计算机科学
COMPUTER SCIENCE

一种基于影响力预测的节点排序模型

段顺然, 尹美娟, 刘粉林, 焦隆隆, 于岚岚

引用本文

段顺然, 尹美娟, 刘粉林, 焦隆隆, 于岚岚.一种基于影响力预测的节点排序模型[J]. 计算机科学, 2023, 50(3): 155-163.

DUAN Shunran, YIN Meijuan, LIU Fenlin, JIAO Longlong, YU Lanlan. Nodes' Ranking Model Based on Influence Prediction [J]. Computer Science, 2023, 50(3): 155-163.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200261

一种基于影响力预测的节点排序模型

段顺然　尹美娟　刘粉林　焦隆隆　于岚岚
战略支援部队信息工程大学网络空间安全学院　郑州 450001
(874117366@qq.com)

摘　要　节点影响力排序一直是复杂网络研究的热点问题。Susceptible-Infected-Recovered(SIR)模型是一种较为理想的节点影响力排序方法,业内常将其用于评价其他的节点影响力排序方法,但该方法时间复杂度较高,难以实际应用。文中提出一个基于sir值学习的节点影响力排序模型,模型综合节点的局部和全局结构信息描述节点特征,利用机器学习方法构建sir值学习模型,以构建的同等规模网络的节点特征和sir值对模型进行训练,训练后的模型能够基于节点特征预测节点的sir值,进而实现节点影响力排序。文中基于该模型实现了一个具体的节点影响力排序方法,并在真实数据集上进行了实验,结果表明,基于该模型得到的影响力排序结果,其准确性和单调性相比度中心性、Kshell、Weighted Kshell degree neighborhood等基于结构特征的方法均有所提升。

关键词:复杂网络;节点影响力;SIR;影响力排序
中图法分类号　TP391

Nodes' Ranking Model Based on Influence Prediction

DUAN Shunran,YIN Meijuan,LIU Fenlin,JIAO Longlong and YU Lanlan
School of Cyberspace Security,Information Engineering University,Zhengzhou 450001,China

Abstract　The ranking of nodes' influence has always been a hot issue in the research area of complex networks. Susceptible-infected-recovered(SIR) model is an ideal nodes' influence ranking method,which is commonly used to evaluate other nodes' influence ranking methods. But it is difficult to be applied in practice due to its high time complexity. This paper proposes a nodes' influence ranking model based on sir value learning. Both the local structure and global structure information of nodes are used as features in the model. The sir value learning model is constructed by means of a deep learning model,which is trained on nodes' features and sir data set in synthetic graphs with the same size. The trained model can predict sir value based on nodes' features, and then rank nodes' influence based on predicted sir. In this paper,a specific nodes' influence ranking method is implemented based on the proposed model,and experiments are carried out on five real networks to verify the effectiveness of the method. The results show that the accuracy and monotonicity of nodes' influence ranking results are improved compared with degree centrality,Kshell and Weighted Kshell degree neighborhood.

Keywords　Complex networks,Nodes' influence,SIR,Influence ranking

1　引言

从电力网络到万维网,从生态系统到动物群体社会关系,人类社会及自然界中存在着大量的复杂系统,而复杂系统可借助复杂网络来描述[1]。研究复杂网络上的信息传播,发现复杂网络上的高影响力节点在理论研究和实际应用中都具有十分重要的价值和意义。例如,发现社交网络中的高影响力节点对引导正面舆情、控制谣言传播具有重要作用;发现疾病传播网络中的高影响力节点对传染病的防控具有重要意义;发现计算机网络中的高影响力节点对防止病毒的扩散具有重要作用。

复杂网络中的高影响力节点发现问题属于最优化问题,是NP难的[2],现有的高影响力节点发现方法的主要思想是基于节点影响力对节点排序,根据排序结果选取top-k个节点作为高影响力节点[3]。现有的高影响力节点发现方法主要有两类[3]:一类是基于节点在网络结构上的特征评估节点影响力,另一类是基于渗流模型评估节点影响力。基于网络结构信息发现高影响力节点的代表性方法有度中心性[4]、介数中心性[5]和Kshell值[6],这类方法计算复杂度低,可在较短时间内实现节点排序。渗流模型通过模拟真实的信息传播过程评估节点影响力,通常需要针对网络上的每一个节点模拟

到稿日期:2021-12-24　返修日期:2022-04-07
基金项目:国家自然科学基金(U1804263);中原科技创新领军人才计划(214200510019)
This work was supported by the National Natural Science Foundation of China(U1804263) and Zhongyuan Science and Technology Innovation Leading Talent Project(214200510019).
通信作者:尹美娟(raindot_ymj@163.com)

数百乃至数千次传播过程，虽然此类方法对节点影响力的评估准确客观，但在大规模网络上进行传播模拟却是一个非常耗时的过程，因此该类方法一般被用于对其他节点影响力排序方法进行评价[3]。传染病模型 SIR[7] 是其中最常用的模型，如 Hierarchical Kshell(HKS)[8]，Local-and-Global-Centrality(LGC)[9]，Weighted Kshell degree neighborhood(KSDW)[10] 等利用 SIR 模型的排序结果与所提方法的排序结果的序列相关性来衡量方法的准确性；Kitsak 等[6] 利用 SIR 模型确定影响力排名靠前与靠后的节点，进而分析节点影响力与节点度中心性和 Kshell 的相关性。

SIR 模型对信息传播的模拟贴近实际，得到了领域内研究者的广泛认可，但较高的时间复杂度限制了其在真实网络中的应用。针对 SIR 模型对传播过程的模拟时间复杂度高、难以实际应用的问题，本文提出了一个基于 sir 值学习的节点影响力排序模型(Predicted Sir-based Ranking Model，PSRM)。该模型利用深度学习模型学习节点的结构特征与节点的 sir 值之间的关系，从而使得模型能够依据节点在网络中的局部和全局结构特征，预测节点的 sir 值。首先在生成的同等规模、类型多样的网络上计算节点结构特征和节点 sir 值，构建训练集，利用深度学习方法构建 sir 值学习模型，训练好的模型可以根据目标网络的节点结构特征预测节点 sir 值，进而实现节点影响力排序。

2 相关工作

基于不同的角度，利用网络结构信息进行高影响力节点发现的方法有不同的分类标准，本文根据判断节点影响力时所利用网络结构信息的不同类型，将其分为基于局部特征、基于全局特征和基于混合特征的方法。

利用网络局部信息的方法有度中心性[4]、Evidential centrality[11]、Neighborhood centrality[12]、H-index[13] 等。度中心性[4] 认为，节点具有的邻居数量越多，节点的影响力就越大。该方法计算简单，但忽略了对节点影响力具有影响的其他信息，如节点所处的网络位置、邻域内邻居的结构信息等。Evidential centrality[11] 权衡了节点的度和强度对节点影响力的影响程度，基于 Dempster-Shafer evidence theory 来刻画节点的传播能力。Neighborhood centrality[12] 认为节点的影响力由节点本身以及节点的一阶邻居、二阶邻居的结构特征确定，其以不同的权重将这些特征累加来衡量节点的影响力。H-index[13] 可用来评价研究者的学术影响力，反映了引用该文章的其他文章被引数情况，本质上是利用节点的邻居节点特征来衡量节点的影响力。上述方法利用节点的局部信息来评估节点的影响力，计算复杂度低，可实际应用，但此类方法忽略了会影响节点信息传播能力的其他因素，如节点在网络全局结构中的位置等。

介数中心性[5]、Kshell[6]、接近中心性[14]、PageRank[15]、Coreness[16] 等方法是典型的利用网络全局结构信息进行节点影响力排序的方法。接近中心性[14] 用网络所有其他节点到某节点的最短路径和的倒数来衡量该节点的影响力，其值越大表示该节点到其他节点的平均最短距离越小，信息传播能力越强。介数中心性[5] 用所有经过某节点的最短路径在所有最短路径中所占的比例来刻画该节点的影响力，其所占比例越大，表示该节点在网络结构中所承担的桥梁作用就越明显，对信息传播的控制能力越强。PageRank[15] 利用随机游走的方法，通过输出概率分布来体现某人随机点击某个链接(节点)的概率，通过概率来刻画节点的重要程度。基于此，Chawla 等[17] 将新兴的量子计算与 PageRank 的游走思想相结合，提出了离散时间量子漫游算法，理论上可对整个互联网上的页面节点进行排序。Coreness[16] 将网络节点分为节点间连接紧密的核心和连接稀疏的周围两个部分，认为位于网络核心的节点能够促进信息在网络间的传播。Kshell[6] 在网络的分解过程中将网络节点划分到不同的 Ks 层级中，认为层级数越高的节点越靠近网络中心，对信息的传播能力越强。这些方法在评估单个节点的影响力借助了网络全局的结构信息，但忽略了节点局部结构对信息传播的作用，使得排序结果缺乏良好的单调性。

基于混合特征的方法指在衡量节点的影响力时同时利用网络节点的局部结构信息和全局结构信息。Bae 等[18] 利用某节点所有邻居节点的 Kshell 值的和来衡量该节点的影响力。Namtirtha 等[10] 提出的基于“Weighted Kshell degree neighborhood(KSDW)”的节点影响力排序方法，首先利用节点的度和 Kshell 值对网络中的边进行赋权，用某节点与其一阶邻居边的权重之和来衡量该节点的影响力。Luo 等[19] 综合考虑了节点的功能属性及结构属性，提出了一种基于功能贡献度的节点重要性度量方法，并在“一般性社会网络”及“功能性社会网络”上分别进行了实验，证明了该方法在适应性及准确性方面均具有一定优势。Zeng 等[20] 对 Kshell 分解过程进行了改进，在每一步去除节点的过程中，同时考虑剩余节点与已去除节点的连接情况，利用该节点和剩余节点的连接个数与 λ 倍的与已去除节点的连接个数之和来衡量节点的影响力，λ 是可调参数。Zareie 等[21] 受邻域规则的启发，通过精心设计的算法综合考虑了节点及其邻居节点的度中心性和 Kshell 值，提出了 ECRM 指标，提高了节点排序的准确性。Zhao 等[22] 针对 Kshell 方法进行了改进，将 sigmod 函数和 Kshell 迭代分解的过程相结合，并采用信息熵加权的方法对节点位置信息和邻域信息进行加权，提出了 PN 方法，并验证了所提方法的正确性和有效性。Maji 等[23] 通过设计一种启发式算法综合了节点的度、接近中心性以及改进后的节点核数来评价节点的影响力，进而发现高影响力节点。此类方法大多通过设计巧妙的算法，以不同权重综合多种特征，借助了网络局部与全局的特征，考虑信息较为全面，但如何针对不同网络选择最合适的权重是此类方法面临的挑战。

本文从 sir 值预测的角度来探索高影响力节点的发现问题，属于基于混合特征的方法，通过构建训练集，利用深度学习方法构建 sir 值学习模型并进行训练，使得训练好的模型能够根据网络的结构特征预测节点的 sir 值，实现节点影响力排序。

3 PSRM

为方便描述，记：$G=\langle V,E\rangle$ 表示网络，其中，$V=\{v_1,v_2,v_3,\cdots,v_n\}$ 表示网络 G 的节点集合；v_i 表示第 i 个节点；$n=$

$|V|$表示网络 G 中的节点总数；$E \subseteq V \times V$ 表示网络连边集合，如果 v_i 和 v_j 存在连边，则将该边记为 $e_{i,j}$，并称 v_i 和 v_j 互为邻居节点；$neighbors(v_i)$ 表示 v_i 的邻居节点集合；$k_i = |neighbors(v_i)|$ 表示 v_i 的度；sir_i 表示 v_i 的 sir 值，文中用 SIR（大写）表示 SIR 传染过程，用 sir（小写）表示具体的 sir 值。

3.1 模型框架

本文提出的基于 sir 值学习的节点影响力排序模型 PSRM 的框架主要包括构建模拟网络集、获取训练数据、训练 sir 值学习模型、目标网络节点影响力排序 4 个部分，具体架构如图 1 所示。

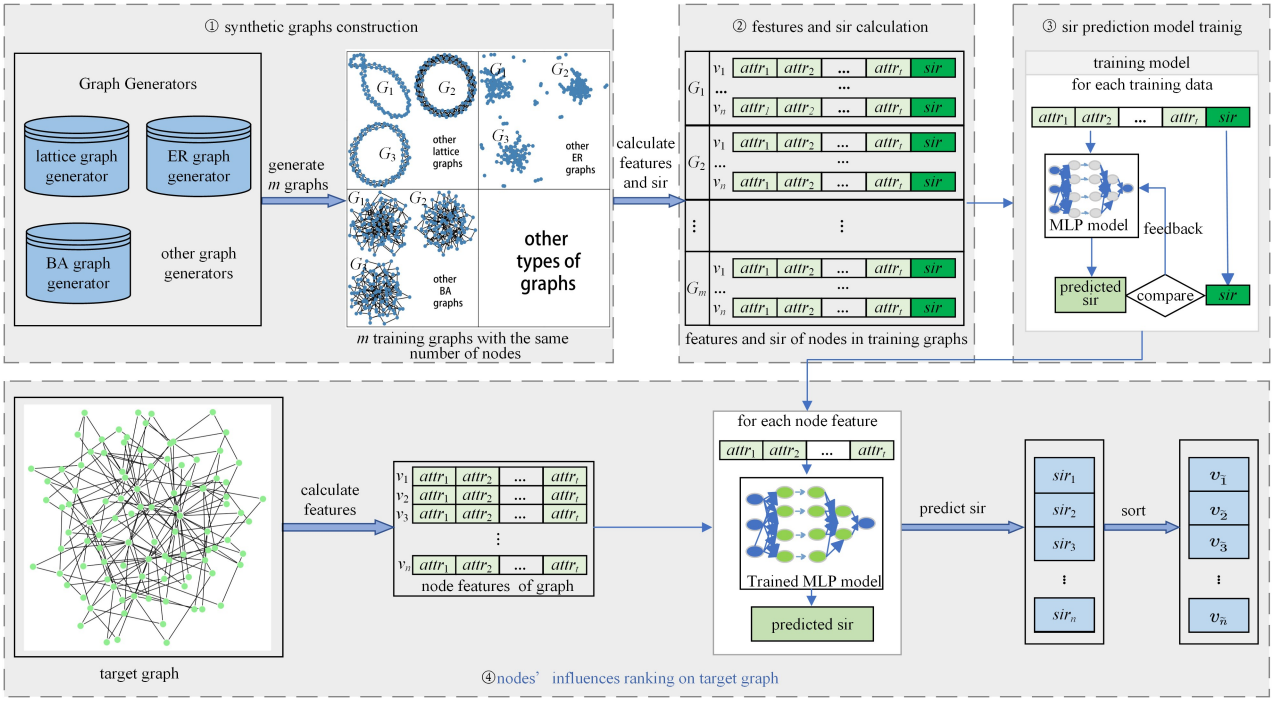


图 1 PSRM 流程图

Fig. 1 PSRM flow chart

PSRM 利用现有的网络生成模型生成模拟网络，计算模拟网络各个节点特征及 sir 值，以此为训练数据训练 sir 值预测模型，训练好的 sir 值预测模型能够根据节点的特征预测节点的 sir 值，并依据预测的 sir 值实现节点影响力排序。

PSRM 的主要步骤如下：

（1）构建模拟网络集

利用规则网络模型、随机网络模型、小世界网络模型、无标度网络模型等多种网络生成模型，构建类型多样、数量充足的同等规模的网络，网络个数记为 m。

（2）计算模拟网络集的节点特征值和 sir 值

1）计算节点的特征向量。选择节点 v_i 在网络局部结构和网络全局结构上的 t 个特征，生成节点 v_i 的特征向量，记作 $\vec{f}_i = \{attr_1, attr_2, \cdots, attr_t\}$，其中 $attr_i$ 表示节点的第 i 维特征。

2）计算节点的 sir 值。针对每个模拟网络进行 SIR 过程的模拟，得到各个节点的 sir 值，同一节点对应的特征向量及 sir 值构成一个训练样本，节点 v_i 对应的训练样本形式为 $(\vec{f}_i, sir_i) = (attr_1, attr_2, \cdots, attr_t, sir_i)$，同一个网络的所有节点对应的训练样本作为一组训练样本。

（3）训练 sir 值学习模型

选择合适的深度学习模型，如多层感知机，以 $(\vec{f}_i, sir_i) = (attr_1, attr_2, \cdots, attr_t, sir_i)$ 为输入，以最小化节点真实 sir 值与模型预测 sir 值之间的偏差为目标，学习 $\vec{f}_i$ 和 sir_i 之间的关系，使得在全体训练集上 sir 预测值与真实值的偏差最小，从而得到 sir 值学习模型。

（4）目标网络节点影响力排序

1）计算特征向量。计算目标网络中各个节点对应的特征向量 $\vec{f}_i = (attr_1, attr_2, \cdots, attr_t)$。

2）sir 值预测与节点排序。利用第三步中训练好的模型根据节点的特征向量 $\vec{f}_i$ 预测节点的 sir_i 值，较大的 sir 值代表节点影响力较大，因此可以依据预测的节点 sir 值的大小关系来实现对网络节点影响力的排序。

3.2 模型合理性分析

在第一步构建模拟网络集时，考虑到导致节点特征不同的原因主要有两个方面：一是节点间的连接关系不同；二是网络规模不同。由 SIR 模型中的传播模拟过程可以看出，传播仅可能通过节点之间的连边进行，因此网络节点间连接关系的不同是导致节点间 sir 值不同的根本原因，而节点间 sir 值之间的大小关系准确反映了节点间影响力的大小关系，因此可以推测网络节点间连接关系的不同是导致节点间影响力不同的根本原因。另一方面，考虑网络规模对不同节点特征的影响，以图 2 为例，网络 G_1 中的浅绿色节点 n 的度中心性为 $2/4=0.5$，Kshell 值为 2；而在网络 G_2 中，浅绿色节点 n 的度中心性为 $2/8=0.25$，Kshell 值为 2。由图 2 可以看出，网络规模的变化使得浅绿色节点的度中心性降低为原来的 1/2，

而 Kshell 值却保持不变。由此可知，当网络规模不同时，比较节点特征是没有意义的。因此，由网络规模引起的节点特征的变化是根据节点特征评估节点影响力的干扰因素。另一方面，从 sir 值本身的意义来讲，同一规模的网络只有结构上的不一样才有可能造成其节点 sir 值不一样，比较两个不同规模网络的 sir 值没有意义，因此在构建网络数据时需要构建同等规模的网络。同时网络数据应类型多样、数量充足，使得训练数据能够覆盖多种网络类型。网络数量充足是保证训练数据足够多的前提，在应用中，应根据实际情况选择合适的训练网络数量。

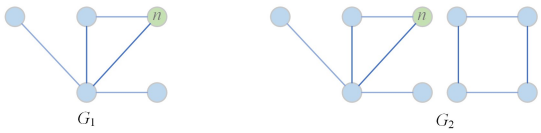


图 2 网络规模对节点特征的影响(电子版为彩图)

Fig. 2 Influence of network scale on nodes' features

在第二步获得训练数据时，需要考虑的节点特征应至少包含节点全局结构和局部结构两方面的特征(SIR 的计算利用网络的全局结构和局部结构)。节点的全局结构特征反映了节点在网络整体结构中的相对位置，对节点的影响力具有重要意义[6]；节点的局部结构特征反映了节点的局部信息传播能力，同时也有助于提高对节点影响力的区分粒度。只考虑网络全局结构特征会忽略节点局部结构给节点影响力带来的差别，使得排序单调性较差；只考虑局部结构特征会造成排序结果的准确性较低。例如，Namtirtha 等[10]综合考虑了节点的度和 KS 值，提高了排序的准确性和单调性；Namtirtha 等[24]基于节点的度、Kshell 和接触距离[25]提出了 Kshell 混合方法，提高了排序的准确性和单调性。

在第三步模型训练过程中，所选模型应能够很好地学习到节点特征向量和节点 sir 值之间的关系。Hornik 等[26]证明，包含足够多隐层神经元的多层前馈神经网络能够以任意精度逼近任意复杂度的连续函数，因此，为更好地学习节点特征与节点 sir 值之间的关系，本文选择了深度学习模型中的多层感知机(Multi Layer Perceptron，MLP，一种多层前馈神经网络)作为 sir 值的学习模型。

PSRM 模型针对目标网络构建同等规模的模拟网络集，避免了网络规模对节点特征的影响，在模拟网络集上，PSRM 模型通过训练，学习到节点特征与节点 sir 值之间的映射关系，其本质是特定规模下的网络节点间连接关系与节点影响力之间的映射关系。这种映射关系同样适用于同等规模的其他网络，因此可以根据同等规模的目标网络节点特征，基于 PSRM 模型学习到的映射关系预测节点影响力，进而实现节点影响力的排序。

4 实验验证和结果分析

本节介绍基于 PSRM 实现的节点影响力排序方法(简记为 psrm)有效性的实验设计。实验选择了 5 个真实网络数据作为实验验证数据，从算法的有效性和单调性两个方面进行对比实验。实验对比方法为基于度中心性的排序方法(简记为 dc)、基于 Kshell 值的排序方法(简记为 ks)和基于 KSDW 的排序方法(简记为 ksdw)。选择这 3 种方法作为对比方法有两方面的原因：一方面，本文基于 PSRM 实现的节点影响力排序方法利用的节点特征涉及了以上 3 种方法；另一方面，dc 是具有代表性的基于网络局部结构的节点影响力排序方法，ks 是具有代表性的基于网络全局结构的节点影响力排序方法，ksdw 是近年来提出的综合了网络局部和全局特征的节点影响力排序方法。

4.1 PSRM 参数设置

本文基于 PSRM 模型实现节点影响力排序算法的参数设置及具体步骤如下。

4.1.1 构建模拟网络集

在构建模拟网络集时，针对每一个具有 n 个节点的目标网络 G 生成 $10n$ 个节点数为 n 的网络，网络类型包含随机网络、小世界网络、无标度网络等，以此为基础构建训练数据集。在 SIR 模型中恢复率 α 设置为 1，传染率 β 设置为 $\theta(\theta=1.2, 1.6, 2.0)$ 倍的传染阈值，SIR 模型重复 1000 次。有关 SIR 模型的详细介绍参见 4.1.4 节。

4.1.2 获得训练数据

实验所选用的局部节点特征中节点的度中心性 $DC_i=\frac{k_i}{n-1}$、传递中心性 $Transitivity_i=\frac{2|\{e_{pq}:v_p,v_q\in neighbors(v_i),e_{pq}\in E\}|}{k_i(k_i-1)}$，全局特征选择 Kshell，并选取了混合特征 KSDW[10]。

网络节点的 Kshell 值在网络的 Kshell 分解的过程中被赋予，在 Kshell 分解中，依次移除所有度为 $k(k=1,2,3,\cdots)$ 的节点以及与之相连的连边，直到剩余网络中不存在度为 k 的节点，被移除节点的 $Kshell=k$。KSDW 利用节点的度和 Kshell 值对网络中的边进行赋权，而后用节点的一阶邻居边的权重之和衡量该节点的影响力。由于论文中的权重是针对实验所用的特定网络设计的，并没有给出统一的设置标准，因此本文在计算 KSDW 时将节点度和节点的 Kshell 值的权重均设置为 0.5。

选择这几个特征的原因在于，信息需要在局部区域充分传播，并且局部区域应具备很好的向外扩散能力，这样才能使得影响力较高。度中心性和传递中心性可以衡量节点在局部的信息扩散能力，而 Kshell 反映了节点在网络核心的程度，可以用来衡量信息向外部扩散的能力，而 KSDW 方法作为近年来衡量节点影响力效果较为出色的方法，可以对节点特征进行补充。

4.1.3 训练 sir 值学习模型

实验中选择的 sir 值学习模型为多层感知机，模型含有两个隐藏层，隐藏层个数分别是 6 和 2，节点激活函数采用 Relu 函数，优化算法采用 Adam 算法，训练数据样式为 $(\vec{f}_i, sir_i)=(KS_i, D_i, T_i, KSDW_i, sir_i)$，其中 KS 表示节点 Kshell 值，D 表示节点的度中心性，T 表示节点的传递中心性，$KSDW$ 表示节点的 KSDW 值，下标 i 表示节点序号，多层感知机结构如图 3 所示。

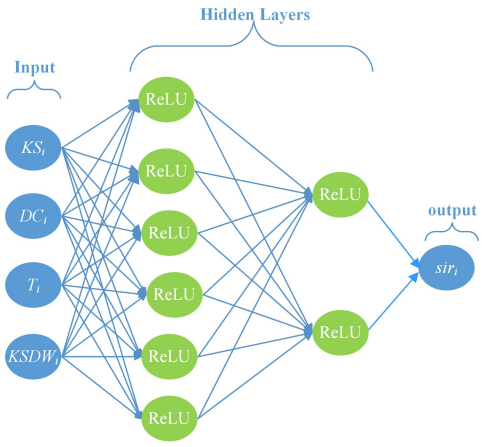


图 3 多层感知机结构

Fig. 3 Multilayer perceptron architecture

4.1.4 SIR 模型

SIR 模型是一种疾病传播模型，该模型把网络中的所有节点划分为 3 种状态：易感染状态(S)、感染状态(I)、恢复并免疫状态(R)，并满足 $|S \cup I \cup R| = n$，且 S, I, R 两两不相交。初始时，网络中只有一个节点处于 I 状态，其余所有节点均处于 S 状态。传染通过节点之间的连边进行，节点 sir 值的计算需要进行多轮 SIR 模拟，在一轮传染过程中，每一个时间步处于 I 状态的节点以概率 β 感染其直接邻居中所有处于 S 状态的节点，同时自身以概率 μ 进入 R 状态，处于 R 状态的节点不会再被感染，当处于 I 状态的节点为空时，即所有节点均处于 S 状态或 R 状态时，则该轮传染过程结束。r_i 表示初始时，仅有节点 v_i 处于 I 状态的 SIR 传染过程结束时，状态集合 R 中的节点数占网络节点总数的比例，即 $r_i = \frac{|R|}{n}$，若传染过程总计进行了 w 轮，则 $sir_i = \frac{1}{w}\sum_{i=1}^{w} r_i$，SIR 模型根据 sir_i 进行节点重要性排序。传染率 β 的设置一般建议略大于传染病阈值 β_{th}[18,27]，在 SIR 模型中，$\beta_{th} = \frac{1}{\langle k^2\rangle/\langle k\rangle - 1}$[28]，其中 $\langle k\rangle$ 和 $\langle k^2\rangle$ 分别表示节点的平均度和二阶平均度。

网络中信息的传播过程可类比于 SIR 模型中疾病的传播，节点 a 感染节点 b，意味着信息由 a 传递到 b，节点康复并免疫意味着信息对节点过时，节点不再传播信息也不再接受信息。在信息传播过程中，sir_i 代表了信息由节点 v_i 进行传播，最终网络中成功接收信息的节点占网络总节点数的比例。

4.2 实验数据

实验选择的真实网络来源[29]如下。

dolphins：dolphins 网络是一种动物社交网络，图中节点代表海豚，两个节点之间的连边表示相应的两只海豚频繁地一起活动。

football：football 网络是根据美国大学生足球联赛而创建的一个社会网络，图中的节点代表足球队，两个节点之间的连边表示两支球队之间进行过一场比赛。

karate：karate 网络描述了美国一所大学空手道俱乐部成员之间的朋友关系，图中节点代表俱乐部成员，两个节点之间的连边代表相应的两个成员之间是交往频繁的朋友关系。

eco-stmarks：eco-stmarks 网络是一个食物网络，图中节点代表物种，两个节点之间的连边代表物种之间具有捕食关系。

reptilia-tortoise-network-bsv：reptilia-tortoise-network-bsv 是一个动物社交网络，图中节点表示海龟，两个节点之间的连边代表与之相关的海龟曾共用一个巢穴。

4.3 评价指标

算法有效性的评价指标选择常用来度量两个序列相似程度的指标：肯德尔秩相关系数和 Rank Biased Overlap(RBO)相关系数。参考文献[8,18]，本文选择的算法单调性评价指标为辅助累计分布函数(Complementary Cumulative Distribution Function，CCDF)。

4.3.1 肯德尔秩相关系数[18]

假设两个序列 x 和 y，它们的元素个数均为 N，两个序列中的第 $i(1\leqslant i\leqslant N)$ 个元素分别用 x_i 和 y_i 表示。x 与 y 中的对应元素组成一个元素对集合，其包含的元素为 $(x_i, y_i)(1\leqslant i\leqslant N)$。当集合 x 和 y 中任意两个元素 (x_i, y_i) 与 (x_j, y_j) 的顺序相同时，即 $x_i > x_j$ 且 $y_i > y_j$ 或 $x_i < x_j$ 且 $y_i < y_j$，这两个元素是一致的。当出现 $x_i < x_j$ 且 $y_i > y_j$ 或 $x_i > x_j$ 且 $y_i < y_j$ 时，这两个元素是不一致的。当出现 $x_i = x_j$ 或 $y_i = y_j$ 时，这两个元素既不是一致的也不是不一致的。则肯德尔秩相关系数 τ 的计算如下：

$$\tau = \frac{C - D}{\sqrt{(N_0 - N_1)(N_0 - N_2)}}$$

其中，C 表示一致元素对的个数，D 表示不一致元素对的个数，N 表示元素的总个数。$N_0 = \frac{1}{2}N(N-1)$，$N_1 = \sum_{i=1}^{s_1}\frac{1}{2}U_i(U_i - 1)$，$N_2 = \sum_{i=1}^{s_2}\frac{1}{2}V_i(V_i - 1)$，其中，$s_1$ 和 s_2 表示 x 与 y 中重复出现的元素种类数，U_i 和 V_i 分别表示 x 与 y 中由相同元素组成的第 i 个集合的元素个数。

4.3.2 RBO 相关系数[30]

假设序列 x 和 y 为两个无限长度的序列，x_i 为序列 x 第 i 个元素，$x_{m:n} = \{x_i : m\leqslant i\leqslant n\}$ 表示序列从位置 m 到位置 n 的所有元素组成的集合。序列 x 和 y 在深度为 d 时的交集记为 $I_d = x_{1:d} \cap y_{1:d}$，交集的元素个数称为序列 x 和 y 在深度为 d 时的交叠，该交叠相对于深度 d 的比值称为序列 x 和 y 的一致度，即一致度 $A_d = \frac{|I_d|}{d} = \frac{|x_{1:d} \cap y_{1:d}|}{d}$。RBO 通过给每个深度的一致度赋予权重 w_d，再计算加权和，即：

$$RBO(x, y, w) = \sum_{d=1}^{\infty} w_d \cdot A_d$$

在 RBO 的应用中，权重 w_d 应满足 $\sum_d w_d = 1$，一般将权重 w_d 设置为 $w_d = (1-p)\cdot p^{d-1}$，其中 $p\in(0,1)$ 是一个可调参数，本文实验中设置 $p = 0.5$。RBO 中的权重 w_d 一方面实现了归一化，另一方面也使得权重的分配可以调节。

4.3.3 CCDF

单调性反映了对节点等级的划分粒度，最好的情况是所有节点都被划分到不同的等级。本文使用辅助累计分布函数 CCDF[8,18]来分析不同排序方式的单调性。等级 r 的 CCDF 函数值由下式给出：

$$CCDF(r)=\frac{n-\sum_{i=1}^{r}n_i}{n}$$

其中，n_i 表示属于 i 等级的节点数量。通过绘制 CCDF 函数值随等级 r 变化的曲线图可以分析排序方式的单调性，如果函数值趋向于 0 的速度越快，即曲线越接近坐标轴，代表有越多的节点被划分到了同一等级，则划分的单调性越差；反之，曲线斜率越接近 -1，则单调性越好。

4.4 实验结果与分析

本文从有效性和单调性两方面分析实验结果，首先在不同 θ 下计算各个网络上基于 SIR 模型的影响力排序结果，作为基准排序，θ 表示 SIR 模型中传染率相对于传染阈值的倍数；而后计算由 4 种排序方法(psrm，ks，dc，ksdw)获得的各个网络节点影响力排序结果与基准排序的相关性(肯德尔秩相关系数和 RBO 相关系数)及单调性，并分析了肯德尔秩相关系数和 RBO 相关系数在评估序列相关性方面的不同。

4.4.1 有效性实验

(1)肯德尔秩相关系数

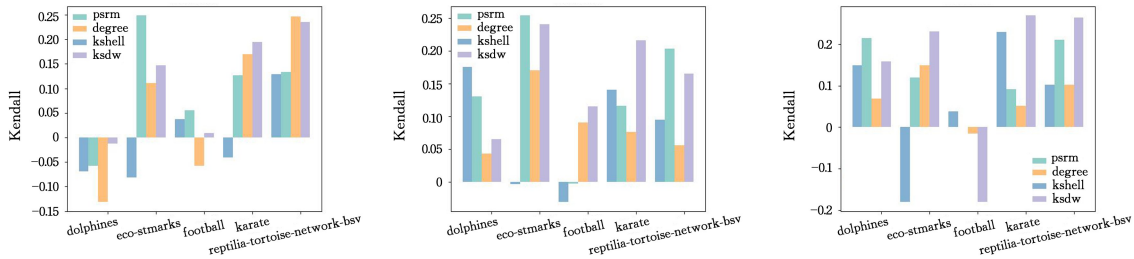
表 1 给出了不同 θ 下，4 种节点影响力排序方法的排序结果与基准排序的肯德尔秩相关系数。

表 1 肯德尔秩相关系数

Table 1 Kendall rank correlation coefficients

graph	θ	psrm	dc	ks	ksdw
dolphins	1.2	−0.0576	−0.1317	−0.0693	**−0.0122**
eco-stmarks	1.2	**0.2485**	0.1111	−0.0815	0.1475
football	1.2	**0.0560**	−0.0578	0.0374	0.0093
karate	1.2	0.1266	0.1693	−0.0410	**0.1943**
reptilia-tortoise-network-bsv	1.2	0.1337	**0.2465**	0.1288	0.2353
dolphins	1.6	0.1306	0.0428	**0.1750**	0.0650
eco-stmarks	1.6	**0.2539**	0.1704	−0.0034	0.2404
football	1.6	−0.0026	0.0902	−0.0310	**0.1155**
karate	1.6	0.1159	0.0766	0.1408	**0.2157**
reptilia-tortoise-network-bsv	1.6	**0.2035**	0.0556	0.0945	0.1655
dolphins	2.0	**0.2142**	0.0682	0.1486	0.1581
eco-stmarks	2.0	0.1192	0.1488	−0.1811	**0.2310**
football	2.0	0.0008	−0.0154	**0.0380**	−0.1808
karate	2.0	0.0909	0.0517	0.2299	**0.2692**
reptilia-tortoise-network-bsv	2.0	0.2102	0.1026	0.1018	**0.2641**

图 4 分别给出了在不同 θ 下，4 种节点影响力排序方法的排序结果与基准排序的肯德尔秩相关系数的柱状图。图中横坐标表示不同的网络，纵坐标表示肯德尔秩相关系数。



(a)kendall 1.2 (b)kendall 1.6 (c)kendall 2.0

图 4 肯德尔秩相关系数柱状图

Fig. 4 Histogram of Kendall rank correlation coefficients

由表 1 可以看出，parm 和 ksdw 在肯德尔秩相关系数方面准确性比 dc 和 ks 更高。肯德尔秩相关系数在衡量序列相关性时没有考虑元素序列位置对相关性的影响，而在高影响力节点发现研究中更关注影响力靠前的节点，这使得当部分影响力靠后的节点次序错误时，对肯德尔秩相关系数的影响较大，使得肯德尔秩相关系数数值较小。

(2)RBO 相关系数

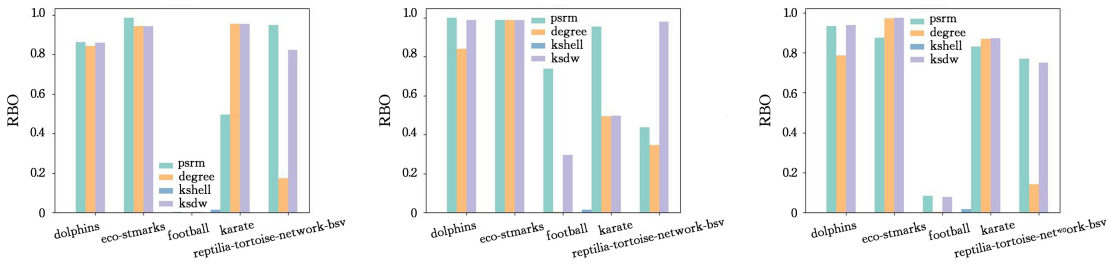
表 2 列出了不同 θ 下，4 种节点影响力排序方法的排序结果与基准排序的 RBO 相关系数。

图 5 分别给出了在不同 θ 下，4 种节点影响力排序方法的排序结果与基准排序的 RBO 相关系数的柱状图。图中横坐标对应不同的网络，纵坐标表示 RBO 相关系数。结合图 5 和表 2 可以看出，在不同传染率 β 下，基于 sir 值预测的方法 psrm，dc，ksdw 表现较好，其中 psrm 取得了 10 个最优效果，ksdw 取得了 5 个最优效果，ks 表现欠佳。对比肯德尔秩相关系数可以发现，在 RBO 相关系数的衡量下，各种方法的准确率均有提高。RBO 相关系数在衡量序列相关性时，考虑了元素序列位置对相关性的影响，在 RBO 相关系数的计算中，序列中位置靠前的两个元素排序错误对 RBO 的负面影响大于位置靠后的两个元素排序错误对 RBO 的负面影响。高影响力节点发现方法中更看重影响力靠前节点的排序准确性，因此 RBO 相关系数的数值要大于肯德尔秩相关系数的数值。

表 2 RBO 相关系数

Table 2 RBO correlation coefficients

graph	θ	psrm	dc	ks	ksdw
dolphins	1.2	**0.8624**	0.8413	0	0.8595
eco-stmarks	1.2	**0.9843**	0.9428	0.0007	0.9426
football	1.2	**0.0041**	0.0008	0	0.0040
karate	1.2	0.4956	0.9536	0.0142	**0.9541**
reptilia-tortoise-network-bsv	1.2	**0.9494**	0.1736	0.0007	0.8233
dolphins	1.6	**0.9990**	0.8420	0	0.9889
eco-stmarks	1.6	**0.9884**	0.9872	0.0007	0.9881
football	1.6	**0.7379**	0.0047	0	0.2952
karate	1.6	**0.9541**	0.4949	0.0158	0.4962
reptilia-tortoise-network-bsv	1.6	0.4384	0.3460	0.0007	**0.9801**
dolphins	2.0	0.9319	0.7845	0	**0.9378**
eco-stmarks	2.0	0.8740	0.9716	0.0007	**0.9733**
football	2.0	**0.0832**	0.0029	0	0.0768
karate	2.0	0.8291	0.8699	0.0158	**0.8712**
reptilia-tortoise-network-bsv	2.0	**0.7692**	0.1414	0.0007	0.7508



(a) ROB 1.2 (b) ROB 1.6 (c) ROB 2.0

图 5 RBO 相关系数柱状图

Fig. 5 Histogram of RBO correlation coefficients

(3)算法有效性分析

本节从实验的角度介绍 RBO 相关系数和肯德尔秩相关系数在评估节点影响力排序方面的适用性。

网络节点影响力排序的最终目的是发现网络中高影响力的节点，节点影响力排序只是发现高影响力节点的辅助方式，通过对节点的影响力进行排序，能够较为方便地发现影响力 top-k(k 根据需要指定)的节点。因此在对节点影响力排序结果准确性进行评估时，应当考虑节点次序之间的差别，在准确性评估中对影响力靠前的节点赋予较高的权重。本文以一个具体实例阐释这种赋权评估的必要性，例如，在经典的跆拳道俱乐部网络中对每个节点分别模拟 SIR 传播过程(传染率 $\beta=0.25$，恢复率 $\alpha=1$)2 000 次后取平均值，将各个节点 sir 值按降序排列并绘制成折线图，如图 6 所示。

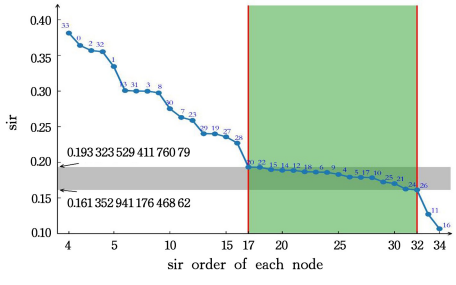
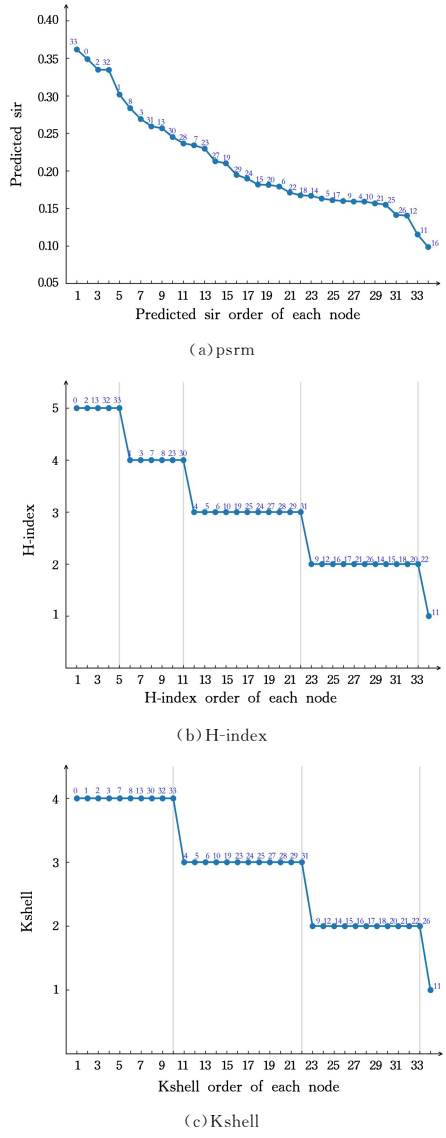


图 6 karate 网络节点 sir 值随排序等级的变化(电子版为彩图)

Fig. 6 sir value of nodes changing with rank order in karate network

图 6 横坐标表示节点 sir 值的排序次序，纵坐标表示节点的 sir 值，图中折线附近的数字分别表示各个节点的编号。从图中可以看出，传播能力排序靠前的节点(图 6 中绿色区域左侧的节点)的 sir 值变化幅度比传播能力排序靠后的节点(图 6 中绿色区域的节点)的 sir 值变化幅度大。如图 6 所示，网络传播能力排名 17－32 位的 16 个节点(图 6 中的灰色区域)的传播能力仅相差 0.032，那么这 16 个节点之间的排序结果出错带来的影响应小于排名靠前的节点之间排序出错带来的影响，因此在评估不同方法排序准确性时排序靠前的节点的权重应该更高。另一方面，在实际应用中，节点影响力排序只是发现网络 top-k 个高影响力节点的手段，因此，传播能力靠前的节点排序是否准确，对实际应用产生的影响更为明显。

为了更直观地展示 RBO 相关系数和肯德尔秩相关系数之间的差异，本小节在跆拳道俱乐部网络上进行了 2 000 次 SIR 模拟，记所得节点影响力排序结果为 R_{sir}。计算 psrm 方法、H-index 方法和 Kshell 方法所得节点影响力排序结果与 R_{sir} 之间的 RBO 相关系数和肯德尔秩相关系数，并绘制了 psrm 方法、H-index 方法和 Kshell 方法所得节点影响力排序结果的折线图，如图 7 所示。



(a) psrm (b) H-index (c) Kshell

图 7 psrm，H-index 和 Kshell 节点影响力排序折线图

Fig. 7 Line chart of nodes' influence ranking results of psrm, H-index and Kshell

从图 7 可以看出，psrm 方法对节点影响力的区分能力优于 H-index 方法和 Kshell 方法。以 top-5 高影响力节点为例，对比图 6 和图 7 可知，psrm 方法不仅发现了影响力排名

前5的节点，并能进一步区分其先后次序；Kshell方法虽然也将影响力排名前5的节点视为最高影响力节点，但具有最高影响力的节点有10个，无法做进一步区分；H-index方法对影响力前5的节点出现了一个误判，如图7所示，误将13号节点的影响力认为大于1号节点，而且同样存在节点影响力区分能力不强的缺陷。从整体变化趋势来看，3种方法均能在一定程度上区分节点影响力，然而通过计算3种方法所得节点影响力排序与 R_{sir} 的肯德尔秩相关系数和RBO相关系数(SIR中重复模拟次数设为2000，感染率设置为1.6倍的传染阈值，RBO中 p 设置为0.95)发现，H-index方法和Kshell方法的肯德尔秩相关系数为负值，psrm的肯德尔秩相关系数虽为正值，但也接近于0，如表3所列。

表3 肯德尔秩相关系数与RBO相关系数

Table 3 Kendall rank coefficients versus RBO coefficients

Ranking methods	Kendall rank coefficients	RBO coefficients
psrm	0.090909	0.946331
Kshell	−0.055258	0.687599
H-index	−0.172905	0.744614

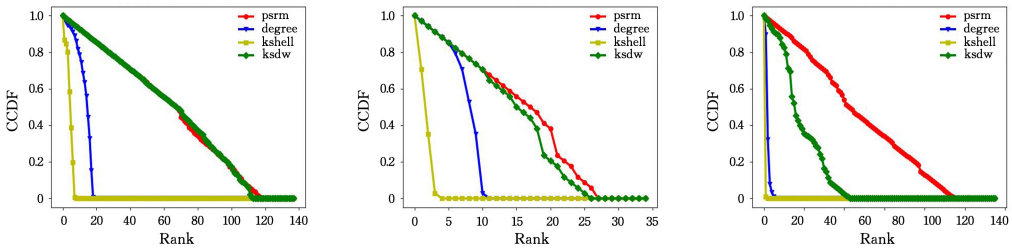
肯德尔秩相关系数接近于0，说明在肯德尔秩相关系数的衡量标准下，3种方法均不能很好地区分节点影响力，与实际情况不符。与肯德尔秩相关系数相比，RBO相关系数在评价方法准确性时给排名靠前的节点赋予较高权重，计算结果表明，3种方法的RBO相关系数均大于0.6，说明3种方法均能在一定程度上区分节点影响力。psrm方法的RBO相关系数最高，说明psrm方法的准确性更好，与实际结果相符。

因此，无论是理论上还是实际应用中，在评估节点影响力排序的准确性时应该赋予影响力排序靠前的节点(top节点)较高的权重。肯德尔秩相关系数在衡量序列的相关性时没有考虑排序位置的差异，而RBO相关系数给影响力排名靠前的节点赋予了更高的权重。因此，相比肯德尔相关系数，RBO相关系数更适合用于对节点影响力排序结果的准确性评估。

结合图4和图5可以看出，PSRM在RBO系数上相比其他方法更有优势，这也说明了基于sir值预测的方法在节点影响力排序方面的优势。

4.4.2 单调性实验

不同排序方式对应的CCDF曲线图如图8所示。



(a) reptilia-tortoise-network-bsv (b) karate (c) football

图8 CCDF曲线图

Fig. 8 Line chart of CCDF

图8中横坐标表示不同的节点等级，纵坐标表示对应的CCDF函数值，从图中可以看出，基于sir值预测的排序方法psrm的单调性最优，其后依次是ksdw，dc，ks。dc和ks排序单调性较差，原因在于网络中具有相同度或相同Kshell值的节点数量较多，对节点影响力的区分粒度较粗，导致最终的结果排序单调性较差，ksdw在衡量节点影响力时考虑了节点及邻居节点的度和Kshell值，区分粒度有了明显提高，而基于sir值预测的方法在节点特征不同时，预测的节点sir值也往往不同，因此单调性较好。

结束语 本文提出了基于sir值学习的节点影响力排序模型，模型的可扩展性高，当更好的描述节点结构特征的指标被提出时，可以直接将其加入到节点特征向量中进行训练。本文使用的深度神经网络模型能够较好地学习到节点特征向量和节点影响力之间的复杂关系，不需要经验知识对节点特征进行赋权，而权重的选择是基于不同权重综合多特征方法难以有效解决的问题。

下一步我们将对本文模型的两个方面做进一步研究。一是在模型的训练中，对于不同的任务，如何选择更有针对性的特征指标以及如何获得相应的标注数据。例如在重要节点发现方面，在不同领域对节点重要性的定义可能不同，当节点的重要性不再体现于信息传播能力上时，譬如在公路交通网中，重要节点可能被定义为对网络鲁棒性影响最大的节点，这时仍然根据SIR模型对数据进行标注是不合适的。二是本文考虑的sir值学习模型是较为基础的深度神经网络模型，如果针对特定学习任务设计更为复杂的、合适的深度学习网络模型，或将有助于提高sir值学习效率及准确率。

参 考 文 献

[1] FANG J Q, WANG X F, ZHENG Z G, et al. A New Interdisciplinary Science: Network Science (I)[J]. Progress in Physics, 2007(3): 239-343.

[2] KEMPE D, KLEINBERG J M, TARDOS É. Maximizing the Spread of Influence through a Social Network[J]. Journal of Chemical Theory and Computation, 2015, 11: 105-147.

[3] MAJI G, MANDAL S, SEN S. A systematic survey on influential spreaders identification in complex networks with a focus on K-shell based techniques[J]. Expert Systems With Applications, 2020, 161(Dec.): 113681.1-113681.18.

[4] ALBERT R, JEONG H, BARABÁSI A L. Diameter of the World-Wide Web[J]. Nature, 1999, 401(6749): 130-131.

[5] FREEMAN L C. A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1977, 40(1): 35-41.

[6] KITSAK M,GALLOS L,HAVLIN S,et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010,6(11):888-893.

[7] PASTOR-SATORRAS R,VESPIGNANI A. Epidemic dynamics and endemic states in complex networks[J]. Physical Review E, 2001,63(6):066117.

[8] ZAREIE A,SHEIKHAHMADI A. A hierarchical approach for influential node ranking in complex social networks[J]. Expert Systems with Applications,2018,93:200-211.

[9] ZHAO J,WANG Y,DENG Y. Identifying influential nodes in complex networks from global perspective[J]. Chaos, Solitons & Fractals,2020,133(7261):109637.

[10] NAMTIRTHA A,DUTTA A,DUTTA B. Weighted kshell degree neighborhood:A new method for identifying the influential spreaders from a variety of complex network connectivity structures[J]. Expert Systems with Application, 2020, 139(Jan.): 112859.1-112859.15.

[11] WEI D,DENG X,ZHANG X,et al. Identifying influential nodes in weighted networks based on evidence theory[J]. Physica A: Statistical Mechanics and its Applications,2013,392(10):2564-2575.

[12] LIU Y,TANG M,ZHOU T,et al. Identify influential spreaders in complex networks,the role of neighborhood[J]. CoRR,2015, 452:289-298.

[13] LÜ L Y,ZHOU T,ZHANG Q M,et al. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications,2016,7(1):10168.

[14] SABIDUSSI G. The centrality index of a graph [J]. Psychometrika,1966,31(4):581-603.

[15] PAGE L,BRIN S,MOTWANI R,et al. The PageRank Citation Ranking:Bringing Order to the Web[R]. Stanford:Stanford InfoLab,1999.

[16] BORGATTI S P,EVERETT M G. Models of core/periphery structures[J]. Social Networks,2000,21(4):375-395.

[17] CHAWLA P,MANGAL R,CHANDRASHEKAR C M. Discrete-time quantum walk algorithm for ranking nodes on a network[J]. arXiv:1905.06575,2020.

[18] BAE J,KIM S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness[J]. Physica A:Statistical Mechanics and its Applications,2014,395:549-559.

[19] LUO J L,JIN J C,WANG L. Evaluation Method for Node Importance in Air Defense Networks Based on Functional Contribution Degree[J]. Computer Science,2018,45(2):175-180,202.

[20] ZENG A,ZHANG C J. Ranking spreaders by decomposing complex networks[J]. Physics Letters A, 2013, 377(14): 1031-1035.

[21] ZAREIE A,SHEIKHAHMADI A,JALILI M,et al. Finding influential nodes in social networks based on neighborhood correlation coefficient[J]. Knowledge-Based Systems, 2020, 194: 105580.

[22] ZHAO N,BAO J,CHEN N. Ranking Influential Nodes in Complex Networks with Information Entropy Method[J]. Complexity,2020,2020(3):1-15.

[23] MAJI G,DUTTA A,CURADO MALTA M,et al. Identifying and ranking super spreaders in real world complex networks without influence overlap[J]. Expert Systems with Applications,2021,179:115061.

[24] NAMTIRTHA A,DUTTA A,DUTTA B. Identifying influential spreaders in complex networks based on kshell hybrid method[J]. Physica A:Statistical Mechanics and its Applications, 2018,499:310-324.

[25] KEELING M J,EAMES K T D. Networks and epidemic models [J]. Journal of the Royal Society, Interface, 2005, 2(4): 295-307.

[26] HORNIK K,STINCHCOMBE M,WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks,1989,2(5):359-366.

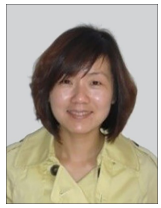
[27] MA L,MA C,ZHANG H F,et al. Identifying influential spreaders in complex networks based on gravity formula[J]. Physica A:Statistical Mechanics and its Applications, 2016, 451: 205-212.

[28] BARABAÁSI A L. Networks Science [M]. Cambridge:Cambridge University Press,2016.

[29] ROSSI R A,AHMED N K. The Network Data Repository with Interactive Graph Analytics and Visualization[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[30] WEBBER W,MOFFAT A,ZOBEL J. A Similarity Measure for Indefinite Rankings[J]. ACM Transactions on Information Systems,2010,28(4):1-38.

DUAN Shunran, born in 1996, postgraduate. His main research interests include social network analysis and so on.



YIN Meijuan, born in 1977, Ph.D, associate professor, master supervisor. Her main research interests include network data analysis and cyberspace security.

(责任编辑:杨雪敏)