Computer Science and Application 计算机科学与应用, 2019, 9(9), 1753-1760
Published Online September 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.99197

Hans汉斯

A Network Embedding Method Based on Graph Folding

Xiaoshuo Feng[1], Dongqi Wang[2*]

[1]Naval Research Academy, Beijing
[2]Software College, Northeastern University, Shenyang Liaoning
Email: 125470246@qq.com, *wangdq@swc.neu.edu.cn

Received: Sep. 2nd, 2019; accepted: Sep. 17th, 2019; published: Sep. 24th, 2019

Abstract

With the full application of information technology, information networks are becoming ubiquitous. Social networks, citation networks, telecommunication networks, and even biological networks have made information network research attract the attention of researchers in many disciplines. Network embedding is a low-dimensional vector representation learning method of nodes that preserves information such as network topology and node content. In the low-dimensional space, network analysis and mining tasks may be easier to solve, and the computational complexity of tasks may also be lowered. This paper designs and implements a network embedding method based on complete subgraph folding. This method regards the *k*-complete subgraphs of the target network as supernodes and uses an arbitrary network embedding algorithm to learn the vector representation of supernodes in the new network of supernodes. Then we use the learned vector representation to initialize the members' vector representation of the supernode, after that, the target network will be fed into an arbitrary network embedding algorithm and learn to get the final vector representation of nodes. In this paper, we select to use Deepwalk algorithm in the experiments. Experiment results show that the proposed method significantly improved the speed of network embedding. At the same time, the node vectors learned by the proposed method also outperformed the original Deepwalk algorithm in selected downstream applications.

Keywords

Network Embedding, Graph Folding, *k*-Clique, Deepwalk

一种基于图折叠的网络嵌入方法

冯晓硕[1]，王冬琦[2*]

[1]海军研究院，北京
[2]东北大学软件学院，辽宁 沈阳

*通讯作者。

文章引用: 冯晓硕, 王冬琦. 一种基于图折叠的网络嵌入方法[J]. 计算机科学与应用, 2019, 9(9): 1753-1760.
DOI: 10.12677/csa.2019.99197

Email: 125470246@qq.com, *wangdq@swc.neu.edu.cn

收稿日期：2019年9月2日；录用日期：2019年9月17日；发布日期：2019年9月24日

摘　要

随着信息技术的广泛应用，信息网络正在变得无处不在，社交网络、引文网络、电信网络乃至生物网络等各类网络让信息网络研究受到了众多学科研究人员的关注。网络嵌入是一种保留网络拓扑信息和节点内容等其他附带信息的网络节点低维向量表示学习方法，在新的低维空间中网络分析挖掘任务可能更容易被解决，任务的运算复杂性也有可能降低。本文设计实现了一种基于完全子图折叠的网络嵌入方法，该方法把目标网络的*k*完全子图视为超节点，在以超节点为单位的新网络上使用任意网络嵌入算法学习超节点的向量表示，之后把超节点的向量表示作为对应*k*-完全子图中所有节点输入到任意网络嵌入学习算法的初始值，重新学习获得节点最终的向量表示。本文使用Deepwalk算法进行了实验，实验结果表明，本方法不但大幅提升了网络嵌入的速度，而且本方法学到的节点向量在一些下游应用中的表现也优于纯粹的Deepwalk算法。

关键词

网络嵌入，图折叠，*k*完全子图，Deepwalk

1. 引言

从社交网络到万维网，图是呈现各种真实世界信息的普遍方式。给定网络结构，通常需要预测与图中每个节点相关联的信息(通常称为属性或标签)。这些信息可以代表数据的各个方面。例如，在社交网络上，它们可以代表一个人所属的社团，或者代表网络上文档内容的类别。在现实的世界中存在着很多大型的网络，这些网络可能包含数十亿个节点和边，因此很难在整个网络上执行复杂的推理过程。为解决这一问题，降维的技术应运而生。其核心思想是将图中的每个节点转换为一个低维的潜在表示，利用这些表示可以在网络图上进行操作，如社团划分、链路预测等。

传统的图降维的方法在小规模网络上效果良好[1]，但是传统的方法存在着复杂度高的问题，因此使得传统的图降维的方法在大规模网络上并不适用。近年来随着图表示学习的发展，DeepWalk [2]、Line [3]、Node2vec [4]等算法被提出，这些基于神经网络的方法已被证明具有高度的可扩展性和较高的性能，在大型网络中的分类和链路预测任务上取得了很好的结果。

基于神经网络的方法虽然取得了较好的效果，但是这些方法存在着一些共同的弱点。它们都是局部方法，更加关注于节点周围的结构，例如 DeepWalk 和 Node2vec 利用短随机游走来获取节点的本地邻居。这种对局部结构的关注导致忽略了全局关系，这就使得学习到的表示可能忽略或根本无法发现重要的全局结构。现实生活中，许多实际网络中通常包含由若干结点组成的完全子图，一些实际网络甚至是由一些完全子图通过公共节点链接而成的[5]，因此在社团划分的角度考虑，在一个完全图内的节点在大概率上可以视为一个社团结构内的节点。因此通过在网络中寻找完全子图进而折叠，是一件很有意义的事情。

在这篇文章中，我们提出了一种基于图折叠的网络嵌入方法，它保留了原有结构的高阶特征。该方法把目标网络的 k 完全子图视为超节点，在以超节点为单位的新网络上使用任意网络嵌入算法学习超节点的向量表示，之后把超节点的向量表示作为对应 k-完全子图中所有节点输入到任意网络嵌入学习算法的初始值，重新学习获得节点最终的向量表示。

2. 相关工作

2.1. 超节点

完全图是每对顶点之间都恰连有一条边的简单图。n 个端点的完全图有 n 个端点及 $n(n-1)/2$ 条边，$(n-1)$-正则图。所有完全图都是它本身的团(clique)。在这里我们把一个 k-完全图视为一个超节点，即保留原有的网络结构不变，将一个完全图中的所有节点折叠为一个结点。这里选取 $k=6$ 为例，如图 1 所示，节点 0、1、2、3、4、5 构成了一个完全图，此外完全图与节点 7、8 相连，我们将 0、1、2、3、4、5 折叠为超节点 sub_0，并与节点 7、8 相连的原有结构保留[6]。

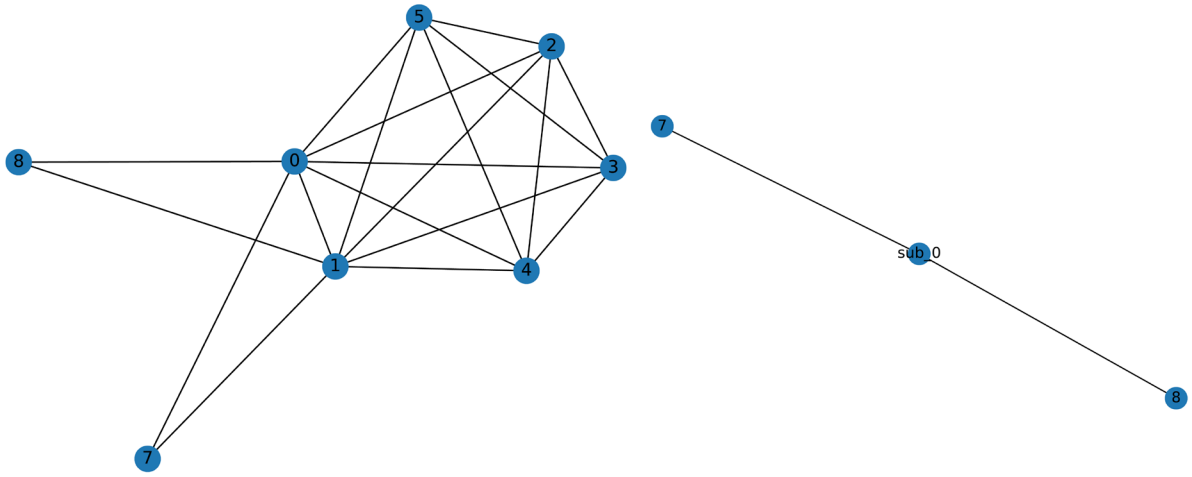

Figure 1. Supernode collapsing diagram
图 1. 超节点折叠示意图

2.2. 折叠规则

社团结构存在于复杂的网络中。节点在社团内紧密相连，而在社团之间的连接相对稀疏。现实的网络结构远远比图 1 所示的情况复杂的多，节点之间存在错综复杂的连接关系。在网络中 $k=3$ 的完全子图普遍存在，$k=3$ 时进行折叠会使整个网络折叠的很小，这样会粗化了网络的局部信息，甚至可能将整个网络折叠为一个超节点，因此在这里 k 至少为 4。

对于完全图的折叠问题归结为以下两类：

1. 完全图与完全图相连；
2. 完全图与非完全图相连。

完全图与完全图相连会存在公共节点，公共节点的个数会影响完全图折叠为超节点的效果，因此针对不同的公共结点的个数定义了不同的折叠规则。如图 2 所示，当公共结点个数为 1 时，认为两个完全图不在一个社团内，因此将两个完全图分别折叠为两个超节点，之后根据原结构将两超节点连接。当公共结点个数为 2 个及以上时，我们认为两个完全图存在于一个社团内，此时将两个完全图折叠为一个超节点，如图 3 所示，以 2 个公共节点为例。

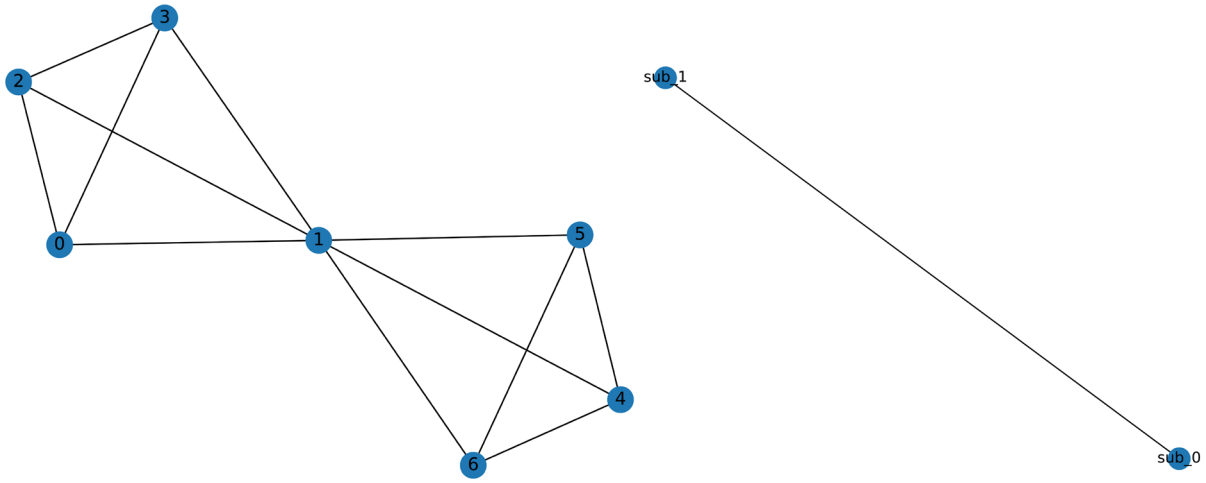

Figure 2. The diagram of one common node
图 2. 公共结点数量为 1 示意图

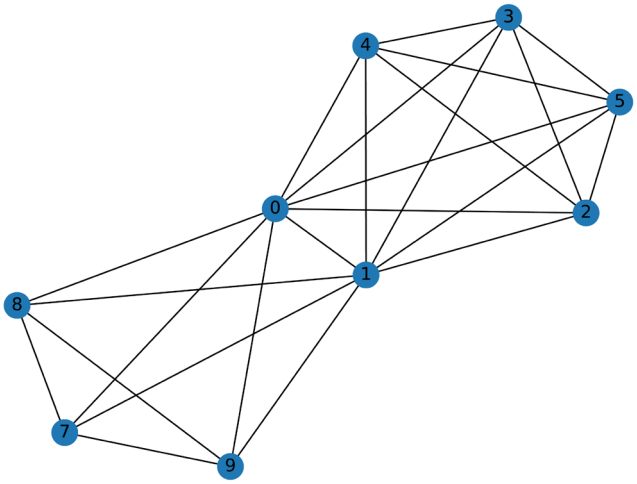

Figure 3. The diagram of two common nodes
图 3. 公共结点数量为 2 示意图

完全图与非完全图相连会存在非完全图的点与完全图中的点构成完全图的情况。如图 4 所示，节点 7 与完全图中的点 0、1、2 构成了一个新的完全图，我们认为两个完全图在一个社团内，因此将两个完全图折叠为一个超节点。若完全图与非完全图相连没有构成新的完全图，即如图 1 所示的情况进行折叠即可。

DeepWalk 算法

据我们所知，DeepWalk 算法[2]是近几年被提出的最具开创性的网络嵌入方法之一，该算法提供了一种能够把网络数据作为神经网络输入的通用表示学习解决方案。假设无权无向网络 $G = (V, E)$，其中 V 是节点集，E 是边集，Deepwalk 主要由两个部分组成：

(1) 对节点 $v \in V$，从 v 开始使用随机游走方法构建由 v 的上下文节点组成的节点序列，其中 v 称为锚点，由扫描形成的节点序列组成训练集；

(2) 使用 SkipGram 算法[7]基于(1)形成的上下文节点序列学习并获得锚点的向量表示。

其中 SkipGram (SG)算法是谷歌提出的 Word2vec 词嵌入方法中的一种算法，词嵌入属于自然语言模

型学习算法，SG 算法与传统语言模型学习算法的不同在于其根据给定的单词去预测去上下文单词。Word2vec 算法的提出激发了相关领域的研究热潮，受该算法启发，Perozzi，B 等把网络的节点看成单词，把通过随机游生成的节点序列看成自然语言的句子，借用 SG 算法解决网络嵌入问题并提出了 Deepwalk 算法。据我们所知，Deepwalk 算法是近几年网络嵌入研究最具有代表性的成果之一，所以本文选择其作为测试对象具有较高的学术价值。

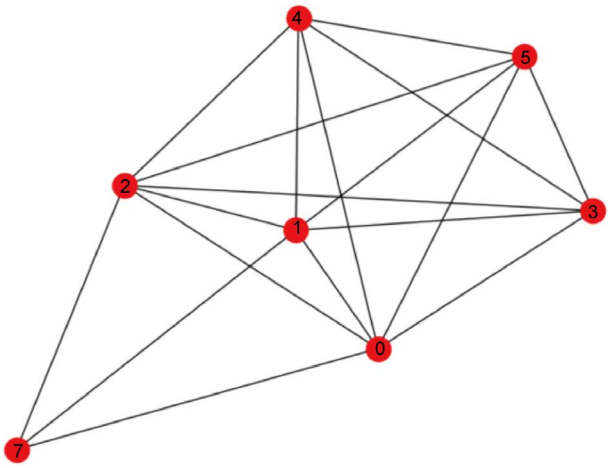

Figure 4. The diagram for generating a new complete graph
图 4. 生成新的完全图示意图

3. 实验

3.1. 实验设计

为了测试本文提出的基于图折叠的网络嵌入方法的有效性，我们在常用的两个数据集上进行实验，现对实验进行了如下设置。首先基于图折叠的方法将原始网络折叠为一个包含超节点的新网络，在新网络上使用 DeepWalk 算法学习新网络的节点表示，其中新网络中超节点的向量表示作为对应超节点的原始网络节点的向量表示，之后将得到的所有节点向量作为 DeepWalk 的初始向量，重新学习获得所有节点的向量表示。使用这些向量进行社团划分，并将不进行折叠的同一网络进行社团划分，对比二者的 NMI 值。

3.2. 实验数据集

(1) 美国政治图书网络[8]

美国政治图书网络(Polbooks network)来源于 21 世纪初选举美国总统时期，科学家从亚马逊网站上统计出政治图书的销售情况所构成。该网络包括 105 个节点，441 条边，每个节点代表销售的图书，边表示两本书的购买者是同一个人。该网络因为政治观点不同而形成三个群体：自由主义、保守主义以及无明显政治观点群体。

(2) 美国大学足球俱乐部网络[9]

Newman 根据美国大学生足球联赛而创建的一个复杂的社会网络(American College football network)。该网络包含 115 个节点和 613 条边，其中网络中的结点代表足球队，两个结点之间的边表示两只球队之间进行过一场比赛。参赛的 115 支大学生代表队被分为 12 个联盟。

本文选取的两个数据集都具有网络社区结构的地板真值，所谓社区(Community)是网络科学领域的一

个概念[5]，是指由相似性较高的节点聚成的节点子群，子群内部节点间连接紧密，子群之间节点间连接稀疏。NMI (Normalized Mutual Information)标准化互信息[10]，常用在聚类任务中度量聚类结果与实际聚类真值的相近程度。在把网络节点表示成向量之后对节点聚类实际上是一种社团划分操作，因此可以使用 NMI 来评估地板真值与划分结果的接近程度。NMI 的值域是 0 到 1，越高代表划分得越准，越低代表划分的效果越差，其计算方法如公式(1)所示，其中 Y 是节点所属社团的地板真值标签，C 是聚类结果中节点所属社团标签，H 是信息熵计算，$I(Y;C)=H(Y)-H(Y|C)$ 被称为交互信息量[10]。

$$\mathrm{NMI}(Y,C)=\frac{2\times I(Y;C)}{\left[H(Y)+H(C)\right]} \tag{1}$$

3.3. 实验结果

根据所提出的折叠原则，两个数据集折叠后的效果分别如图 5、图 6 所示，不难看出美国政治图书网络具有明显的社团结构，与之对应折叠后的新网络也具有与原始网络大致相似的社团结构；美国大学足球俱乐部网络实际情况为 12 个社团，折叠之后网络中超节点的个数与实际社团数接近，这说明折叠的原则是可行的。

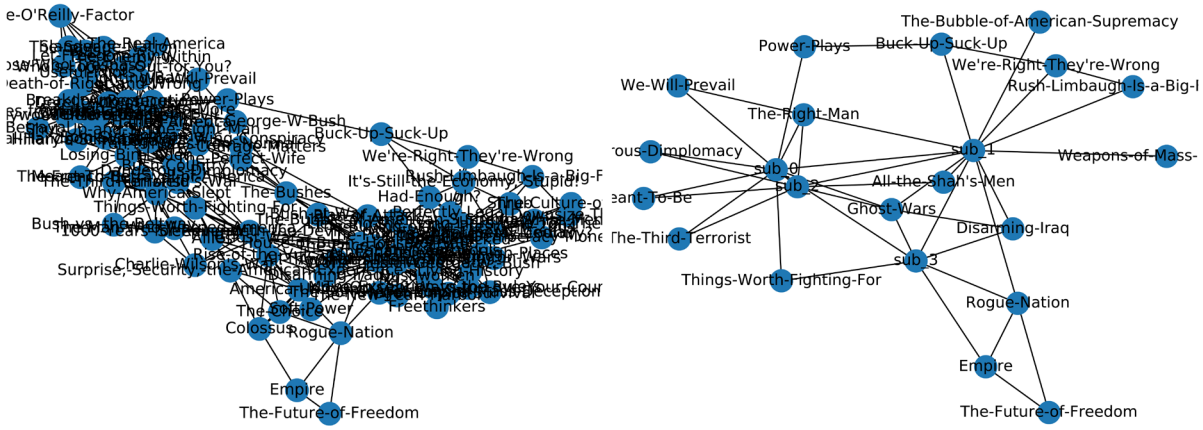

Figure 5. Folding graph of the Polbooks network
图 5. 美国政治图书网络折叠图

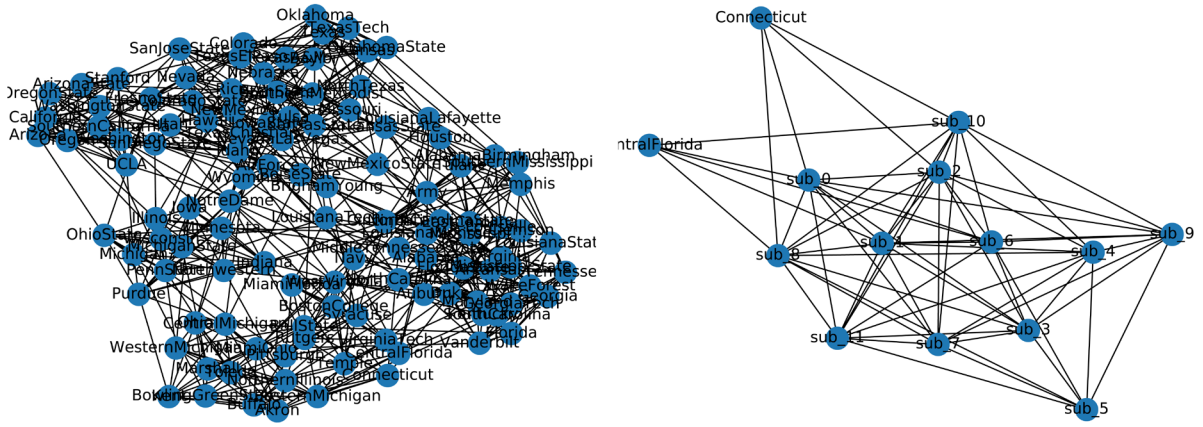

Figure 6. Folding graph of the American College football network
图 6. 美国大学足球俱乐部网络折叠图

我们进行了比较实验。在实验中分别设置完全子图中节点个数 k 为 4、5、6。结果如下所述。

Table 1. System resulting data of standard experiment
表 1. 标准试验系统结果数据

数据集	$k = 4$	$k = 5$	$k = 6$
Polbooks	62.14%	62.05%	61.89%
Football	94.55%	93.88%	93.28

由表 1 所示，对于 Polbooks 及 Football 数据集，传统的 Deepwalk 网络嵌入后进行社团划分的 NMI 结果分别为 62.03%和 93.26%，基于图折叠的网络嵌入方法对其进行社团划分，通过选取不同完全子图中节点个数，来计算 NMI 的值。当节点个数 k 为 4 时 NMI 值最高，在两个数据集上得到的结果均优于传统的方法，这时便得到了最优的社团划分结果。在 k 取 5 和 6 的时候相比于传统方法也并不逊色。根据表中数据可以发现，伴随着完全子图中节点数 k 的增加，相应的社团划分效果有所下滑，这是由于 k 的增大导致在折叠为超节点时，图折叠的效果随之明显，折叠之后的网络忽略了一些局部信息。

此外网络的嵌入速度相比于纯粹的 Deepwalk 算法，在 Polbooks 及 Football 数据集上分别提升了 21%和 18%。

4. 结论

图表示学习是复杂网络中的一个重要研究方向。在本文中，我们提出了一种图折叠的新方法，进而提出了基于图折叠的网络嵌入方法。实验结果表明，该方法具有运算速度快的特点，并且所得到的向量表示更优，在社团划分等下游应用中效果更好。

基金项目

辽宁省博士启动基金 20170520358。

参考文献

[1] Belkin, M. and Niyogi, P. (2002) Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Vancouver, 3-8 December 2001, 585-591.

[2] Perozzi, B., Al-Rfou, R. and Skiena, S. (2014) Deepwalk: Online Learning of Social Representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 701-710. https://doi.org/10.1145/2623330.2623732

[3] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. and Mei, Q. (2015) Line: Large-Scale Information Network Embedding. In: *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, Geneva, 1067-1077. https://doi.org/10.1145/2736277.2741093

[4] Grover, A. and Leskovec, J. (2016) node2vec: Scalable Feature Learning for Networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 855-864. https://doi.org/10.1145/2939672.2939754

[5] Newman, M. (2018) Networks. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198805090.001.0001

[6] Hu, Y. (2005) Efficient, High-Quality Force-Directed Graph Drawing. *Mathematica Journal*, **10**, 37-71.

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Volume 2, Lake Tahoe, 5-10 December 2013, 3111-3119.

[8] Newman, M. (2013) Books about US Politics. http://www-personal.umich.edu/~mejn/netdata

[9] Girvan, M. and Newman, M.E. (2002) Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences*, **99**, 7821-7826. https://doi.org/10.1073/pnas.122653799

[10] Estévez, P.A., Tesmer, M., Perez, C.A. and Zurada, J.M. (2009) Normalized Mutual Information Feature Selection. *IEEE Transactions on Neural Networks*, **20**, 189-201. https://doi.org/10.1109/TNN.2008.2005601