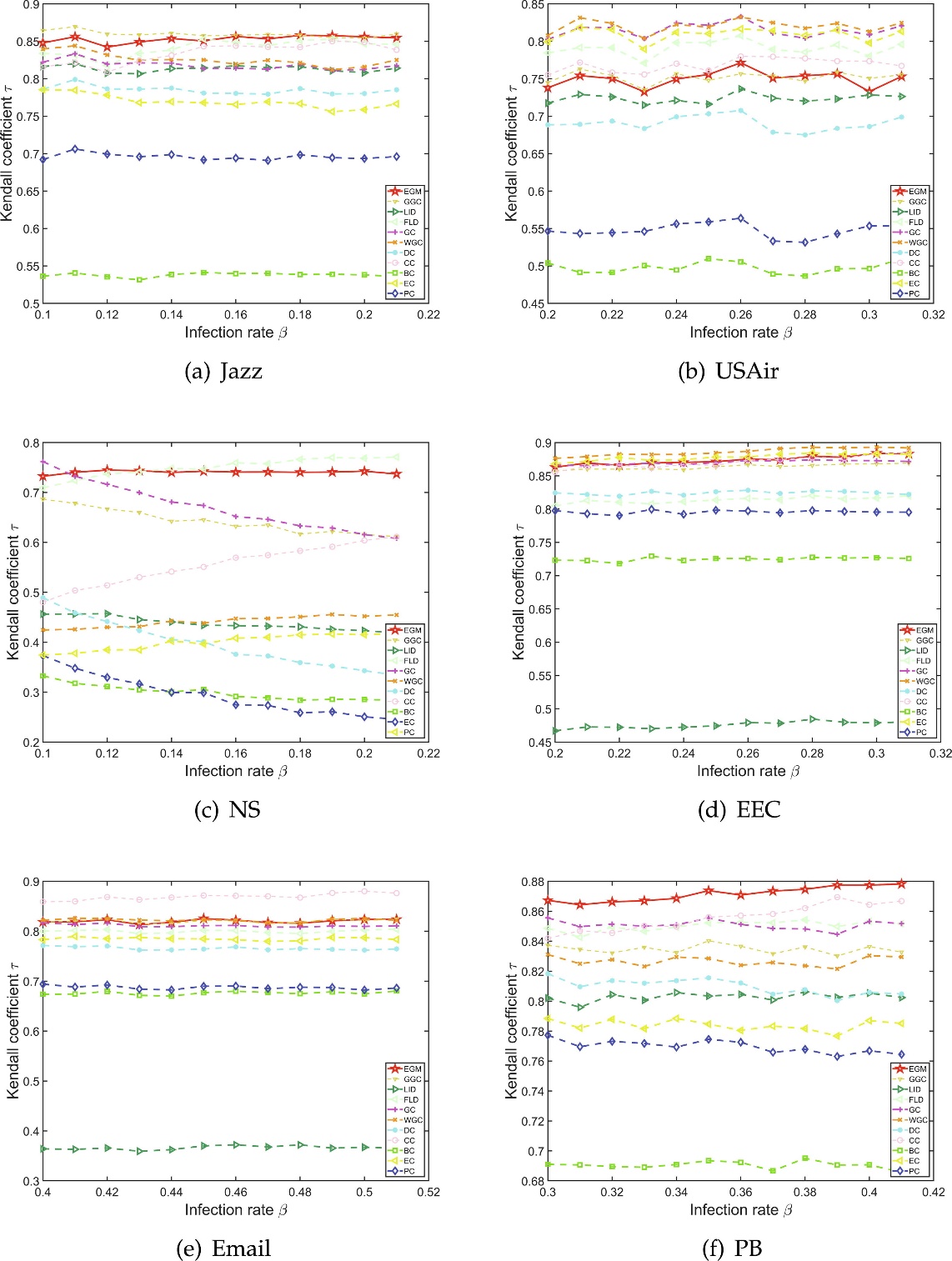
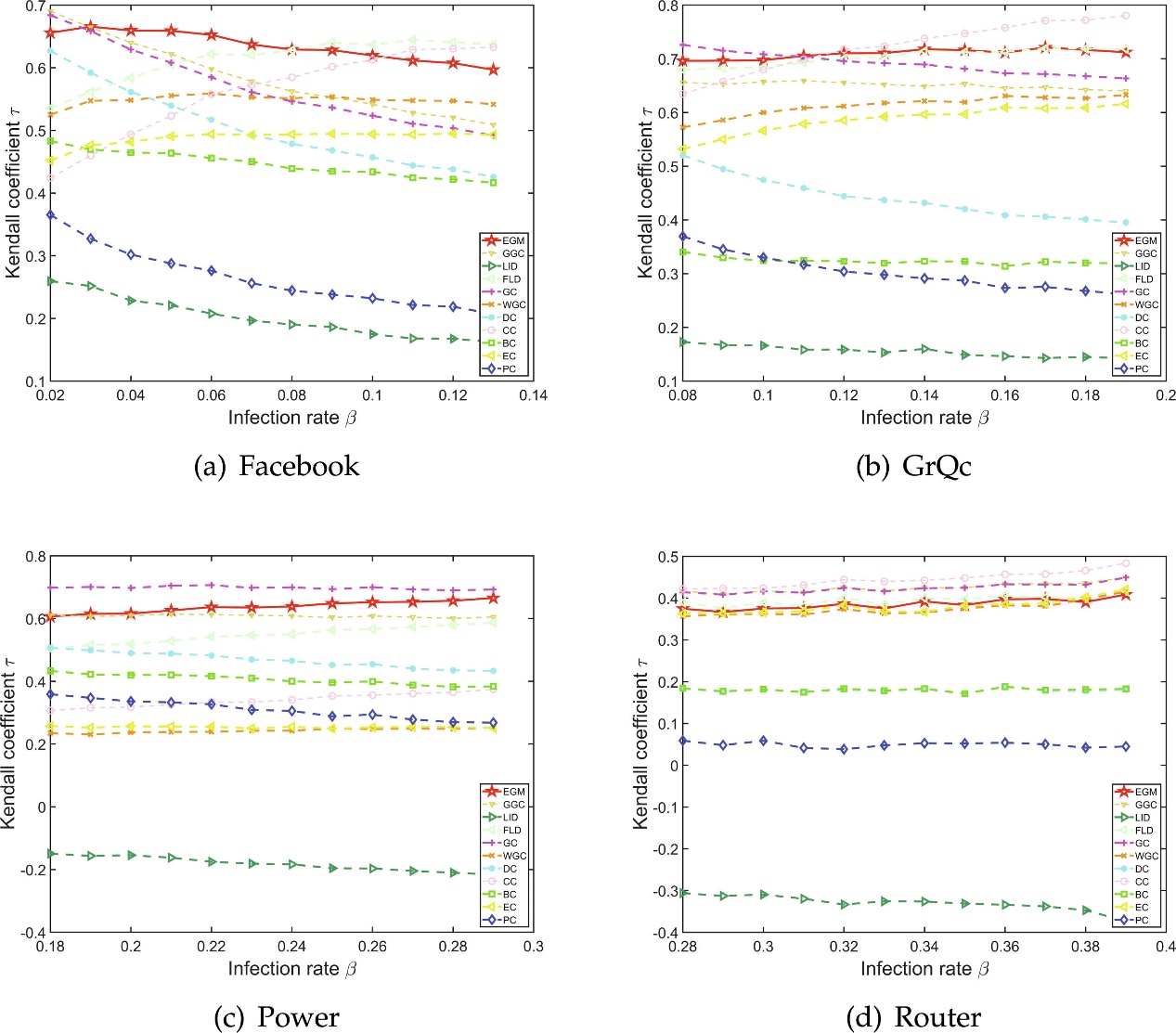


图19描述了在四个大规模网络上的实验结果，在Facebook中，当感染率为0.03 ≤ b ≤ 0.08时，EGM的能力远远优于其他措施。在GrQc中，当β= 0:11时，EGM有最好的表现。在Power中，EGM打败了除GC以外的其他方法。对于Router，当0.28 ≤ b ≤ 0.37时，EGM排名第五。总之，在Jazz、NS、PB、Facebook和GrQc中，在某些感染率下，我们的方法获得了最高的τ值。特别是在NS、PB、Facebook和GrQc等网络中，EGM实现了对引力模型的较大改善。

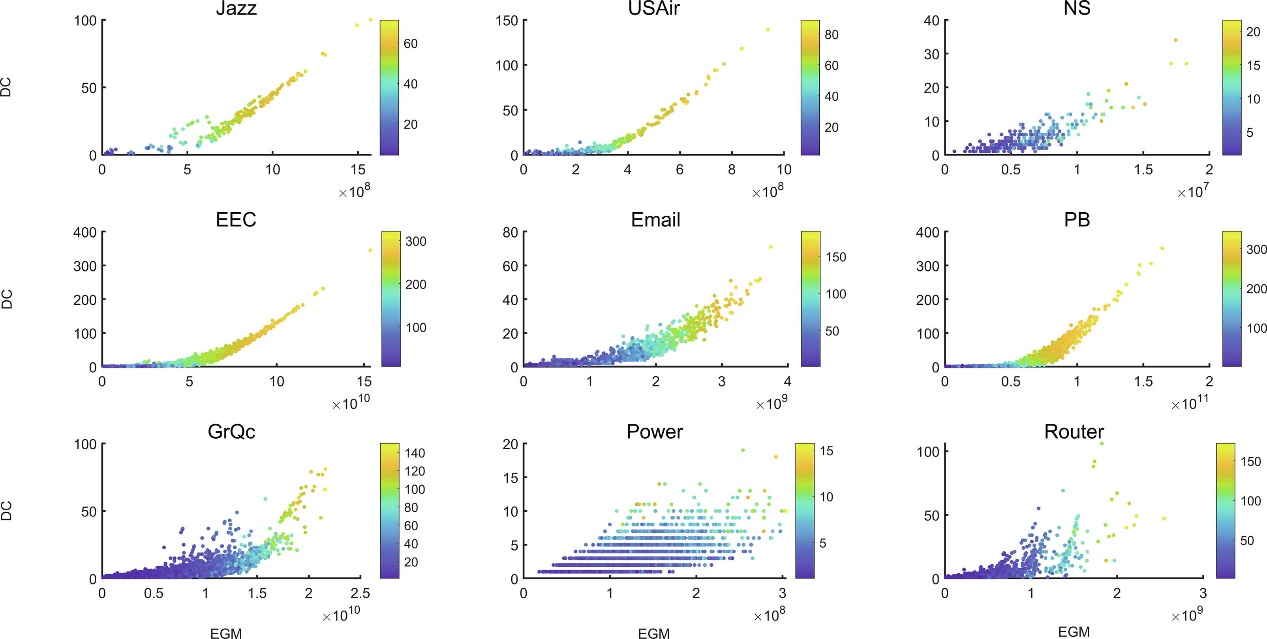


1. **相关性实验**

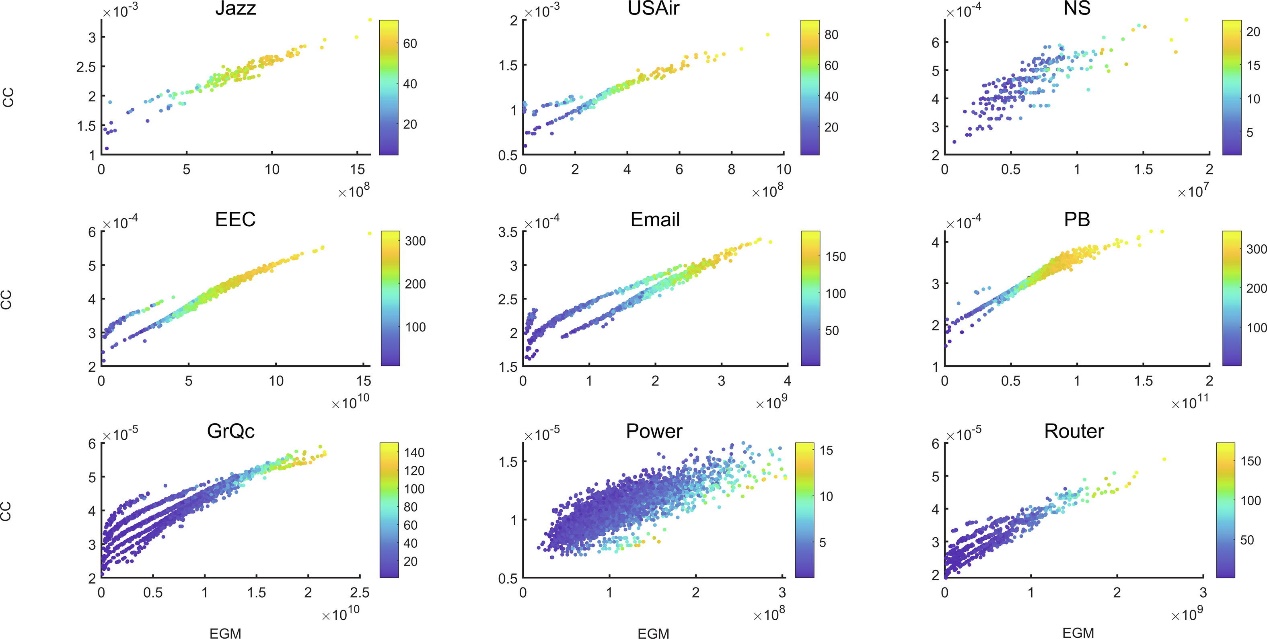
本实验的目的是通过评估EGM和其他方法之间的关系来证明所提出方法的合理性。此外，引入SI模型来评估节点之间的意义，其中传播概率β被设定为0.1，时间t被设定为10。每个实验都在九个不同的网络上进行测试，以全面衡量EGM和其他方法之间的相关性。所有的实验都进行了一百次，平均情况在图1、图8中表示。图中的每一点都表示现实世界网络中的节点。节点的感染能力可以通过F（10）当β=0得到，它由点的颜色表示。颜色越深，意义越大。如果EGM大的节点的另一个测量值也很大，这意味着EGM与该方法正相关。相反，如果EGM小的节点的另一个测量值很大，这意味着EGM与该方法是负相关的。

如图1、图8所示，网络中大多数节点的中心度都很小，这遵循了复杂网络的无标度特征。

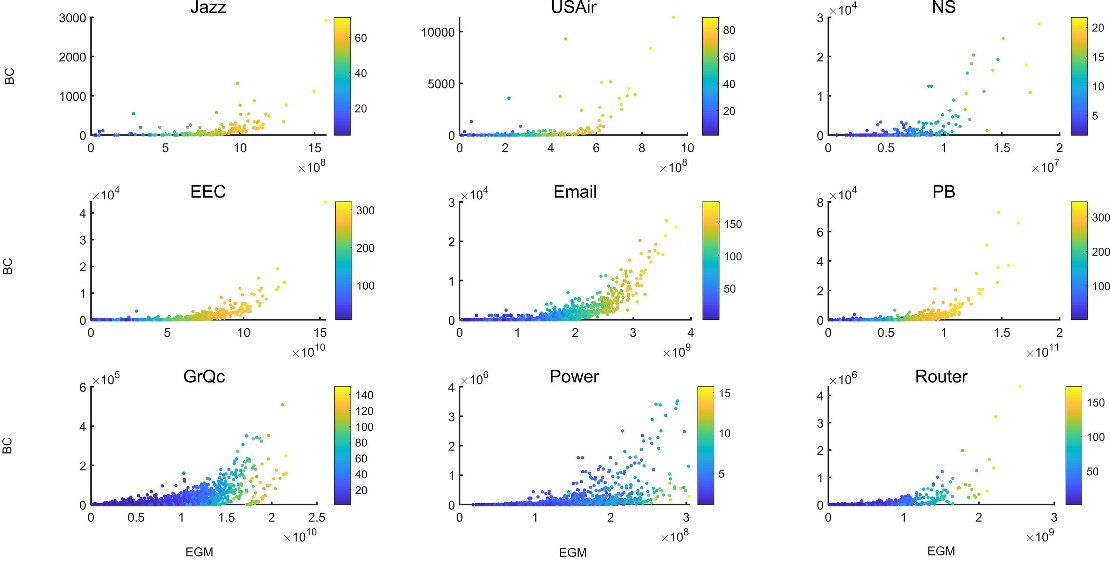
在图1中，DC和EGM是正相关的。特别是除GrQc、Power和Router外，它们之间的关联性更为明显。



如图2所示，CC和EGM之间的正相关关系稍强。



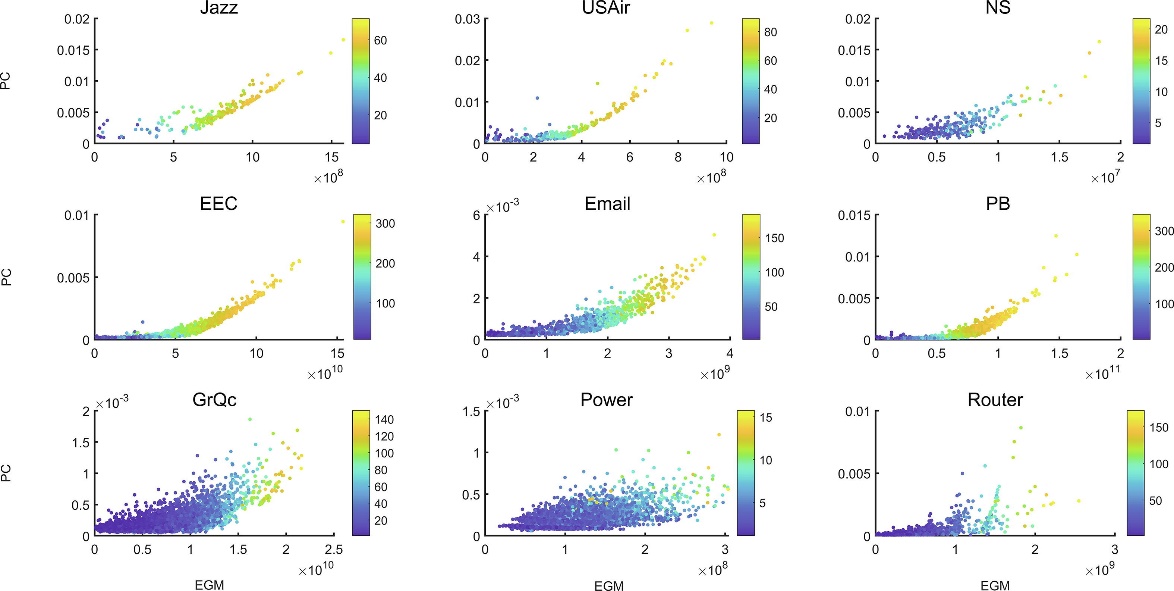
在图3中，BC和EGM之间的相互关系比较弱。



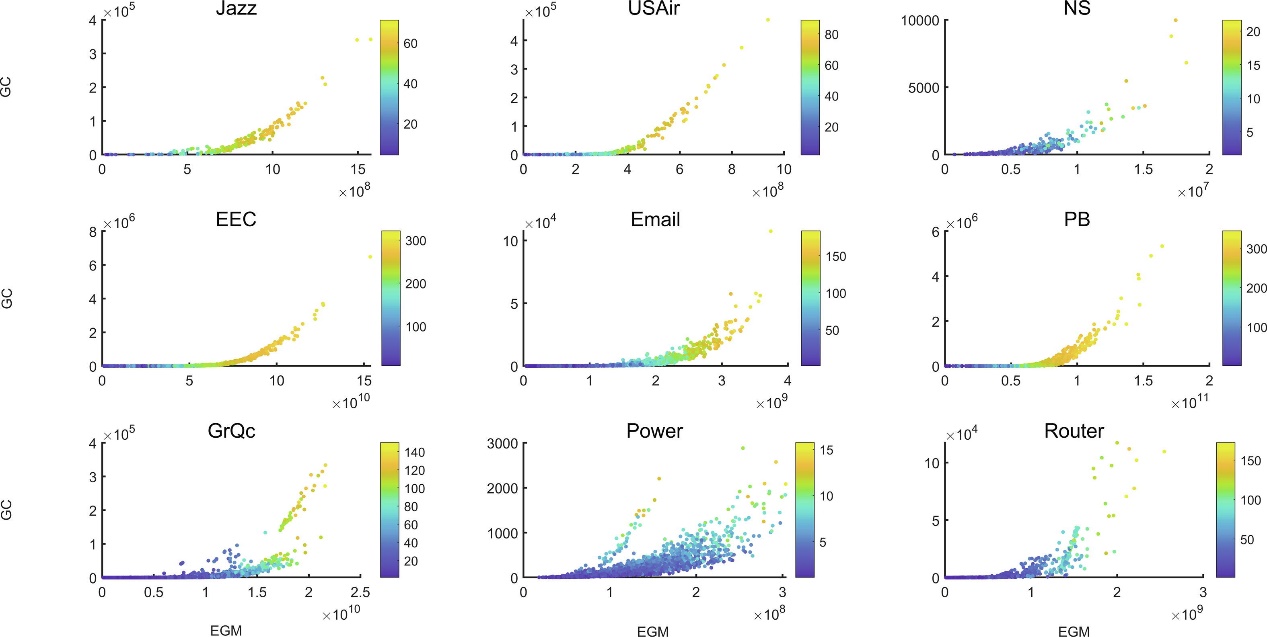
对于图4，在NS、GrQc和Router中，EC和EGM之间的关系较弱。



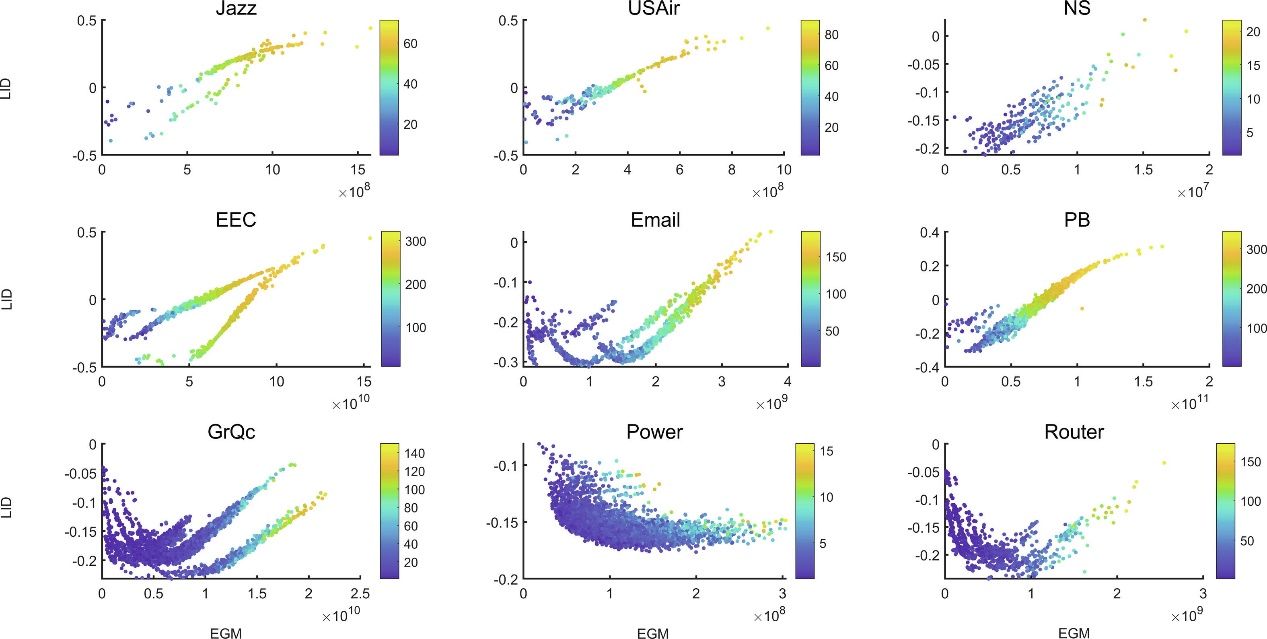
在图5中，PC和EGM之间的关系在GrQc、Power和Router中不明显。



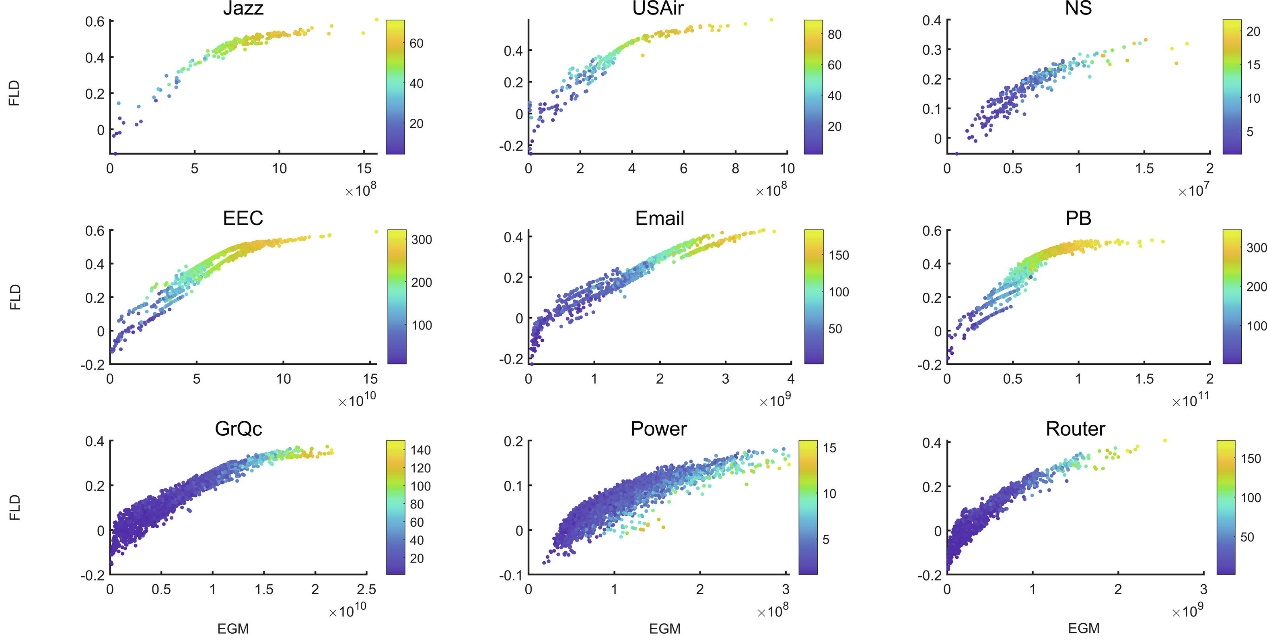
在图6中，EGM是对GC的改善，它与GC有很强的相互关系。



如图7所示，LID和EGM在Jazz、USAir、NS、EEC和PB中是正相关的。相反，LID和EGM在Power中是负相关的。特别是，LID和EGM之间的负相关在Email、GrQc和Router中变成了正相关。



对于图8，EGM与FLD有很强的正相关关系。可以得出结论，EGM与DC、CC、PC、GC、LID和FLD表现出很强的相关性，这反映了我们识别影响者的方法的合理性。



### 1.4 结论

本文提出的EGM以精确的半径和数值信息为基础，用于识别复杂网络中的关键传播者。本文的两个创新点是：精确的影响半径被确定，节点质量被客观地选择。在影响半径方面，EGM通过计算引力之间的相互作用范围，解决了基于重力的模型中模糊截断半径的问题。在质量选择方面，引入了信息熵来全面衡量邻居节点的影响，从而得出节点的价值信息。更重要的是，通过在11个真实世界的网络上进行6个不同的实验，比较了提出的EGM方法，包含DC、CC、BC、EC、PC的经典方法，包含GC、WGC、GGC的类似方法，以及包含LID、FLD的最先进措施。总之，EGM已经被证明在识别复杂网络中的关键传播者方面具有卓越的有效性和稳定性。

一些正在进行的工作如下。在应用方面，EGM目前只适用于无定向网络。对于有向加权网络，在EGM中计算影响半径时，需要参考连边的方向来衡量节点的影响范围。而在给节点分配价值信息时，可以引入权重来控制节点的价值。此外，就准确性而言，在测量节点的价值信息时，只有一阶邻居节点被EGM考虑。二阶邻域甚至聚类系数因子都可以用来代替节点的质量。最终，一个不可避免的问题是，当两个节点的EGM相同时，如何排名。在未来的研究中可以引入k-shell算法，该算法可以很好地考虑节点的位置。在未来，我们将继续改进EGM，以更好地识别复杂情况下的影响者。