Journal of System Simulation

Volume 32 | Issue 7 Article 6

7-15-2020

Modeling and Simulation On Influence of Complex Network Nodes Based on Data Field in

Chenxi Shao
School of Computer Science and Technology of University of Science and Technology of China, Hefei 230022, China;

Xiaoqi Chen
School of Computer Science and Technology of University of Science and Technology of China, Hefei 230022, China;

Xingfu Wang
School of Computer Science and Technology of University of Science and Technology of China, Hefei 230022, China;

Fuyou Miao
School of Computer Science and Technology of University of Science and Technology of China, Hefei 230022, China;

Modeling and Simulation On Influence of Complex Network Nodes Based on Data Field in

Abstract

Abstract: Research on the influence of complex network nodes is an important part of data mining. Mining the influential nodes in complex networks not only has important academic significance, but also helps to suppress the outbreak of epidemics, control the spread of rumors, and promote e-commercial products and so on. *By selecting the Mixed Degree Decomposition* (*MDD*) *value of each node as its mass, the complex network is abstracted into a data field, the influential nodes are identified by combining the data field model, and some well-known centralities are compares with.* The classical Susceptible-Infected-Recovered (SIR) epidemic model is used to evaluate the simulation performance by comparing the number of infected nodes. Simulations on real networks show that the data field can effectively identify the influential nodes.

Keywords

complex networks, influential nodes, mixed degree decomposition, data field, simulation

Recommended Citation

Shao Chenxi, Chen Xiaoqi, Wang Xingfu, Miao Fuyou. Modeling and Simulation On Influence of Complex Network Nodes Based on Data Field in[J]. Journal of System Simulation, 2020, 32(7): 1257-1266.

第 32 卷第 7 期 系统仿真学报© Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

基于数据场的复杂网络节点影响力建模与仿真

邵晨曦，陈小齐，王行甫，苗付友
(中国科学技术大学计算机科学与技术学院，安徽 合肥 230022)

摘要：复杂网络节点影响力的研究是数据挖掘的重要组成部分。挖掘出复杂网络中有影响力的节点不仅具有重要的学术意义，且有助于抑制流行病的爆发、控制谣言的传播和推广电子商务产品等。*通过选取每个节点的混合度分解值(Mixed Degree Decomposition，MDD)作为质量，将复杂网络抽象为数据场，结合数据场模型来识别有影响力的节点，并与一些著名的节点中心性方法进行对比。*使用经典的传染病模型(Susceptible-Infected-Recovered，SIR)通过对比感染节点的数量来评估仿真性能。对实际网络的仿真实验结果表明，数据场模型能够有效的识别网络中有影响力的节点。

关键词：复杂网络；节点影响力；混合度分解；数据场；仿真
中图分类号：TP391.9 文献标识码：A 文章编号：1004-731X (2020) 07-1257-10
DOI: 10.16182/j.issn1004731x.joss.18-0837

Modeling and Simulation On Influence of Complex Network Nodes Based on Data Field in

Shao Chenxi, *Chen Xiaoqi*, *Wang Xingfu*, *Miao Fuyou*
(School of Computer Science and Technology of University of Science and Technology of China, Hefei 230022, China)

Abstract: Research on the influence of complex network nodes is an important part of data mining. Mining the influential nodes in complex networks not only has important academic significance, but also helps to suppress the outbreak of epidemics, control the spread of rumors, and promote e-commercial products and so on. *By selecting the Mixed Degree Decomposition* (*MDD*) *value of each node as its mass, the complex network is abstracted into a data field, the influential nodes are identified by combining the data field model, and some well-known centralities are compares with.* The classical Susceptible-Infected-Recovered (SIR) epidemic model is used to evaluate the simulation performance by comparing the number of infected nodes. Simulations on real networks show that the data field can effectively identify the influential nodes.

Keywords: complex networks; influential nodes; mixed degree decomposition; data field; simulation

引言

随着科技的进步与发展，数据规模以指数级的速度增长，数据挖掘成为炙手可热的研究方向[1]。

挖掘关键节点已经成为数据挖掘的重要前提，具有重要的理论意义和实际应用价值，主要表现在抑制流行病的蔓延、控制谣言的传播和市场营销等方面[2]。通常，网络中具有更高可能性被感染或感染大多数邻居的节点被认为是影响力较高的节点。

分析复杂网络拓扑结构已经成为挖掘重要节点必不可少的一部分。许多经典的拓扑度量指标可用来评估节点影响力，例如度中心性[3]、介数中心性[4]和接近度中心性[5]等。其中度中心性简单直接，但是没有考虑到网络的全局结构；介数中心性和接

收稿日期：2018-12-17 修回日期：2019-04-30；
基金项目：国家自然科学基金(61472381, 61472382)；
作者简介：邵晨曦(1954-)，男，浙江杭州，硕士，副教授，研究方向为系统仿真；陈小齐(1994-)，男，安徽滁州，硕士生，研究方向为复杂网络。

近度中心性考虑了网络的全局结构，但由于计算复杂度颇高，因此不适用于规模庞大的网络。近年来，Kitsak 等[6]提出一种基于相同 shell 中的节点具有相似影响且较高 shell 中的节点可能感染更多节点的假设来识别最具影响力的节点的 *k*-shell 分解方法。然而，*k*-shell 分解方法倾向于将位于相同 shell 的但是彼此之间影响力不同的节点分配相同的 *k*-shell 值，因此通过 *k*-shell 分解方法得到的节点影响力单调性很差。为了解决单调性的问题，进一步改善 *k*-shell 方法的性能，Zeng 等[7]提出一种通过结合剩余度和删除度来预测节点传播能力的混合度分解方法(Mixed Degree Decomposition，MDD)，可以有效地降低 *k*-shell 算法的简并性；Chen 等[8]通过考虑次近邻来设计一个半局部指标；Lin 等[9]通过考虑目标节点与具有最高 *k*-shell 值的节点集合之间的最短路径，提出一种改进的节点影响力排序方法。

通常来说，复杂网络中的节点之间相互影响相互联系，这与物理场中物质粒子之间的相互作用类似，因此可以将物理学的知识应用到复杂网络中[10]。我们将数域空间中的数据对象看作是具有质量的物质粒子，所以每个数据对象会形成一个虚拟场，附近的数据对象会受到它场力作用的影响，且由所有数据对象相互叠加产生一个数域空间上的虚拟场。由于形成的虚拟场实质上是稳定的有源场，因此可以采用标量势函数以及矢量场强函数来描述其分布规律。这里采用 MDD 指标，将每个节点的 k_m 值视为其质量，并且将任意两个节点之间的最短路径距离定义为它们之间的距离。受这些工作的启发，提出一种新的中心性指标来衡量节点的影响力，称之为数据场中心性指标(Data Field-based Centrality，DFC)。为了评估算法的性能，可以使用 SIR 传染病模型来检查不同中心性指标的节点传播影响力[11]。对实际网络的仿真结果表明，提出的 DFC 算法可以很好地挖掘出网络中的有影响力的节点，并且与度中心性[3]、介数中心性[4]、接近度中心性[5]和 MDD[7]方法相比较，提出的 DFC 算法性能要好。

1 复杂网络数据场建模

从相关物理方面的知识，我们可以得到稳定的有源场主要包括核力场、静电场和重力场。在物理场的作用下，可以用势函数对数据场模型进行描述，该函数描绘了数据的贡献如何从单个样本范围扩散到整个集群范围。

1.1 势函数

假设在数域空间 Ω 中由数据集 D 生成的数据场中存在任意 M 维对象 $x=[x_1,x_2,...,x_M]^T$，它的势能值不仅与其质量成正比，而且与数据点 x 到其场源 X_i 之间的距离成反比。对于稳定有源场来说，势函数是关于场模型中空间距离的单值递减函数。通过与物理场中势函数定义的类比，可以得到数据场中势函数的定义如下：

$$\varphi(x)=\sum_{i=1}^{n} m_i \times K\left(\frac{x-X_i}{\sigma}\right) \tag{1}$$

式中：$m_i(m_i \geqslant 0)$为对象 X_i 的质量；$\| x-X_i \|$为数据 x 到对象 X_i 之间的距离；σ 为影响因子，具有极其重要的控制作用；$K(x)$为单位势函数。

1.2 数据场模型

借鉴物理中的场论概念，我们将描述数域空间中物质粒子之间的相互作用关系的方法模型应用到复杂网络中。给定图结构 $G=(V, E)$，其中节点集 V 包含 n 个节点，且每个节点被视为质量大于零的物质粒子。构建数据场模型最重要的问题就是如何衡量物质粒子之间相互作用的形式，接下来将详细讨论。

这里采用 F 来刻画复杂网络中两个节点之间的相互作用关系，且 F 与它们的质量呈正相关，与它们之间的距离呈负相关。因此吸引力 F 可以采用如下形式：

$$F(x)=\sum_{i=1}^{n} \frac{k_1 . m_i}{d^{k_2}} \tag{2}$$

式中：m_i 为节点 i 的质量；d 为节点 i 与其他节点

之间的距离；k_1 和 k_2 为相关系数。F 值越大，则说明节点间的相似性程度越深。

如果将数域空间中的数据集 D 的每个数据对象 X_i 视为核力场中的核子，则对应的 F 的扩展形式如式(3)所示。

$$F(x)_{\text{nuclear}}=\sum_{i=1}^{n}km_i\cdot\exp\left(-\left(\frac{\|x-X_i\|}{\sigma}\right)^2\right) \quad (3)$$

式中：k 为相关系数；$\|x-X_i\|$ 为场点 x 到核子 X_i 的距离；σ 为影响因子。

如果将静电场应用于数域空间中，则整体作用力 F 计算公式为：

$$F(x)_{\text{electrostatic}}=\sum_{i=1}^{n}\frac{kq_i}{\|x-X_i\|^2} \quad (4)$$

式中：q_i 为点电荷 X_i 的电荷量。

同理，如果将重力场应用到数域空间中，则势函数 F 为：

$$F(x)_{\text{gravity}}=\sum_{i=1}^{n}\frac{G\cdot m_i}{\|x-X_i\|} \quad (5)$$

式中：$\|x-X_i\|$ 为场点 x 到质点 X_i 的距离。

分析公式(3)~(5)可以发现都可以保证 F 是在数域空间内光滑、连续的，且具有各向同性。但为了保证收敛性，即能够保证作用力 F 随着节点间距离的增加而快速衰减直至为零，因此只有公式(3)满足要求，因为 F 与 $e^{\left(\frac{x-X_i}{\sigma}\right)^2}$ 成反比，意味着 F 会随着距离的增加而急剧下降，最终趋近于零。基于以上分析，作用力 F 采用核力场形式来描述。

1.3 影响因子

分析公式(3)可知 σ 作为节点属性唯一的自变量，将其称为影响因子，能够控制数据对象间的交互距离，具有极其重要的控制作用[12]。这里介绍一种基于拓扑势熵的 σ 选取方式。

根据信息论知识，熵越大，不确定性就越大。如果势值分布不均匀，则不确定性最小，则采用势值完全可以反映出数域空间中的节点分布。数据的不确定性通常采用熵来衡量，因此采用熵理论来最优化 σ 的值[13]。对于数据集 D，任意对象 X_i 的势值为 $\varphi(i)$，则势熵 H 的计算公式如式(6)所示。

$$H=-\sum_{i=1}^{n}\frac{\varphi(i)}{Z}\cdot\log\left(\frac{\varphi(i)}{Z}\right) \quad (6)$$

式中：$Z=\sum\varphi(i)$ 为一个标准化因子。

以 Facebook 关系网络为例，势熵 H 和影响因子 σ 之间的关系如图 1 所示。H 与 σ 之间的关系呈现先快速减小然后缓慢的增加直到最后趋于稳定的态势，因此势熵存在一个最小值，此时数据场的分布最为稳定，最能体现出节点之间的相互作用，即此时对应的影响因子 σ 为最优值。由高斯函数的数学性质可得，大约有 99.73%的数据随机分布在 3σ 范围内[14]。所以可以认为在数据场中，任意的数据对象的影响半径为 $3\sigma/\sqrt{2}$，即与指定数据对象之间的距离超出影响半径范围的数据对象之间的相互影响力可以忽略不计。

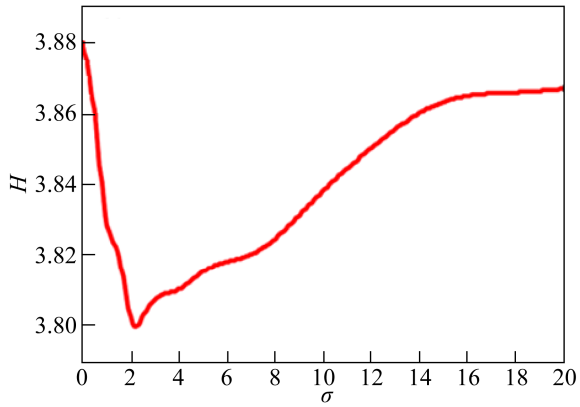

图 1 H 与 σ 的关系图

Fig. 1 Curve of potential entropy and impact factor

2 节点影响力度量指标

本文主要关注的是无权无向的简单网络。一个无向网络由具有 n 个节点和 m 条边的 $G=(n, m)$表示，并且网络结构可用相邻矩阵 $\boldsymbol{A}=(a_{ij})_{n\times n}$ 来描述，其中如果节点 i 与节点 j 相连，则 $a_{ij}=1$；否则，$a_{ij}=0$。这里，简单介绍一下将在本文中讨论的节点中心性指标。

度中心性(D)[3]可以表示为节点在网络中具有的邻居或边的数量。介数中心性(B)[4]是网络中用于识别核心节点的流行方法之一，被定义为通过指

定节点的最短路径。接近度中心性(C)[5]被定义为与其他每个节点之间测地距离长度之和的倒数。*k*-shell 分解方法(Ks)[6]通过以下步骤实现：首先，将度为 1 的所有节点删除，并继续删除现有节点，直到所有剩余节点的度数都大于 1，所有已删除的节点都标记为 1-shell；然后移除所有度数为 2 的节点，直到所有节点的度数都大于 2，将这些已删除节点标记为 2-shell；继续该过程直到所有的节点都被分配了一个指定的 shell。*k*-shell 方法仅仅考虑了剩余节点的度值，会将许多影响力不同的节点分配到相同的 shell，因此无法区分这些节点的影响力。为了解决这个问题，混合度分解方法(*MDD*)[7]同时考虑了剩余度和删除度，可以有效地降低 *k*-shell 算法的简并性，计算公式如式(7)所示。

$$k_m(i)=k_r(i)+\lambda\cdot k_e(i) \tag{7}$$

式中：$k_m(i)$为节点 i 的混合度；$k_r(i)$为剩余度；$k_e(i)$为删除度；λ 为介于 0~1 之间的可调参数。

在 *MDD* 的每个步骤中，根据 k_m 删除节点，并且剩余节点的 k_m 也随之更新。这里设定参数 λ=0.7 以充分考虑移除节点对节点中心性的影响。为了更好地说明 *MDD* 的过程，一个简单的例子如图 2 所示。

受到先前这些研究的启发，可以得出 2 个结论：一方面，如果节点的邻居具有更高的 *MDD* 值，则节点的影响力会提高；另一方面，任意 2 个节点之间的相互作用会随着它们之间的距离增加而快速衰减直至为零。可以看出这与复杂网络数据场模型的相关性质类似，因此将前面构建的复杂网络数据场模型应用到度量节点重要性的研究工作中。接下来使用 *MDD* 方法的结果来表示节点的质量且利用最短路径长度表示节点间的距离大小。基于以上的工作，网络中任意节点 i 的影响力可以通过式(8)来衡量(标记为 *DFC*)。

$$DFC(i)=\sum\nolimits_{j\in S(i)} m_i\cdot\exp\left(-\left(\frac{d_{ij}}{\sigma_{opt}}\right)^2\right) \tag{8}$$

式中：$m_i(m_i>0)$为通过 *MDD* 方法求出的节点的质量；$S(i)$为与节点 i 距离小于等于 $3\sigma_{opt}/\sqrt{2}$ 的邻域集合，其中 σ_{opt} 为影响因子的最优值；d_{ij} 表示节点 i 与 j 之间的最短路径长度。

由上述结论可知，如果某节点的邻居具有更高的影响力，则该节点的影响力会显著提高。因此提出一个综合考虑节点 i 的邻居影响力的扩展中心性指标如式(9)所示。

$$DFC_{\text{plus}}(i)=\sum\nolimits_{v\in\Gamma(i)} DFC(v) \tag{9}$$

式中：$\Gamma(i)$ 为节点 i 的最近邻节点集合。

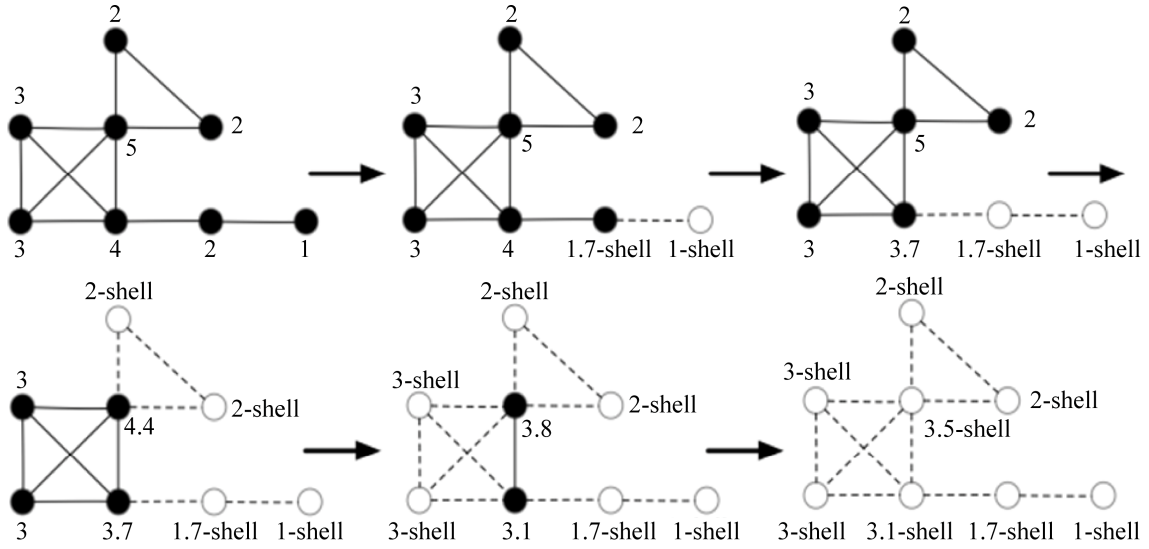

图 2 混合度分解(MDD)的过程
Fig. 2 Procedure of MDD

DOI: 10.16182/j.issn1004731x.joss.18-0837

3 实验与分析

本章从不同的方面及不同的网络(包括真实世界网络和仿真网络)上比较了我们的算法与其他算法的有效性和准确性。

3.1 数据集

为了验证 DFC_{plus} 算法的有效性和准确性，应用 8 个经典的真实世界网络来做仿真对比实验。包括：Zachary's Karate Club，从 1970~1972 年，Wayne W. Zachary 研究了空手道俱乐部社交网络[15]；Dolphins，海豚网络是居住在新西兰的 62 只海豚之间定期通信形成的无向社交网络[16]；Facebook，该网络是由美国著名社交平台脸书上用户之间的相互交通形成的社交网络[17]；NetSci，科学家合作网络是由科学家之间相互合作形成的网络[18]；Email，邮件网络是伦敦大学学院计算机科学系之间的电子邮件联系的拓扑结构[19]；Protein，蛋白质网络反映了酵母中包含的蛋白质之间的相互作用[7]；Power，美国西部各州的电网结构[20]；PGP，这是使用完美隐私(Pretty Good Privacy, PGP)算法的用户之间的交互网络[21]。这 8 个网络的详细拓扑结构属性如表 1 所示。

表 1 8 个真实网络的拓扑结构属性
Tab. 1 Topological properties of 8 real-world networks

网络	节点	边	平均度	密度	聚集系数
Karate	34	156	4.59	0.278	0.588
Dolphins	62	159	5.11	0.084	0.303
Facebook	324	2 218	13.69	0.042	0.466
Netsci	379	914	4.82	0.013	0.371
Email	1 133	5 451	9.622	0.009	0.220
Protein	1 870	2 277	2.44	0.001	0.055
Power	4 941	6 594	2.67	0.001	0.107
PGP	10 680	24 316	4.55	0.001	0.378

除了真实网络，还比较了算法在 LFR 仿真网络[22]上的性能。LFR 网络是以一个含有 8 个元素的参数向量$[n,<k>,k_{\max},C_{\min},C_{\max},\mu,\gamma,\beta]$生成的。其中，$\gamma$ 和 β 分别表示网络中度和社团满足的幂律分布的指数，n、$<k>$ 和 $k_{\max}$ 分别表示网络节点的个数、平均度和最大度，$C_{\min}$ 和 $C_{\max}$ 分别表示网络中社团规模的最小值和最大值，μ 是一个重要的混合参数，表示不属于基准节点所属的任何社团的节点的最近邻节点的平均分数值，这个参数控制不同社团之间边的分数值，能够反映出网络中噪声数据点的数量。当 μ=0 时，网络中所有的边都在相同社团内；当 μ=1 时，所有的边都位于不同社团之间。随着 μ 的增加，社团结构变得更加难以识别，也更难度量节点的影响力值。在实验中，为了保证结果的准确性以及减少噪声节点的影响，实验中为每个 μ 值生成 100 个 LFR 网络，然后运行算法并对每个算法的结果取平均值。

3.2 仿真评估指标

为了能够有效评估 DFC_{plus} 算法的性能，本文选取了排序单调性(Monotonicity，M)[10]、肯德尔相关性系数(Kendall Correlation Coefficient，τ)[23]和 SIR (Susceptible-Infected-Recovered)传染病模型[24]3 个指标来衡量算法的性能。

为了定量地提高关于节点影响力的排序区分度，利用单调性来进一步测量不同方法的性能，这种单调性计算公式如式(10)所示。

$$M(\boldsymbol{V})=\left[1-\frac{\sum_{t\in\boldsymbol{V}} n_t(n_t-1)}{n(n-1)}\right]^2 \tag{10}$$

式中：n 为网络大小；$\boldsymbol{V}$ 为排序向量；t 为排序值；n_t 为排序值为 t 的节点数目。当 $M(\boldsymbol{V})$接近 1 时，意味着该排序列表具有更好的单调性并且可以很好地区分节点；反之，如果 $M(\boldsymbol{V})$接近零时，则网络中节点只有一个排序值，即所有节点具有相同的影响力。

肯德尔相关性系数[23]用于测量基于某种方法得到的节点排序向量与实际节点影响力之间的相关性。假设$(x_1, y_1), (x_2, y_2),..., (x_n, y_n)$分别表示来自两个排序列表 X 和 Y 的一组组合排列。如果两个元素等级彼此一致，则说任意一对组合(x_i, y_i)和(x_i, y_i)是一致的。即，如果 $x_i>x_j$ 且 $y_i>y_j$ 或者 $x_i<x_j$ 且 $y_i<y_j$，则它们被认为是一致的；如果 $x_i>x_j$ 且 $y_i<y_j$

或者 $x_i<x_j$ 且 $y_i>y_j$，则它们被认为是不一致的；如果 $x_i=x_j$ 或者 $y_i=y_j$，则认为它们既不是一致的也不是非一致的。因此肯德尔系数可以通过计算两个数据集 X 和 Y 之间一致的和不一致的组合对数来定义，计算公式如式(11)所示。一般来说，τ 位于[–1,1]之间，$\tau>0$ 表示正相关，$\tau<0$ 表示负相关。因此，τ 是用于量化两个排序表之间排序相似性的对应属性的指标。

$$\tau(X,Y)=\frac{n_c-n_d}{\frac{1}{2}\cdot n(n-1)} \tag{11}$$

式中：n_c 为一致的组合对数的数目；n_d 为不一致的组合对数的数目。

经典的 SIR 传染病模型涉及 3 个不同的状态：Susceptible(S)、Infected(I)和 Recovered(R)。在这个模型中，首先设置一个节点被感染从而研究该节点的影响力，其他的节点被设置为 S 状态；然后在每一步中，每个 I 状态节点在试图以感染概率 λ 感染其 S 状态的最近邻节点后被恢复，重复这个过程直至网络中没有被感染的节点；接下来，记录 R 状态的节点数作为初始节点的影响力值。对于感染概率 λ 来说，可以利用平均场理论推导出关于种群网络中感染率 λ 的阈值，计算公式如式(12)所示。

$$\lambda_{th}=\frac{<k>}{<k^2>-<k>} \tag{12}$$

式中：$<k>$ 为平均度；$<k^2>$ 为二阶平均度。为了使感染过程更加充分，λ 通常被设置为稍微大于 λ_{th}。

3.3 结果与分析

通过实验数据集和性能评价指标的详细介绍，现在评估在一个示例网络和 8 个规模大小不同的真实世界网络中采用不同节点影响力排序方法的性能。如图 3 所示给出一个具有 18 个节点和 26 条边的示例网络以直观地比较这些性能指标，来自不同性能指标的排序值如表 2 所示。

从表 2 可以看出，k-shell 中心性不能很好地区分节点的影响力，即在同一个 shell 中，存在着许多影响力不同的节点。此外，结果表明，DFC_{plus} 可以很好地区分节点的影响力，并且与通过 SIR 模型获得的排序列表很好地保持一致。

在表 3 中，描绘了所有的节点中心性指标在 8 个真实世界网络中的排序单调性。

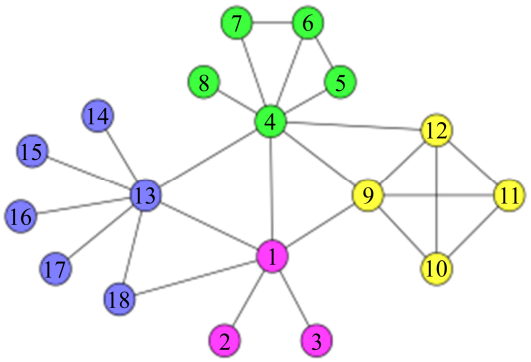

图 3 示例网络的结构图
Fig. 3 Structure of example network

表 2 不同指标决定的排序列表
Tab. 2 Ordering lists determined by different indices

排名	D	B	C	Ks	*MDD*	DFC_{plus}
1	4	13	4	9,10,11,12	4	4
2	13	1	1	1,4,5,6,7,13,18	13	1
3	1	4	13	其他	1	13
4	9	9	9	-	9	9
5	12	12	12	-	10,11,12	12
6	6,10,11	6	18	-	6	10,11
7	5,7,18	-	6	-	5,7,18	18
8	-	-	5	-	其他	6

表 3 不同中心性指标的单调性
Tab. 3 Monotonicity of different centrality indices

网络	D	B	C	Ks	k_m	DFC_{plus}
Karate	0.52	0.68	0.85	0.43	0.88	0.95
Dolphins	0.55	0.70	0.88	0.29	0.90	0.96
Facebook	0.43	0.49	0.78	0.14	0.81	0.93
Netsci	0.38	0.32	0.80	0.23	0.92	0.99
Email	0.31	0.33	0.77	0.16	0.78	0.90
Protein	0.56	0.60	0.75	0.49	0.72	0.99
Power	0.37	0.45	0.67	0.19	0.71	0.91
PGP	0.40	0.59	0.70	0.22	0.63	0.92

由表 3 可以看出，首先，对于 k-shell 中心性指标来说，因为此方法经常将影响力不同的节点分配到相同的 shell，因此 k-shell 方法应用于不同的

DOI: 10.16182/j.issn1004731x.joss.18-0837

网络获得的排序单调性是最差的；其次，结果表明 DFC_{plus} 可以得出更高的 M 值，而且在某些网络中更是非常接近 1，因此数据场模型方法 DFC_{plus} 拥有比其他影响力指标更好区分节点影响力的性能。

接下来，利用肯德尔系数 τ 探索 DFC_{plus} 方法的性能。当使用 SIR 模型来评估节点的影响力时，通过公式(11)可以计算出不同网络的感染率阈值，感染率 λ 设置为稍微大于阈值 λ_{th}。这 8 个网络的感染率阈值分别为 0.129，0.147，0.047，0.125，0.054，0.061，0.258 和 0.053，因此将实际的感染率 λ 分别设置为 0.13，0.15，0.05，0.13，0.06，0.07，0.26 和 0.06。然后，利用肯德尔系数 τ 测量了不同排序方法与 SIR 模型得到的节点传播影响力之间的相关系数，详细结果如表 4 所示。

表 4 不同方法对应的 Kendall 相关系数
Tab. 4 Kendall correlation coefficient of different indices

网络	τ(D)	τ(B)	τ(C)	τ(Ks)	$\tau(k_m)$	$\tau(DFC_{plus})$
Karate	0.399	0.609	0.753	0.360	0.720	0.832
Dolphins	0.411	0.606	0.575	0.309	0.699	0.913
Facebook	0.358	0.673	0.631	0.757	0.673	0.864
Netsci	0.528	0.679	0.713	0.401	0.817	0.948
Email	0.444	0.561	0.624	0.554	0.614	0.749
Protein	0.412	0.654	0.533	0.480	0.750	0.874
Power	0.463	0.538	0.372	0.374	0.812	0.798
PGP	0.548	0.649	0.622	0.459	0.813	0.936

通过观察表 4，可以总结出，在绝大多数情况下，DFC_{plus} 方法表现优于其他的方法，即通过 DFC_{plus} 方法能够识别出影响力更高的传播节点，因此可以说 DFC_{plus} 方法在识别节点影响力方面表现出更佳的性能。

同时，为了进一步评估感染率 λ 是如何影响不同性能指标的有效性，可以将相关系数值 τ 看作是变量感染率 λ 的函数，然后在 4 个流行真实世界网络上进行仿真实验，结果如图 4 所示。

观察图 4 可以发现，在绝大多数情况下，当 $\lambda>\lambda_{th}$ 时，DFC_{plus} 方法在识别网络中核心节点方面具有比其他方法更高的性能；其次，在 Facebook 网络上进行的仿真实验结果表明，随着感染率 λ 的不断增加，扩展的中心性指标 DFC_{plus} 的相关性系数 τ 会随之下降，因此可以推断出不同网络具有的不同的拓扑结构属性会影响节点的传播影响力。

排序单调性 M[10]、Kendall 相关性系数 τ[23]和 SIR 传染病模型[9]在评估节点影响力方面具有典型的代表作用。通过 3 种评价指标对比实验可以看出，DFC_{plus} 模型具有更好的性能，且会受到不同网络拓扑结构的影响，因此对于真实世界网络，DFC_{plus} 算法在度量节点影响力方面具有较好的表现。

接下来，还在具有 5 000 个节点的 LFR 基准测试网络上进行仿真实验，从而进一步评估感染率 λ 是如何影响不同性能指标的准确性，也是通过将相关系数值 τ 视为变量感染率 λ 的函数，详细的实验结果如图 5 所示。

从图 5 可以清晰地看出，首先，不管是在 μ=0.1 还是 μ=0.4 生成的 LFR 网络上得出的实验结果都表明，随着感染率 λ 的增加，所有算法的相关性系数 τ 呈现出先缓慢上升然后快速下降的现象，这时相关系数 τ 存在一个极大值，此时对应的极大值点绝大多数位于稍稍大于感染率阈值的位置(μ=0.1、μ=0.4 生成的 LFR 网络感染率阈值分别为 0.118 和 0.129)，这主要是因为感染率 λ 的取值略微大于阈值 λ_{th} 时，可以使得感染过程更加充分；其次，随着感染率 λ 的增加，在绝大多数情况下，提出的 DFC_{plus} 算法相对于其他算法都具有较好的性能，而 k-shell 算法在所有情况下的性能都是最差的，主要是因为 k-shell 算法只考虑剩余节点的度值，无法区分具有相同 shell 值节点的影响力；最后，当 μ 为 0.1 时得出的相关系数结果相对于 μ 为 0.4 时更好，这主要是因为混合参数 μ 能够控制不同社团之间边的分数值，且 μ 值越大，社团结构变得越难识别，节点的影响力值也就越难度量。

综合上述实验结果与详细分析，可以表明，我们的方法不仅可以识别真实世界网络上的有影响力的节点，还可以用于识别仿真网络。

第 32 卷第 7 期 系统仿真学报 Vol. 32 No. 7
2020 年 7 月 Journal of System Simulation Jul., 2020

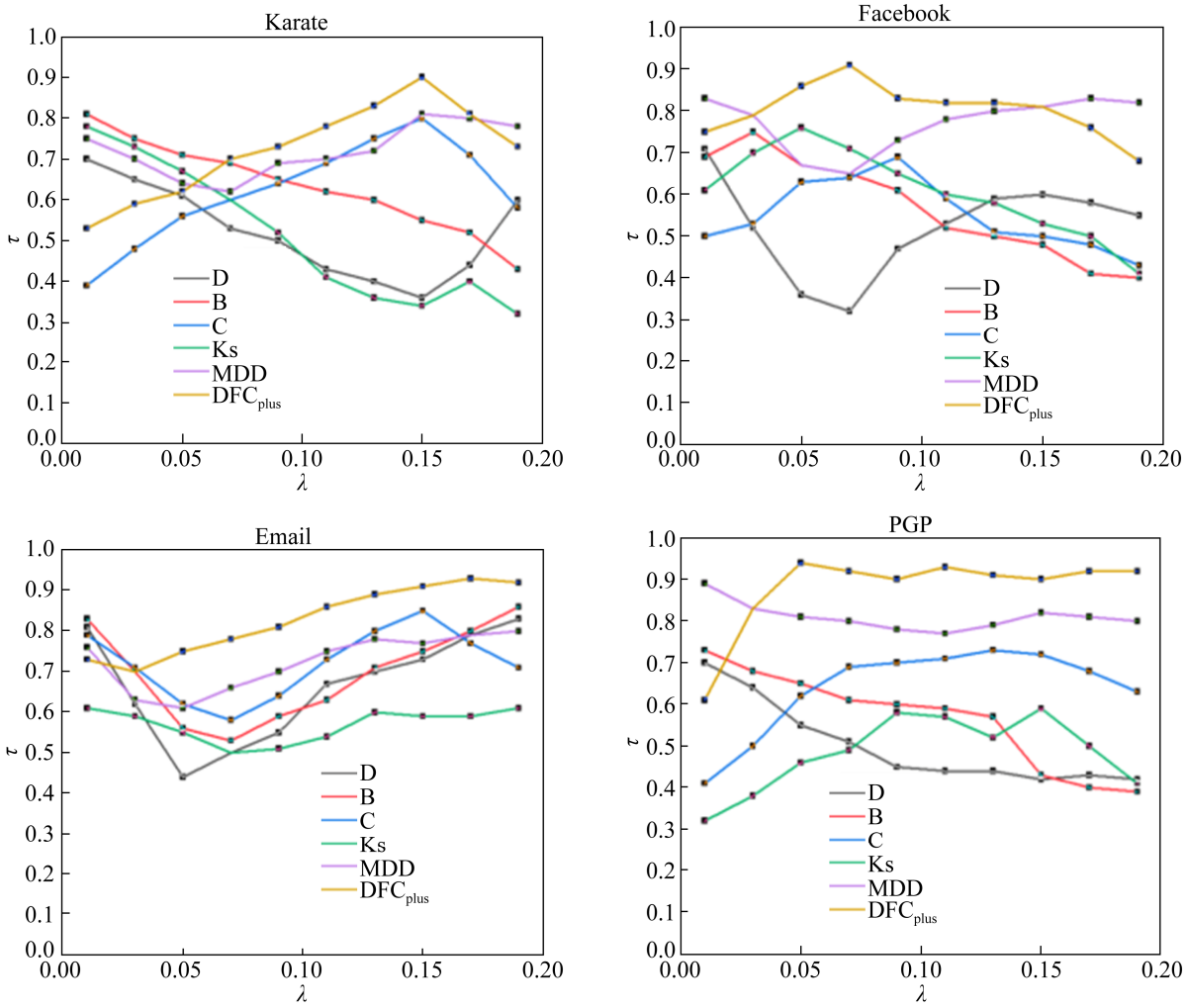

图 4 真实世界网络上 τ 与 λ 之间的变化曲线

Fig. 4 Curve between τ and λ on real-world networks

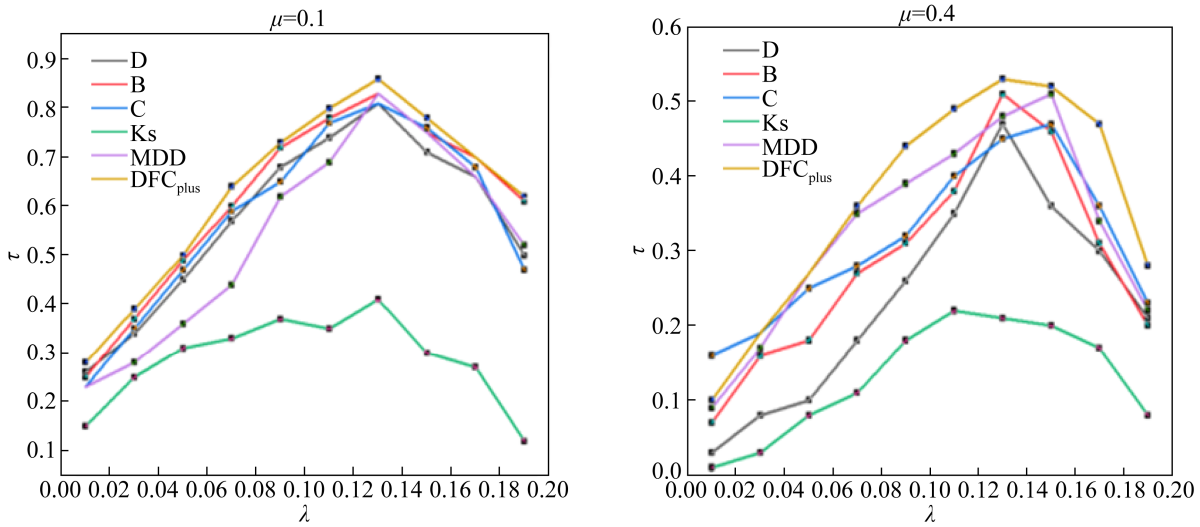

图 5 LFR 网络上 τ 与 λ 之间的变化曲线

Fig. 5 Curve between τ and λ on LFR networks

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/6
DOI: 10.16182/j.issn1004731x.joss.18-0837

4 结论

综上所述，本文提出了一种结合数据场模型的度量复杂网络中节点影响力的有效中心性排序方法：DFC_{plus} 方法。结合提出的数据集网络，经过实验方法设计、验证与结果分析。具体采用了混合度分解方法将每个节点的 k_m 值视为其质量，并且将两个节点之间的最短距离视为它们之间的距离；通过研究复杂网络的性质可将其抽象为数据场模型，并引用具有短程场性质的高斯函数来衡量节点的影响力，该数据场模型能够反映任意 2 个节点之间的相互影响力与其质量成正比；在此基础上，提出一些真实世界数据集的评价指标，将不同的方法应用到不同的网络中，通过计算单调性 M，发现 DFC_{plus} 方法能够很好地区分网络中节点影响力的差异；最后计算肯德尔相关性系数 τ。结果表明，在极大多数情况下，DFC_{plus} 算法在度量节点影响力方面比其他度量指标具有更好的性能；通过比较 SIR 传染病模型中各中心性指标对应的处于恢复状态的节点数来评估算法的有效性，可以发现，DFC_{plus} 方法有效性方面优于其他算法。

虽然 DFC_{plus} 算法可以取得较好的结果，但仍然存在着一些问题需要进一步研究和探索。首先，仅仅考虑了一种计算节点质量的方法，因此该方法可以进行一些扩展，比如采用多种中心性指标计算节点的质量进行比较；其次，仅专注于无权无向网络，在现实中，网络的结构则是多种多样的，网络之间不是独立存在的，而是彼此之间相互影响相互作用的，这对于高维数据集的处理结果有待改进，需要深入探索针对高维数据集的降维方法；最后，虽然可以获得满意的结果，但是需要计算每个节点的势值从而获得影响因子的最优值，所以计算时间复杂度稍许偏高。

因此，在接下来的研究工作中，将着手解决上述面临的问题。

参考文献:

[1] Wang S, Li Y, Wang D. Data field for mining big data[J]. Geo-spatial Information Science (S1009-5020), 2016, 19(2): 106-118.

[2] Ma L L, Ma C, Zhang H F, et al. Identifying influential spreaders in complex networks based on gravity formula[J]. Physica A Statistical Mechanics & Its Applications (S0378-4371), 2015, 451: 205-212.

[3] Tang Y, Li M, Wang J, et al. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks[J]. Biosystems (S0303-2647), 2015, 127: 67-72.

[4] Riondato M, Kornaropoulos E M. Fast approximation of betweenness centrality through sampling[J]. Data Mining and Knowledge Discovery (S1384-5810), 2016, 30(2): 438-475.

[5] Solé-Ribalta A, De Domenico M, Gómez S, et al. Random walk centrality in interconnected multilayer networks[J]. Physica D: Nonlinear Phenomena (S0167-2789), 2016, 323: 73-79.

[6] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics (S1745-2473), 2010, 6(11): 888-893.

[7] Zeng A, Zhang C J. Ranking spreaders by decomposing complex networks[J]. Physics Letters A (S0375-9601), 2013, 377(14): 1031-1035.

[8] Chen D, Lü L, Shang M S, et al. Identifying influential nodes in complex networks[J]. Physica a: Statistical mechanics and its applications (S0378-4371), 2012, 391(4): 1777-1787.

[9] Lin J H, Guo Q, Dong W Z, et al. Identifying the node spreading influence with largest k-core values[J]. Physics Letters A (S0375-9601), 2014, 378(45): 3279-3284.

[10] Zhuo Z, Zhang X, Niu W, et al. Improving data field hierarchical clustering using Barnes-Hut algorithm[J]. Pattern Recognition Letters (S0167-8655), 2016, 80(sep.1): 113-120.

[11] Li T, Wang Y, Guan Z H. Spreading dynamics of a SIQRS epidemic model on scale-free networks[J]. Communications in Nonlinear Science and Numerical Simulation (S1007-5704), 2014, 19(3): 686-692.

[12] Sun Z, Miao F, Xu Y, et al. Community detection in complex networks using flow simulation[C]// Proceedings of the 2017 VI International Conference on Network, Communication and Computing. Kunming, PEOPLES R CHINA: ACM, 2017: 150-155.

[13] Bae J, Kim S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness[J]. Physica A Statistical Mechanics & Its

Applications (S0378-4371), 2014, 395(4): 549-559.

[14] Barany I, Vu V. Central limit theorems for Gaussian polytopes[J]. The Annals of Probability (S0091-1798), 2007, 35(4): 1593-1621.

[15] Wang S, Chen Y. HASTA: A Hierarchical-Grid Clustering Algorithm with Data Field[J]. International Journal of Data Warehousing & Mining (S1548-3924), 2014, 10(2): 39-54.

[16] Liu Y, Jin J, Zhang Y, et al. A new clustering algorithm based on data field in complex networks[J]. The Journal of Supercomputing (S1573-0484), 2014, 67(3): 723-737.

[17] Wang S, Wang S, Yuan H, et al. Clustering by differencing potential of data field[J]. Computing (S1436-5057), 2018, 100(4): 403-419.

[18] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research (S0091-7710), 1977, 33(4): 452-473.

[19] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology (S0340-5443), 2003, 54(4): 396-405.

[20] Huang Z, Wang Z, Zhang Z. Detecting Overlapping and Hierarchical Communities in Complex Network Based on Maximal Cliques[C]// Chinese National Conference on Social Media Processing. Guangzhou, PEOPLES R CHINA: Springer Singapore, 2015.

[21] Karimi-Majd A M, Fathian M, Amiri B. A hybrid artificial immune network for detecting communities in complex networks[J]. Computing (S1436-5057), 2015, 97(5): 483-507.

[22] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical review E (S2470-0053), 2008, 78(4): 046110.

[23] Li C, Wang L, Sun S, et al. Identification of influential spreaders based on classified neighbors in real-world complex networks[J]. Applied Mathematics and Computation (S0096-3003), 2018, 320: 512-523.

[24] Fu Y H, Huang C Y, Sun C T. Using global diversity and local features to identify influential social network spreaders[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining. Beijing, PEOPLES R CHINA: IEEE, 2014.

https://dc-china-simulation.researchcommons.org/journal/vol32/iss7/6
DOI: 10.16182/j.issn1004731x.joss.18-0837