**基于距离动态的社区检测**

**[Community Detection based on Distance Dynamics](http://dm.uestc.edu.cn/publication/)**



**目 录**

[摘要(Abstract) 4](#_Toc2334)

[1.引言(Introduction) 6](#_Toc7429)

[1.1基本思想(Basic Idea) 9](#_Toc11802)

[1.2贡献(Contributions) 12](#_Toc16813)

[2.相关工作(Related Work) 15](#_Toc27371)

[3. 基于距离动态的社区检测(Community Detection Based On Distance Dynamics) 17](#_Toc23040)

[3.1距离动态与用户定义的社区标准(Distance Dynamics versus User-defined Community Criteria) 17](#_Toc18501)

[3.2预备(Preliminaries) 19](#_Toc5912)

[3.3本地交互模式（Local Interaction Model） 23](#_Toc14638)

[3.4吸引子算法(The Attractor Algorithm) 28](#_Toc21740)

[3.5复杂性分析Complexity Analysis 31](#_Toc12958)

[4. 实验(Experiments) 34](#_Toc8689)

[4.1综合网络(Synthetic networks) 38](#_Toc10542)

[4.2现实世界数据(Real World Data) 42](#_Toc2185)

[(1) 具有类信息的网络(. Networks with class information) 46](#_Toc27960)

[(2)无类信息的网络(. Networks without class information) 49](#_Toc20760)

[4.3小社区和异常检测( Small Community and Anomaly Detection) 52](#_Toc24862)

[4.4 运行(Runtime) 52](#_Toc27080)

[5.结论(Conclusions) 54](#_Toc8843)

[6.致谢(Acknowledges) 54](#_Toc18735)

[7.参考文献(REFERENCES) 55](#_Toc18151)

**ABSTRACT**

How can we uncover the natural communities in a real-world network that allows insight into its underlying structure and also potential functions? In this paper, we introduce a new community detection algorithm, called Attractor, which automatically spots communities in a network by examining the changes of “distances” among nodes (i.e. distance dynamics).The fundamental idea is to envision the target network as an adaptive dynamical system, where each node interacts with its neighbors. The interaction will change the distances among nodes, while the distances will affect the interactions. Such interplay eventually leads to a steady distribution of distances, where the nodes sharing the same community move together and the nodes in different communities keep far away from each other. Building upon the distance dynamics, Attractor has several remarkable advantages: (a) It provides an intuitive way to analyze the community structure of a network, and more importantly, faithfully captures the natural communities (with high quality). (b)

Attractor allows detecting communities on large-scale networks due to its low time complexity (O(|E|)). (c) Attractor is capable of discovering communities of arbitrary size, and thus small-size communities or anomalies, usually existing in real-world networks, can be well pinpointed. Extensive experiments show that our algorithm allows the effective and efficient community detection and has good performance compared to state-of-the-art algorithms.

**Keywords**

Community detection; Interaction model; Network

# 摘要(Abstract)

我们如何发现现实世界网络中的自然社区，从而能洞察其底层结构和潜在的功能？在本文中，我们引入了一种称为吸引子(Attractor)的新的社区检测算法，它通过检查节点之间的“距离”的变化（即距离动态）来自动识别网络中的社区。基本思想是将目标网络设想为自适应动态系统，其中每个节点与其邻居交互。 交互将改变节点之间的距离，而距离将影响相互作用。其中共享同一社区的节点一起移动，并且不同社区中的节点彼此远离，这种相互作用最终导致距离的稳定分布。基于距离动态，吸引子具有几个显着的优势：（a）它提供了一种直观的方法来分析网络的社区结构，更重要的是忠实地捕捉到自然社区（高品质）。（b）吸引子允许在大规模网络上检测社区，因为它的时间复杂度低（O（| E |））。（c）吸引子能够发现任意大小的社区，因此通常存在于现实世界网络中的小规模社区或异常现象可以被精确定位。广泛的实验表明，我们的算法允许有效和高效的社区检测，并且与最先进的算法相比具有良好的性能。

**关键词**

社区检测 互动模式(Interaction model) 网络

**1.INTRODUCTION**

During the past decades, community detection [3] (also called graph clustering or graph partitioning) has attracted a lot of attention. Many approaches have been proposed to identify communities based on different criteria (e.g. betweenness [5], normalized cut [16], modularity [11], etc.), and each criterion comes to specific advantages and drawbacks. As an example, the wide-spread modularity based algorithms [11], only yield a good graph partitioning if the network follows the random null assumption that each node has the equal chance to link any other node of the network[4]. This assumption becomes unreasonable for large networks (usually called “resolution limit”) as their connectivity patterns are usually in a local instead of a global fashion. Moreover, the growing large-scale networks in diverse field are posing an increasing challenge for most established community detection algorithms. Therefore, how to identify the community structure in large-scale networks effectively and efficiently remains a big data mining task to date.

In this paper, instead of introducing a new user-defined criterion for community detection like normalized cut [16] or modularity [11], we consider the problem of community detection from a new point of view: distance dynamics. We will see this new viewpoint supplements an intuitive way to identify community structure, and has several attractive properties. But let us first illustrate the basic idea.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

KDD’15, August 10-13, 2015, Sydney, NSW, Australia.

©2015 ACM. ISBN 978-1-4503-3664-2/15/08 ...$15.00

DOI: <http://dx.doi.org/10.1145/2783258.2783301.>

# 1.引言(Introduction)

在过去几十年中，社区检测[3]（也称为图形聚类或图分割）已引起了广泛关注。已经提出了许多方法来根据不同的标准来识别社区（例如中间性[5]，归一化剪切[16]，模块化[11]等），并且每个标准都具有特定的优点和缺点。比如，基于广泛模块化的算法[11]只有在网络遵循随机空值假设时才产生良好的图形分割，每个节点具有相等的机会连接网络的任何其他节点[4]。这种假设对于大型网络（通常称为“分辨率限制”）变得不合理，因为它们的连接模式通常是局部的，而不是全局的。此外，不断增长的大型网络在大多数已建立的社区检测算法中也面临越来越大的挑战。因此，如何有效，高效地识别大型网络中的社区结构仍然是迄今为止的大型数据挖掘任务。

在本文中，我们从新角度考虑了社区检测的问题，而不是为社区检测引入新的用户定义标准，如标准化[16]或模块化[11]。 我们将看到这个新观点补充了一种直观的方式来识别社区结构，并具有几个有吸引力的属性。但是让我们先来说明一下基本思想。

授予个人或教室使用的全部或部分本作品的数字化或复印件的许可，无需付费，只要副本不是为了利润或商业利益而制作或分发的，并且副本在第一页上承担本通知和完整引用。本作品的版权由ACM所有者必须得到尊重。允许提供信用证。要复制或重新发布，在服务器上发布或重新分发到列表，需要事先具体的许可/费用。请求Permissions@acm.org的权限。

KDD'15，2015年8月10日至13日，悉尼，新南威尔士州澳大利亚。

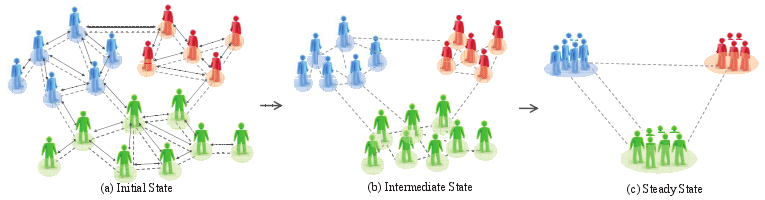
©2015 ACM. ISBN 978-1-4503-3664-2/15/08 ...$15.00

DOI: <http://dx.doi.org/10.1145/2783258.2783301.>

**1.1 Basic Idea**

From the view of sociology, a “community” can be perceived as a group of persons who are connected to each other by relatively durable social relations to form a tight and cohesive social entity, due to the presence of a “unity of will ” or “sharing common values” [7]. It is thus curious for us to know whether the community structure can be automatically revealed by simulating the degree of cohesiveness of persons over time. Namely, we expect all persons in the same community gradually enhance the cohesiveness by influencing each other, and finally converge together (e.g. same opinion, common values, etc.). Inspired by such perception, we present a new method to shed light on the compartmental organization of a given network from the perspective of distance dynamics. The basic idea is to view a network as an adaptive dynamical system, and investigate its dynamics over time. Here, instead of exploiting the node dynamics like traditional dynamical systems in physics, we examine the changes of “distances” among nodes (i.e. distance dynamics) over time. Driven by the local topology-driven interactions (cf. Section 3.3), the distances among nodes change gradually, and often exhibit two distinct ways as time evolves, where the distances among nodes in the same communities tend to decrease while those in different communities increase. Finally, all distances achieve a stable pattern. We call the stable pattern as an attractor, a conceptual metaphor that all nodes attract their adjacent nodes, resulting in the nodes sharing the same communities move together while those in different communities keep far away from each other. Therefore, the community structure finally pops up automatically.

Building upon our proposed interaction model (cf. Section3), the dynamics of each distance between two connecting nodes can be described in the following three phases: Firstly, each distance starts with an initial value. Secondly, as time evolves, each distance gradually shrink or stretch relying on its local topological structure. Finally, each distance converges (0 or 1 eventually). And as a result, the network will be naturally split into several distinct communities by simply removing the edges associated with distances equivalent to 1. To better illustrate the basic idea, let us take a social network as an example. Fig. 1 displays the distance dynamics of an artificial social network composing of a number of persons and a set of inter-relationships (dashed lines). In this network, there exist three groups (representing as cartoon people with different colors) based on their different hobbies. Supposing some new techniques have been plugged into a mobile phone, and persons in this network are discussing their cons and pros. We are interested in knowing how opinion disparities among persons evolve driven by the underlying structure over time. In the beginning, each person usually has their own ideas, and the disparities of opinions with his neighbors are thus different (First phase: initial opinion disparities among people, see Fig. 1(a)). Due to the influence from his/her known persons (i.e. persons having relationships), the disparities of opinions among these persons gradually change (increase or decrease) over time (Second phase: the simulation of distance dynamics, see Fig. 1(b)). Finally the opinion disparities of all people tend to converge, and three communities naturally pop up in terms of the “distances” among persons (Third phase: stable pattern, see Fig. 1(c)).

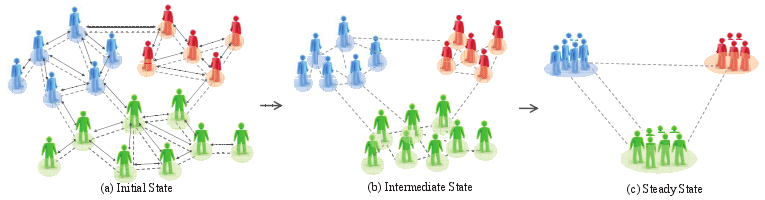


**Figure 1: The illustration of community detection based on distance dynamics. (a) A social network, where the dashed lines indicate the relationships among persons, and arrows demonstrate the direct mutual interactions based on their relationships. (b) Relying on a proposed interaction model, the “distances” among people will change over time, where persons in the same community tend to gradually move together while people in different communities will keep far away from each other. (c) The steady state of persons in terms of the “distances”: three intuitive communities.**

## 1.1基本思想(Basic Idea)

从社会学的角度来看，“社区”可以被认为是通过相对持久的社会关系相互联系形成一个紧密和凝聚力的社会实体的一群人，因为存在“意志统一”或“共享共同价值观”[7]。因此，我们知道社区结构是否可以通过模拟人们随时间的凝聚力的程度而自动显露出来。也就是说，我们期望同一社区的所有人通过相互影响逐渐增强凝聚力，最终汇合在一起（例如同一观点，共同价值观等）。在这种感觉的启发下，我们提出了一种从距离动态的角度阐明给定网络的区间组织的新方法。基本思想是将网络视为自适应动态系统，并随时间调查其动态。在这里，我们不是像物理学中的传统动力学系统那样利用节点动力学，而是随着时间的推移来研究节点之间的“距离”（即距离动力学）的变化。由本地拓扑驱动的交互驱动（参见3.3节），节点之间的距离逐渐变化，并且随着时间的推移经常表现出两种截然不同的方式，同一社区的节点之间的距离倾向于降低，而不同社区的节点间距离增加。最后，所有距离都达到了稳定的格局。我们将稳定模式称为吸引子，即所有节点吸引相邻节点的概念隐喻(a conceptual metaphor)，导致共享相同社区的节点一起移动，而不同社区中的节点彼此远离。因此，社区结构最终会自动弹出。

基于我们提出的交互模型（参见第3节），可以在以下三个阶段描述两个连接节点之间的每个距离的动力学：首先，每个距离以初始值开始。其次，随着时间的推移，每个距离依赖于其局部拓扑结构逐渐缩小或拉长。最后，每个距离收敛（最终0或1）。因此，通过简单地移除与等于1的距离相关联的边，网络将自然地分成几个不同的社区。为了更好地说明基本思想，让我们以社交网络为例。图1显示由多人组成的人造社交网络和一组相互关系（虚线）的距离动态。在这个网络中，根据不同的兴趣分成三组（用不同颜色的卡通人物代表）。假设一些新技术已经插入到手机中，而这个网络中的人正在讨论他们的缺点和专长。我们有兴趣知道人们之间的意见差异如何随着时间的推移而由基础结构驱动。一开始，每个人通常都有自己的想法，与邻居的意见差距是不同的（第一阶段：人们初步意见差异，见图1（a））。由于他/她的知名人士（即有关系的人）的影响，这些人之间的意见差距随时间逐渐变化（增减）（第二阶段：模拟距离动态，见图1（b） ）。最后，所有人的意见差距趋于收敛，三个社区在人际“距离”方面自然弹出（第三阶段：稳定模式，见图1（c））。

图1：基于距离动态的社区检测示意图。 （a）一个社交网络，虚线表示人与人之间的关系，而箭头则表现出基于他们的关系的直接相互作用。 （b）依靠提出的交互模型，人们之间的“距离”将随着时间的推移而变化，同一社区的人们往往逐渐走向一起，而不同社区的人们将彼此远离。 （c）“距离”方面的稳定状态：三个直观的社区。

**1.2 Contributions**

By simulating the distance dynamics, Attractor has several attractive benefits for community detection in networks, most importantly:

**•Intuitive Community Detection:** Instead of optimizing user-specified measures, Attractor investigates the community structure in networks from a new point of view: distance dynamics. Building upon three proposed interaction patterns, Attractor allows automatically spotting communities intuitively, and more importantly, faithfully finding high-quality communities (cf. Fig. 5 - Fig. 9, Table 2 and 3).

**•Small Community and Anomaly Detection:** Relying on the local topology-driven dynamic interactions, the small communities or anomalies usually existing in large-scale networks can be well identified as Attractor allows discovering arbitrary-size communities (cf. Fig. 10, Table 2 and 3).

**•Scalability:** Thanks to the local interaction model, Attractor only needs to investigate the distances of linked nodes over time, which results in a relatively low time complexity of O(|E|) (cf. Section 3.5, Fig.11). This property of Attractor lends itself to handling large real-world networks.

The remainder of this paper is organized as follows: In the following section, we briefly survey related work. Section3 presents our algorithm in detail. Section 4 contains an experimental evaluation. We finally conclude the paper in Section 5.

## 1.2贡献(Contributions)

通过模拟距离动态，Attractor对网络中的社区检测有几个有吸引力的好处，最重要的是：

* **直觉社区检测(Intuitive Community Detection)：**Attractor不是优化用户指定的措施，而是从新的角度来研究网络中的社区结构：距离动态。 基于三个提出的交互模式，Attractor允许自动地直观地发现社区，更重要的是忠实地找到高质量的社区（参见图5 - 图9，表2和3）。
* **小社区和异常检测(Small Community and Anomaly Detection)：**依靠本地拓扑驱动的动态交互，通常存在于大型网络中的小社区或异常可以很好地识别为吸引者允许发现任意大小的社区（参见图10，表2和 3）。
* **可扩展性(Scalability)：**由于本地交互模型，Attractor只需要调查连接节点随时间推移的距离，导致O（| E |）的时间复杂度相对较低（参见第3.5节，图11）。 Attractor的这个属性适用于处理大型现实世界网络。

本文的其余部分组织如下：在下一节中，我们简要介绍相关工作。 第3节详细介绍了我们的算法。 第4节包含实验评估。 我们最后在第5节总结了这篇论文。

**2. RELATED WORK**

During the past several decades, many approaches have been established for community detection, such as [8], [16], [15] [11] etc. Due to space limitation, we only report the closest approaches from the literature. For detailed reviews of graph clustering, please refer to [14][3].

**Cut-Criteria Based Community Detection.** The cut-criterion based community detection algorithms refer to a class of widely used techniques which seek to partition a graph into disjoint subgraphs such that the number of “cuts”across the subgraphs is minimized. Wu and Leahy [19] have proposed a clustering method based on the minimum-cut criterion, where the cut between two subgraphs is computed as the total weights of the edges that have been removed. k−disjoint subgraphs are obtained by recursively finding the minimum cuts that bisect the existing segments. To avoid an unnatural bias towards splitting small-sized subgraphs based on the minimum-cut criterion, Shi and Malik [16] have proposed the popular normalized cut, to compute the cut cost as a fraction of the total edge connections to all the nodes in a graph. To optimize this criterion, a generalized eigenvalue decomposition was used to speed up computation time. In many cases, this class of graph clustering algorithms relying on the eigenvector decomposition of a similarity matrix is also called spectral clustering. Although this type of community detection usually allows identifying the communities with high quality, it is not capable of handling large-scale networks. In addition, it is a non-trivial task to determine the suitable number of communities without prior knowledge. Currently, another mainstream strategy to community detection is based on modularity criterion. Modularity is originally introduced to measure the quality of a division of the network according to the “expected cut ”. It is defined as the fraction of the edges that fall within groups minus the expected such fraction in an equivalent network with edges placed at random. In order to get a graph partition with high modularity, modularity-based community detection methods (e.g. [11, 2, 13]) use various kinds of techniques, such as greedy search and simulated annealing, to optimize the quality function. As mentioned in Section 1, studies have demonstrated that modularity-based community detection algorithms tend to fail on many real-world networks due to the “resolution limit”. The situation becomes worse especially when the network size increases.

**Large-Scale Network Handling.** To identify communities on large-scale networks, many algorithms [9, 8, 18] have been proposed during past decades. Metis is a class

of multi-level scalable partitioning techniques proposed by Karypis and Kumar [9], [8]. The graph clustering starts with constructing a sequence of successively smaller graphs, and a bisection of the coarsest graph is applied. Subsequently, a finer graph is generated in the next level based on the previous bisections. At each level, an iterative refinement algorithm such as Kernighan-Lin (KL) or Fiduccia-Mattheyses (FM) is used to further improve the bisection. A more robust overall multilevel paradigm has been introduced by Karypis and Kumar [8], which presents a powerful graph coarsening scheme. It uses simplified variants of KL and FM to speed up the refinement. Thanks to the multilevel graph construction, Metis allows scaling up large-scale networks, however, the quality of resulting communities may suffer in coarsening. The Markov Cluster algorithm (MCL) [18] is another popular algorithm used in life sciences based on the simulation of (stochastic) flow in graphs. MCL allows identifying high-flowing regions representing the communities using random walk.

In contrast to optimizing user-defined measures, here we provide a more intuitive way to investigate the community structure based on distance dynamics, which is not only capable of uncovering the high-quality communities, but also allows handling large-scale networks.

# 2.相关工作(Related Work)

在过去的几十年中，为社区检测建立了许多方法，如[8]，[16]，[15] [11]等。由于空间限制，我们只报告文献中最接近的方法。 有关图形聚类的详细评论，请参见[14] [3]。

**基于标准的社区检测(Cut-Criteria Based Community Detection)。**基于切割标准(cut-criterion)的社区检测算法是指一类广泛使用的技术，其尝试将图分解成不相交的子图，使得跨越子图的“切割”数量被最小化。 Wu和Leahy [19]提出了基于最小切割标准的聚类方法，其中两个子图之间的切割被计算为已被去除的边的总权重。通过递归地找出平分现有段的最小切割获得k-不相交子图。为了避免基于最小切割标准分割小尺寸子图的不自然的偏差，Shi和Malik [16]提出了普遍的归一化切割，在一张图中将削减成本计算为与所有节点的总边连接的一部分。为了优化该标准，使用广义特征值分解来加快计算时间。在许多情况下，这种依赖于相似矩阵的特征向量分解的图聚类算法也称为频谱聚类。虽然这种类型的社区检测通常允许识别高质量的社区，但它不能处理大规模网络。此外，在没有事先知识的情况下确定适当数量的社区是一项不平凡的任务。目前，社区检测的另一个主流策略是基于模块化标准。模块化最初是为了根据“预期切割”来衡量网络划分的质量。它被定义为落在组内的边的分数减去等效网络中预期的这样的分数，其边随机放置。为了获得具有高模块性的图形分区，基于模块化的社区检测方法（例如[11,2,13]）使用诸如贪婪搜索和模拟退火的各种技术来优化质量函数。如第1节所述，研究表明，由于“分辨率限制”，基于模块化的社区检测算法在许多现实网络中往往会失败。情况变得更糟，特别是当网络规模增加时。

**大规模网络处理(Large-Scale Network Handling)。**为了识别大型网络上的社区，在过去几十年中已经提出了许多算法[9,8,18]。 Metis是由Karypis和Kumar [9]提出的一类多级可扩展分区技术[8]。图形聚类开始于构建连续较小图的序列，并应用最粗糙图的二分。随后，基于先前的二分法，在下一级生成更精细的图。在每一级，使用诸如Kernighan-Lin（KL）或Fiduccia-Mattheyses（FM）的迭代细化算法来进一步改善二等分。 Karypis和Kumar [8]提出了一个更强大的整体多层次模式，它提出了一个强大的图形粗化方案。它使用KL和FM的简化变体来加快细化。由于多层次的图形构建，Metis可以扩展大规模网络，然而，由此产生的社区的质量可能会在粗化中受损。马尔可夫簇算法（MCL）[18]是基于图中（随机）流模拟的生命科学中另一种流行的算法。 MCL允许使用随机游走识别代表社区的高流动地区。

与优化用户定义的措施相比，我们提供了一种更直观的方式来调查基于距离动态的社区结构，这不仅能够揭示高质量的社区，而且还可以处理大型网络。

**3. COMMUNITY DETECTION BASED ON DISTANCE DYNAMICS**

**3.1 Distance Dynamics versus User-defined Community Criteria**

Currently, many criteria have been proposed to qualify community structure from different point of view, and each criterion has its own advantages and drawbacks. In this study, instead of introducing new user-defined criterion, we present a new community detection approach based on the distance dynamics. As stated in Section 1.1, the basic philosophy is to envision a network as a dynamic system, and dynamically investigate the distances among adjacent nodes to uncover its community structure. Compared to most existing algorithms, except for the vivid way to community discovery, the new viewpoint also has some additional desirable prosperities. (a) Exploring the distance dynamics provides an intuitive and comprehensive image to model the real-world network dynamics. For example, a community (e.g. friendship network) is usually established and intensified based on the relationships by interactions (e.g. social activities in friendship networks). (b) Without using user-defined measures, communities are discovered automatically driven by the intrinsic local topology of networks. (c) Insight into distance dynamics also offers a generalized way for network mining in metric space instead of vector space. This is quite beneficial to network analysis as the information of real-world networks we usually can gain is their connectivity patterns.

In the following, we start with some preliminary definitions, and then an interaction model is proposed in Section 3.3. Section 3.4 presents the algorithm Attractor in detail, and we analyze its time complexity in Section 3.5.

# 3. 基于距离动态的社区检测(Community Detection Based On Distance Dynamics)

## 3.1距离动态与用户定义的社区标准(Distance Dynamics versus User-defined Community Criteria)

目前，从不同角度出发，提出了许多标准来确定社区结构，每个标准都有其自身的优缺点。在这项研究中，我们提出了一种基于距离动态的新的社区检测方法，而不是引入新的用户定义标准。如第1.1节所述，基本思想是将网络设想为动态系统，并动态调查相邻节点之间的距离，以发现其社区结构。与大多数现有算法相比，除了生动的社区发现之外，新观点还有一些额外的理想繁荣。（a）探索距离动态提供了直观和全面的图像来建模真实网络动态。例如，社区（例如友谊网络）通常基于通过交互（例如友谊网络中的社交活动）的关系来建立和加强。 （b）在不使用用户自定义措施的情况下，社区被网络本地局部拓扑自动驱动。 （c）对距离动态的洞察还提供了在度量空间而不是向量空间中进行网络挖掘的一般化方法。这对网络分析是非常有益的，因为我们通常可以获得的真实世界网络的信息是它们的连接模式。

在下文中，我们从一些初步定义开始，然后在3.3节提出了一个交互模型。 第3.4节详细介绍了Attractor算法，并在3.5节分析了它的时间复杂度。

**3.2 Preliminaries**

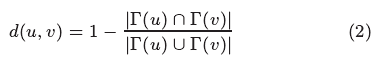
For the purpose of community detection, some necessary definitions are first introduced. Definition 1 (Undirected Graph ) Let G = (V, E, W ) be an undirected graph, where V is the set of nodes, E is the set of edges and W is the corresponding set of weights. e ={u, v} ∈E indicates a connection between the nodes u and v. w(u, v) represents the weight of edge e. ∀e ={u, v} ∈E, w(u, v) = 1, in case of unweighted graph. Definition 2 (Neighbors of node u) Given an undirected graph G = (V, E, W ), the neighborhood of a node u∈V is the set Γ(u) containing node u and its adjacent

nodes.

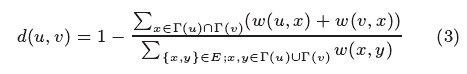
Γ(u) ={v∈V|{u, v} ∈E} ∪ {u} (1)

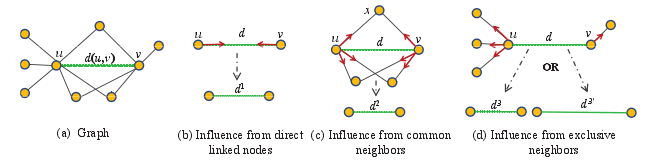
Based on the two definitions, we further use the popular Jaccard distance [6] to quantify the initial distance between two adjacent nodes. Selecting this measure mainly has two reasons. First, Jaccard distance provides an intuitive way to characterize the node similarity. Generally, the more common neighbors two nodes have, the more similar they are. Secondly, Jaccard distance is computed in a local fashion and is thus time efficient. Definition 3 (Jaccard Distance) Given an undirected

graph G = (V, E, W ), the Jaccard distance of two nodes u and v is defined as:



For weighted graph, the Jaccard distance of two nodes u and v is further extended as:





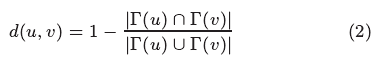
**Figure 2: Illustration of the change of node distances influencing by three distinct interaction patterns.**

## 3.2预备(Preliminaries)

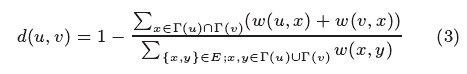
为了社区检测的目的，首先介绍了一些必要的定义。 定义1（无向图）令G =（V，E，W）为无向图，其中V为节点集，E为边集，W为对应的权重集。 e = {u，v}∈E表示节点u和v之间的连接，w（u，v）表示边e的权重。 在未加权图的情况下，∀e= {u，v}∈E，w（u，v）= 1。 定义2（节点u的邻居）给定无向图G =（V，E，W），节点u∈V的邻域是包含节点u及其相邻节点的集合Γ（u）。

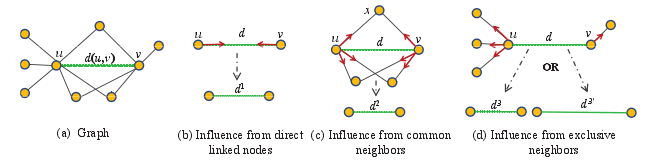
Γ(u) ={v∈V|{u, v} ∈E} ∪ {u} (1)

基于这两个定义，我们进一步使用流行的Jaccard距离[6]来量化两个相邻节点之间的初始距离。 选择这一措施主要有两个原因。 首先，Jaccard距离提供了一种直观的方式来表征节点相似度。 一般来说，两个节点具有更常见的邻居，它们越相似。其次，Jaccard距离以本地方式计算，因此是时间有效的。 定义3（Jaccard距离）给定无向图G =（V，E，W），两个节点u和v的Jaccard距离定义为：



对于加权图，两个节点u和v的Jaccard距离进一步扩展为：





独占邻居的影响

共同邻居的影响

直接连接节点的影响

图2：由三个不同的相互作用模式影响的节点距离变化的图示。

**3.3 Local Interaction Model**

To uncover the community structure in networks based on distance dynamics, we should build up a suitable interaction model. Therefore, the interaction range and interaction patterns need to be first considered.

**Interaction Range.** In order to identify the community structure in networks, the exploring of local topology is essential. Thus, instead of observing the collective interactions, we focus on the distance dynamics in a local way. Obviously, the intrinsic connections (edges) of real-world networks gives a natural way to model the interaction range. Precisely, for each node, it naturally interacts with its adjacent nodes.

**Interaction Patterns.** After specifying the interaction range, the next crucial step is to determine the interaction patterns among nodes to simulate the distance dynamics. Formally, let e ={u, v} ∈E be an edge between two adjacent nodes u and v, and d(u, v) is its initial distance. Obviously, any change of the distance d(u, v) actually results from the variation of node u and node v. In fact, there are three distinct scenarios that allows influencing the distance d(u, v), relying on its local topological structure (see Fig. 2). In the following, we will elaborate how the distance changes in the three different scenarios, respectively.

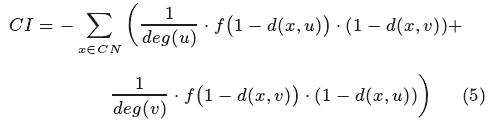
Pattern 1:INFLUENCE FROM DIRECT LINKED NODES. The distance d(u, v) between node u and node v, is obviously influenced by the two direct linked nodes u and v. Through mutual interactions, the one node attracts the other to move towards itself, and thus result in the decrease of d(u, v) (see Fig. 2 (b)). Like a friendship network, each people affects their known people, and tends to increase their cohesiveness gradually (i.e. the “distance” will decrease). Formally, to characterize the change of the distance d(u, v), we define DI , indicating the influence from the interactions of direct linked nodes, as follows:



where deg(u) is the degree of the node u, f (·) is a coupling function and sin(·) is used in this study. 1−d(u, v) indicates the similarity between u and v, and the more similar the two nodes are, the higher influence between each other they will have. The term  is called normalization factor, which is used to consider the different influences between linked nodes with diverse degrees. Namely, the nodes with more links are harder to be influenced comparing to the nodes with less links. Take instructor network as an example. One supervisor usually links to many students while one student only connects to his supervisor. In this situation, the supervisor may have a high influence on each student while the influence for supervisor from each student is relatively low.

Pattern 2: INFLUENCE FROM COMMON NEIGHBORS Here we consider the second scenario: the influence from the common neighbors CN = (Γ(u)−u)∩(Γ(v)−

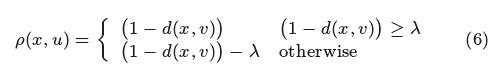
v) of nodes u and v (Fig. 2(c)). As the common neighbors have both links with the two nodes u and v, they attract the two nodes and thus result in the change of the distance d(u, v). Specifically, each common neighbor attracts both node u and node v to move towards itself, and thus leads to the decrease of the distance d(u, v) (See Fig. 2(c)). Formally, we define the change of d(u, v) from the influence of common neighbors, CI , as follows:



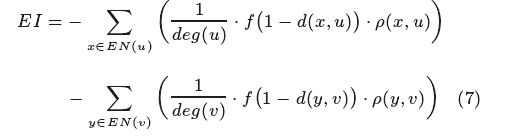
Here the two terms (1−d(x, v)) and (1−d(x, u)) for each common neighbor are used to further quantify the degree of influence compared to the influence from direct linked nodes. For example, considering a common neighbor x interacting with node u (see Fig. 2(c)), if x and v are more similar, the influence from x on u is more similar to the influence from v. Theoretically, once the similarity between x and v equals one (i.e. they can be viewed as the same node), the influence of the node x on the distance d(u, v) simply transfers into the first pattern.

Pattern 3:INFLUENCE FROM EXCLUSIVE NEIGHBORS: The third interaction pattern happens when there exists some neighbors exclusively belong to node u or v, EN (u) =Γ(u)−Γ(u)∩Γ(v) , EN (v) = Γ(v)−Γ(u)∩ Γ(v), respectively. Although, like pattern 1 and pattern 2, each exclusive neighbor of u attracts u to move close to itself, there is no knowledge whether node u is attracted to move closer to node v or attracted to move far away from v (see Fig. 2(d)). To determine the positive or negative influence of exclusive neighbors on the distance, a similarity-based heuristic strategy is proposed. The basic philosophy is to investigate whether each exclusive neighbor of node u is similar with node v, and vice versa. If the exclusive neighbor of node u is similar with node v, the movement of node u towards exclusive neighbor results in the decrease of the distance d(u, v). Similarly, If the exclusive neighbor is not similar with node v, the movement of node u towards the exclusive neighbor will

lead to the opposite effect: moving far away from the node v. Therefore, here we introduce a cohesion parameter λ, to determine the underlying influence as follows. The cohesion parameter λ will be further discussed in Section 3.4.



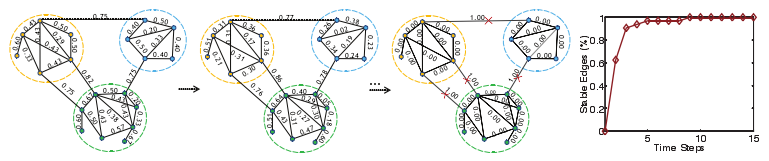
where ρ(x, u) characterizes the degree of positive or negative influence on the distance d(u, v). Then, the change of d(u, v) influencing by exclusive neighbors, EI , is defined as follows:



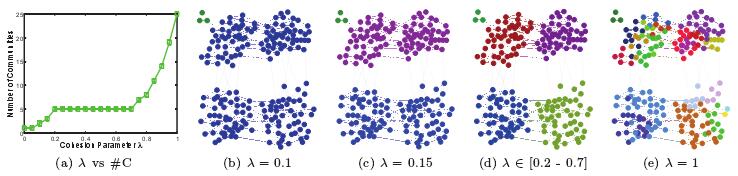
Finally, by considering the three interaction patterns together, the dynamics of the distance d(u, v) between nodes u and v over time is govern by:

d(u, v, t + 1) = d(u, v, t) + DI (t) + CI (t) + EI (t) (8)

where d(u, v, t + 1) is the renewed distance at time step t + 1. DI (t), CI (t) and EI (t) characterize the changes of distance from the direct linked nodes, common neighbors and exclusive neighbors, respectively.



**Figure 3: Illustration of the distance dynamics. (a) The graph representation of the social network of Fig.1(a), where the numbers on edges indicate the initial distances among connected nodes. (b) The updated node distances after one time step. (c) The final state of the network.**



**Figure 4: The sensitivity of cohesion parameter λ on community detection.**

## 3.3本地交互模式（Local Interaction Model）

为了发现基于距离动态的网络社区结构，我们应该建立一个合适的交互模型。因此，需要首先考虑交互范围和交互模式。

**互动范围** 为了识别网络中的社区结构，探索本地拓扑是至关重要的。 因此，我们不是观察集体交往，而是以当地的方式关注距离动态。 显然，现实世界网络的内在联系（边）给出了建立交互范围的自然方式。 精确地说，对于每个节点，它自然地与其相邻的节点进行交互。

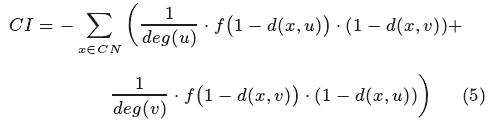
**互动模式** 在指定交互范围之后，下一个关键步骤是确定节点之间的交互模式来模拟距离动态。 形式上，令e = {u，v}∈E是两个相邻节点u和v之间的边，d（u，v）是其初始距离。 显然，距离d（u，v）的任何变化实际上是由于节点u和节点v的变化而产生的，实际上，依赖于其局部拓扑结构（见图2）有三种不同的情景可以影响距离d（u，v）。 在下文中，我们将分别阐述三种不同场景中距离的变化。

**模式1：直接连接的影响。**节点u和节点v之间的距离d（u，v）明显地受到两个直接连接节点u和v的影响，通过相互作用，一个节点吸引另一个节点朝向自身移动，从而导致 d（u，v）（见图2（b））的减少。 像友谊网一样，每个人都会影响他们的知名人士，并逐渐增加他们的凝聚力（即“距离”会减少）。 形式上，为了表征距离d（u，v）的变化，我们定义DI，指示直接连接节点的相互作用的影响如下：



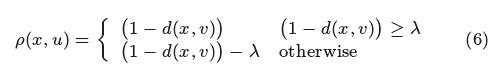
其中deg（u）是节点u的度，f（·）是一个耦合函数，而sin（·）被用于本研究。 1-d(u,v)表示u和v之间的相似度，两个节点之间的相似度越高，它们之间的影响越大。术语 称为归一化因子，用于考虑不同程度的连接节点之间的不同影响。 也就是说，与链路较少的节点相比，具有更多链路的节点更难受影响。 以教员网络为例。 一名主管通常连接到许多学生，而一名学生只能连接到他的主管。 在这种情况下，主管对每个学生的影响可能很大，而每位学生的主管影响相对较小。

**模式2：共同邻域的影响。**在这里我们考虑第二种情况：节点u和v的公共邻居CN =（Γ（u）-u）∩（Γ（v）-v）的影响（图2（c ））。 由于共同的邻居与两个节点u和v都有连接，它们会吸引两个节点，从而导致距离d（u，v）的变化。 具体来说，每个公共邻居都吸引节点u和节点v向自身移动，从而导致距离d（u，v）的减小（见图2（c））。 正式地，我们根据共同邻居的影响CI定义d（u，v）的变化，如下：

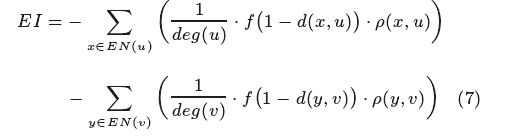


这里使用每个公共邻居的两个项（1-d（x，v））和（1-d（x，u））来进一步量化与直接连接节点的影响相比的影响程度。例如，考虑到与节点u相互作用的公共邻居x（参见图2（c）），如果x和v更相似，x对u的影响更接近于v的影响。理论上，一旦x和v之间的相似等于1（即它们可以被视为相同的节点），节点x对距离d（u，v）的影响简单地转移到第一模式中。

**模式3：独有邻居的影响：**当存在一些邻居仅属于节点u或v时，第三种交互模式发生，EN（u）=Γ（u）-Γ（u）∩Γ（v），EN（v） =Γ（v）-Γ（u）∩Γ（v）。虽然像模式1和模式2一样，u的每个独有邻居都吸引u靠近自身，但是不知道节点u是否被吸引到更靠近节点v的位置，或者被吸引远离v移动（见图2（d））。为了确定独有邻居对距离的正面或负面影响，提出了一种基于相似性的启发式策略。基本思想是调查节点u的每个独有邻居是否与节点v相似，反之亦然。如果节点u的独有邻居与节点v相似，则节点u向独有邻居的移动导致距离d（u，v）的减小。类似地，如果独有邻居与节点v不相似，则节点u朝向独占邻居的移动将导致相反的效果：远离节点v移动，因此，这里我们引入内聚参数λ来确定潜在影响如下。内聚参数λ将在3.4节进一步讨论。



其中ρ（x，u）表示对距离d（u，v）的正或负影响程度。那么，根据独有邻居的影响EI定义d（u，v）的变化，如下：



最后，通过一起考虑三种交互模式，随着时间的推移，节点u和v之间的距离d（u，v）的动态由（8）决定：

d(u, v, t + 1) = d(u, v, t) + DI (t) + CI (t) + EI (t) (8)

其中d（u，v，t + 1）是时间步长t + 1处的更新距离。DI（t），CI（t）和EI（t）分别表示直接连接节点，共同邻居和独有邻居的距离变化。

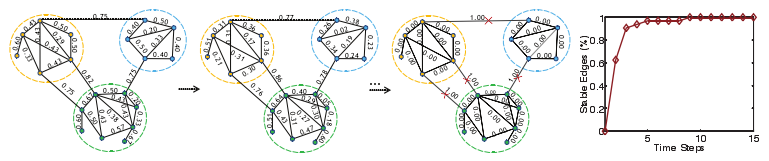
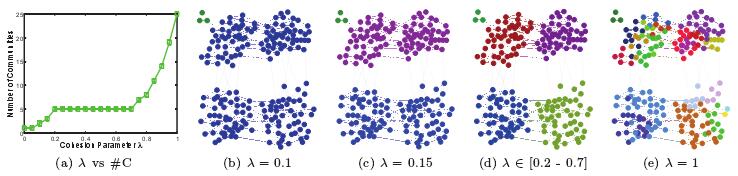
图3：距离动力学的图示。 （a）图1（a）的社交网络的图形表示，其中边数字表示连接的节点之间的初始距离。 （b）经过一段时间后更新的节点距离。 （c）网络的最终状态。

图4：内聚参数λ对社区检测的敏感性。



**3.4 The Attractor Algorithm**

In this section, we present the Attractor algorithm in detail.

**Dynamic Interaction.** Building upon the interaction model (cf. Eq. (8)), the distance dynamics can be simulated, which mainly involves the following steps:

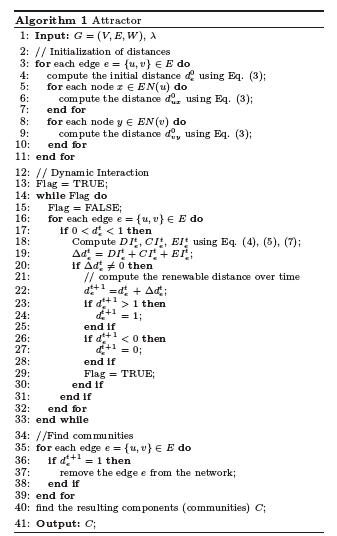
1. At initial time (t = 0), without any interaction, each edge is associated with an initial distance. Here, the initial value is computed according to the Jaccord distance with Definition 2 or Definition 3.

2. As time evolves, relying on local intrinsic topological structure, the dynamics of each distance is simulated according to the three proposed interaction patterns (Eq.(4), Eq.(5) and Eq.(7)). Thanks to the topological-driven influences, the distances among nodes sharing the same community tend to decrease while those in different communities increase gradually.

3. Finally, all distances will converge, and the communities can be easily obtained by removing the edges with maximal distances (i.e. d(u, v) =1).

For illustration, Fig. 3(a)-(c) shows three states for the social network of Fig. 1 from t = 0 to t = 9 during the local dynamic interaction process. T = 0 indicates the initial distances among connected nodes (Fig. 3(a)). From that moment on, each node interacts with its neighbors and influences the corresponding distances based on the proposed interaction model (cf. Eq. 8), and the new distances after one time step are further illustrated in Fig. 3(b). After nine steps, all distances converge, either 0 or 1, and three communities are naturally identified by cutting out all edges with distances equaling to 1.

**Detection of small communities or Anomalies.** In real-world networks, there usually exist many communities with various sizes. Especially in large-scale networks, the size of a large fraction of communities is usually small [1]. However, for many traditional community detection algorithms, such as Modularity or Ncut, they tend to partition the whole network into relatively equal-size groups with cluster size being no less than (n is the number of nodes in a network) [1], and fail to find small communities due to the problem called “resolution limit” [4]. For attractor, as it simulates the distance dynamics and does not rely on any user-defined criterion, it allows intuitively finding the intrinsic communities with arbitrary size in networks. Therefore, it also provides a promising way to handle anomalies/outliers. In this scenario, anomalies are interpreted as the noisy nodes or unusual nodes isolated from all other nodes over time, and finally pop out automatically.



**Cohesion parameter λ.** For the Attractor algorithm, the cohesion parameter λ is used to determine the positive or negative interaction influence on the distances from exclusive neighbors (see Eq. (6)). Generally, with the higher value of λ, it yields more communities while produces bigger communities with lower value of λ. By modulating the cohesion parameter λ, Attractor allows analyzing the community structure from coarse to fine. Moreover, λ is informative and is easy to tune compared to the algorithms requiring to specify the number of clusters. Fig. 4(a) plots the finding number of communities with different λ ranging from 0 to 1 on a synthetic network. From this plot, we can see that Attractor allows yielding perfect partitioning with the parameter λ on a long stable range (0.2 - 0.7). The clustering results with respect to distinct parameters are further illustrated in Fig. 4(b) to Fig. 4(e). Extensive experiments further demonstrate Attractor is not sensitive to clustering results and usually produces a good result within the range λ = [0.4, 0.6]. Finally, the Pseudo code of Attractor is given in Algorithm 1.

## 3.4吸引子算法(The Attractor Algorithm)

在本节中，我们详细介绍了Attractor算法。

**动态互动。**基于交互模型（参见方程（8）），可以模拟距离动力学，其主要涉及以下步骤：

1. 在初始时间（t = 0），没有任何交互，每个边与初始距离相关联。这里，根据与定义2或定义3的Jaccord距离计算初始值。

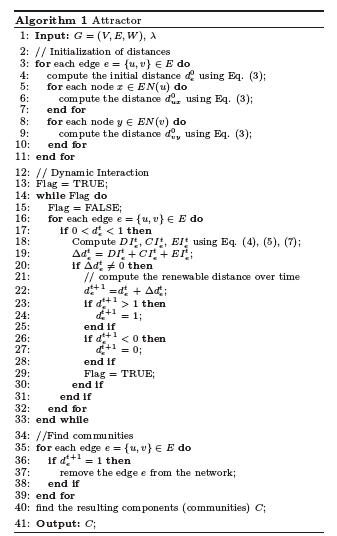
2.随着时间的推移，依赖于局部内在拓扑结构，根据三种提出的相互作用模式（方程（4），式（5）和式（7）），模拟每个距离的动力学。 由于拓扑驱动的影响，共享同一社区的节点之间的距离往往会下降，而不同社区的节点间距离逐渐增加。

3.最后，所有距离都会收敛，并且可以通过去除具有最大距离的边（即d（u，v）= 1）来容易地获得社区。

为了说明，图3（a）-（c）示出了在本地动态交互过程中图1的社交网络从t = 0到t = 9的三种状态。 T = 0表示连接节点之间的初始距离（图3（a））。 从那一刻开始，每个节点与其邻居交互并根据所提出的相互作用模型影响相应的距离（参见方程8），并且在一个时间步长之后的新的距离进一步如图3（b）所示。 经过九个步骤，所有距离都会收敛，0或1，通过切割距离等于1的所有边自然地识别三个社区。

**检测小社区或异常。** 在现实世界的网络中，通常有许多不同规模的社区。 特别是在大型网络中，大部分社区的规模通常很小[1]。 然而，对于诸如模块化或Ncut的许多传统社区检测算法，它们倾向于将整个网络划分成相对大小相等的群组，其中群集大小不小于（n是网络中的节点数量）[1] 并且由于称为“解决限制”的问题，找不到小社区[4]。对于吸引子，由于它模拟距离动力学，并且不依赖于任何用户定义的标准，它允许直观地在网络中查找任意大小的内在社区。 因此，它也提供了处理异常/异常值的有前景的方法。在这种情况下，异常被解释为随着时间的推移与所有其他节点隔离的嘈杂节点或异常节点，最后自动弹出。

**内聚参数λ。**对于吸引力算法，内聚参数λ用于确定对独有邻居的距离的正或负相互作用影响（参见等式（6））。一般来说，随着λ值的增加，它产生更多的社区，而较小的λ值产生较大的社区。Attractor通过调节内聚参数λ，可以从粗到细分析社区结构。此外，与需要指定集群数量的算法相比，λ是信息性的并且易于调整。图4（a）在合成网络上绘制了具有从0到1范围的不同λ的社区的查找数量。从这个图可以看出，吸引力允许在长稳定范围（0.2 - 0.7）下使用参数λ进行完美分割。关于不同参数的聚类结果进一步如图4（b）到图4（e）所示。大量实验进一步证明，吸引力对聚类结果不敏感，通常在λ= [0.4,0.6]范围内产生良好的结果。最后，算法1给出了吸引力的伪代码。



**3.5 Complexity Analysis**

To investigate the distance dynamics, the initial distance of any two linked nodes in a network is required, and thus the time computation is O(|E|). Moreover, for the local dynamic interaction, Attractor also needs to compute the corresponding Jaccard distances for exclusive neighbors (Algorithm 1(Line 5-10)). The time complexity is O(k· |E|), where k is approximately the average number of exclusive neighbors of two linked nodes. During the local dynamic interaction process, as all distances have already existed, Attractor only needs to recall these distances at previous time stamp without any distance computation, and thus the time complexity is O(T· |E|). Totally, the time complexity is O(|E|+ k· |E|+ T· |E|) , where T is the number of time steps. In most cases, T is small with 3≤T≤50.

## 3.5复杂性分析Complexity Analysis

为了研究距离动力学，需要网络中任何两个连接节点的初始距离，因此时间计算是O（| E |）。 此外，对于本地动态交互，Attractor还需要计算独有邻居的相应Jaccard距离（算法1（行5-10））。时间复杂度为O（k·| E |），其中k近似为两个连接节点的独有邻居的平均数。在本地动态交互过程中，由于所有距离都已经存在，所以吸引力只需要在上一个时间戳内调用这些距离，而不需要任何距离计算，因此时间复杂度为O（T·| E |）。 总之，时间复杂度是O（| E | + k·| E | + T·| E |），其中T是时间步长的数量。 在大多数情况下，T小，3≤T≤50。

**4. EXPERIMENTS**

In this section, we evaluate our proposed algorithm Attractor on synthetic as well as real-world networks to demonstrate its benefits.

**Selection of comparison methods.** To evaluate the performance of Attractor, we compare it to several representatives of community detection algorithms.

**Ncut** [16] is a well-known algorithm for graph clustering by optimizing the normalized cut criterion. As the eigenvalue decomposition is applied to speed up finding the optimal cut, it is also usually called as spectral clustering.

**Modularity** [11] is the current most popular community detection algorithm based on the modularity measure, which uses the expected cut to measure clustering quality.

**Metis** [8] is a very fast graph clustering approach for large networks via multi-level partitioning and parallelized implementation.

**MCL** [18] is a popular algorithm used in life sciences based on the simulation of (stochastic) flow in graphs.

**Louvain** [2] is another well-known modularity based algorithm. Compared with the algorithm Modularity proposed by Newman [11], it allows for hierarchical community detection and has lower time complexity.

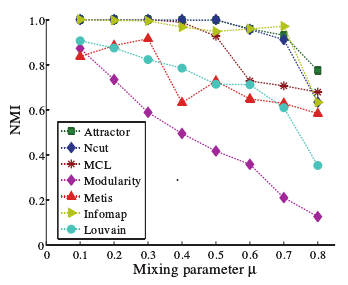
**Infomap** [13] envisions community detection problem as a coding problem, and aims at finding the optimal partitions based on minimum description length principle.

For all experiments, without further statement, Ncut and Metis specify the cluster number K =|C|,|C| is the true number of classes of the network if the ground truth is available. MCL takes the default inflation parameter (i = 2.0) as suggested by authors [18]. For Louvain and Infomap, the default parameters are used. We set the cohesion parameter λ = 0.5 for Attractor as default parameter. All experiments have been performed on a workstation with 3.4 GHz CPU and 32.0 GB RAM.

**Evaluation Matrices.** To extensively compare different community detection algorithms with respect to effectiveness, we evaluate the clustering results in two ways.

•Networks with class label. For networks whose communities are already known, the performance is directly measured by three widely used evaluation measures: Normalized Mutual Information (NMI) [17], Adjusted Rand Index (ARI) [12] and Cluster Purity.

•Networks without class label. Since the ground truth of the community structure is unknown, it is a non-trivial task to compare the performance of distinct algorithms in an objective way. In order to evaluate the quality of communities produced by different algorithms reasonably, two popular internal measures modularity [11] and normalized cut (ncut) [16] have been applied in this study, although they are somehow tailored for modularity -criterion or cut -criterion based algorithms.



**Figure 5: The performance of different algorithms on the LFR benchmark networks by varying the number of inter-cluster edges.**

# 实验(Experiments)

在本节中，我们在合成和现实世界的网络中评估了我们提出的算法吸引力，以显示其优势。

**比较方法的选择。**为了评估Attractor的性能，我们将其与社区检测算法的几个代表进行比较。

**Ncut** [16]是通过优化归一化切割标准的图形聚类的众所周知的算法。由于特征值分解被应用于加速找到最佳切割，通常也被称为光谱聚类。

**模块化**[11]是基于模块化度量的当前最流行的社区检测算法，其使用预期切割来测量聚类质量。

**Metis**[8]是一种通过多级分区和并行实现的大型网络的非常快速的图形聚类方法。

**MCL**[18]是基于图中（随机）流模拟的生命科学中流行的算法。

**Louvain**[2]是另一种着名的基于模块化的算法。 与Newman [11]提出的算法模块化相比，它允许层次化社区检测并具有较低的时间复杂度。

**Infomap**[13]设想社区检测问题作为编码问题，旨在根据最小描述长度原理找到最优分区。

对于所有实验，没有进一步的说明，Ncut和Metis指定簇号K = |C|，|C|是网络的真实数量，如果地面的真相可用。 MCL采用作者建议的默认通货膨胀参数（i = 2.0）[18]。 对于Louvain和Infomap，使用默认参数。我们设置Attractor的内聚参数λ= 0.5作为默认参数。所有实验都在具有3.4 GHz CPU和32.0 GB RAM的工作站上执行。

**评估矩阵** 为了广泛地比较不同社区检测算法的有效性，我们以两种方式评估聚类结果。

* 带有类标签的网络。对于社区已知的网络，性能直接由三个广泛使用的评估措施直接衡量：归一化互信息（NMI）[17]，调整后的兰德指数（ARI）[12]和集群纯度。
* 没有类标签的网络由于社区结构的基本真实性是未知的，因此以客观的方式比较不同算法的性能是一项非常简单的任务。 为了综合评估不同算法产生的社区质量，本研究采用了两种流行的内部措施模块化[11]和归一化切割（ncut）[16]，尽管它们以某种方式为模块化标准而定制， 基于标准的算法。

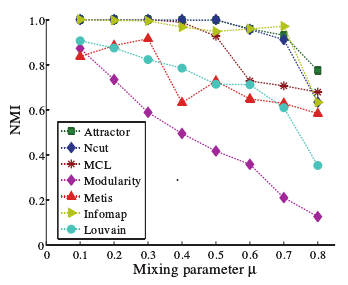


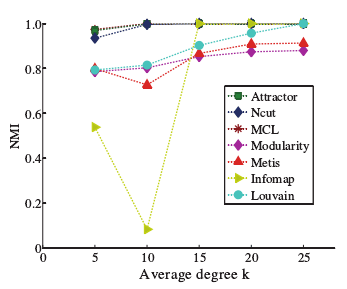
图5：通过改变群集间边数量，在LFR基准网络上执行不同算法的性能。

**4.1 Synthetic networks**

In this section, we first generate several synthetic networks featuring distinct characteristics to compare the performance of various community detection algorithms. For fair comparison and to make the synthetic networks to be more consistent with the real-world networks, the LFR benchmark networks [10] have been applied, where the distributions of degree and community size of networks can be easily controlled. To increase the complexity of networks, the mixing parameter µ [10], defined as the fraction of links of each node outside its community, is used to control the difficulty of community separation.

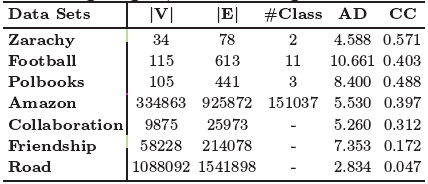
**Noise Edge:** First, we evaluate how well the different graph clustering algorithms allow detecting communities by varying their inter-cluster edges. The inter-cluster edges, which we call noise edges, are added into the network to hamper community separation. We fix node average degree and community size, and change the mixing parameter µ from 0.1 to 0.8 to generate a serial of networks with different inter-cluster edges. All networks consist of 2000 nodes with the average degree k = 20.

With the increase of mixing parameter, the performance (measured by NMI) of all five approaches is shown in Fig. 5. We can see that the algorithms of Attractor, Ncut and MCL almost achieve the perfect clusterings by adding inter-cluster edges with the mixing parameter up to 0.4 (40% edges of each node links to other communities). Their performances begin to decrease with more and more inter-edges added into the network, and Attractor is more robust to these noise edges. For Modularity and Louvain, they are more sensitive to these noise edges, and their performances are not comparable with other five algorithms on these networks. Regarding the Metis algorithm, its performance is fluctuated and starts to decrease dramatically as soon as more inter-edges are added (with µ = 0.4). The performance of infomap is surprising, since its performance decreases first and then increases with µ ranging from 0.5 to 0.7.



**Figure 6: Performance of different algorithms on the LFR benchmark networks by varying community density using the average degree <k>.**

**Table 1: Statistics of real-world data sets, where AD: average degree; CC: clustering coefficient.**



**Community Density:** Next, we evaluate how the algorithms respond to the networks with different average degrees, which we call community density. Here we fix the inter-cluster edges (µ = 0.1), and change the average degree k from 5 to 25 to see the influence of community density on the performance of these algorithms. Fig. 6 shows that Attractor, MCL and Ncut yield good results for all these networks, while the performances of Metis and Modularity are a bit worse. We can see that Attractor, Metis and Ncut allows correctly finding the good communities even with low community density (k = 5). For Louvain, Metis and Modularity, they are more sensitive to the community density on these synthetic networks. As to Infomap, it show an abnormal performance when average degree k is 10.

## 4.1综合网络(Synthetic networks)

在本节中，我们首先生成几个具有不同特征的综合网络，以比较各种社区检测算法的性能。为了公平比较，使合成网络与现实网络更加一致，已经应用了LFR基准网络[10]，其中可以容易地控制网络的度数和社区规模的分布。 为了增加网络的复杂性，混合参数μ[10]被定义为其社区外的每个节点的链路的分数，用于控制社区分离的难度。

**噪声边：**首先，我们评估不同的图形聚类算法如何通过改变它们的跨集群边来检测社区。我们称之为噪声边的群集间边被添加到网络中以阻止社区分离。 我们修复节点平均度和社区大小，并将混合参数μ从0.1改为0.8，以生成具有不同簇间边缘的一系列网络。所有网络由2000个节点组成，平均度k = 20。

随着混合参数的增加，所有五种方法的性能（通过NMI测量）如图1所示。 我们可以看到，Attractor，Ncut和MCL的算法几乎通过将混合参数添加到0.4（每个节点的40％边缘连接到其他社区）来增加簇间边来实现完美的聚类。随着越来越多的边添加到网络中，其性能开始下降，而Attractor对这些噪声边更加活力。对于模块化和Louvain，它们对这些噪声边更敏感，并且它们的性能与这些网络上的其他五种算法不可比。 关于Metis算法，它的性能波动，并且一旦添加更多的边（μ= 0.4）就开始显着降低。 Infomap的表现令人惊讶，因为其性能首先降低，然后随着μ范围从0.5增加到0.7性能提高。

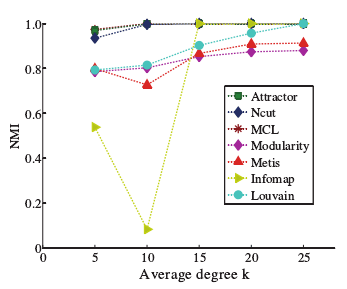
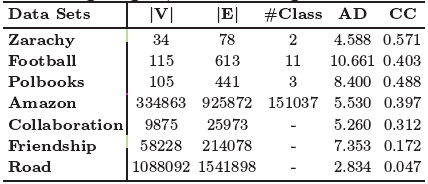


图6：使用平均度<k>改变社区密度对LFR基准网络上的不同算法的性能。

表1：现实世界数据集的统计，其中AD：平均程度; CC：聚类系数

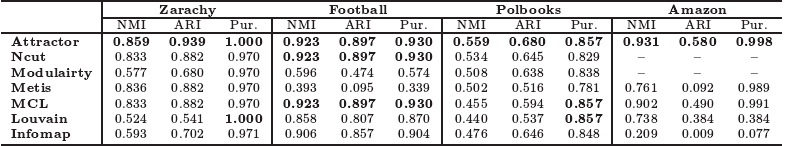


**社区密度：**接下来，我们评估算法如何响应不同平均度的网络，我们称之为社区密度。在这里我们修正了簇间边缘（μ= 0.1），并将平均度k从5改为25，以查看社区密度对这些算法性能的影响。图6显示，吸引力，MCL和Ncut对所有这些网络产生良好的结果，而Metis和Modularity的性能有点差。 我们可以看到，即使社区密度低（k = 5），Attractor，Metis和Ncut也可以正确地找到好社区。 对于Louvain，Metis和Modularity，他们对这些合成网络的社区密度更敏感。至于Infomap，当平均度k为10时，表现出异常表现。

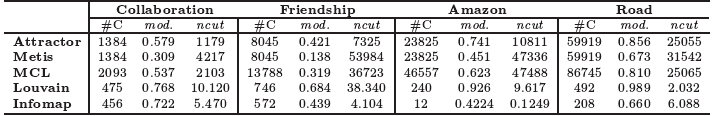
**4.2 Real World Data**

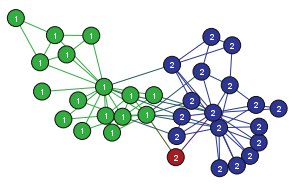
In this section, we evaluate the performances of different community detection algorithms on real-world networks which are all publicly available from the UCI network data repository (https://networkdata.ics.uci.edu/index.php) and Stanford large network dataset collection (http://snap.stanford.edu/data/). The statistics of seven networks are summarized in Table 1.

**Table 2: The performance of different graph clustering algorithms on labeled real-world networks.**

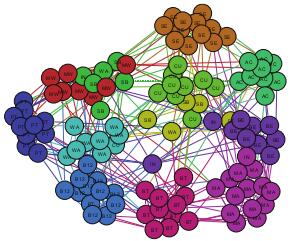


**Table 3: The performance of different algorithms on large real-world networks without class information.**

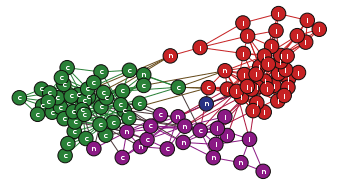




**Figure 7: Attractor on karate club network. Colors of nodes indicate different detected communities.**



**Figure 8: Attractor on American football network.**

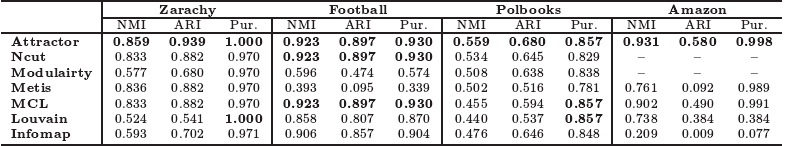


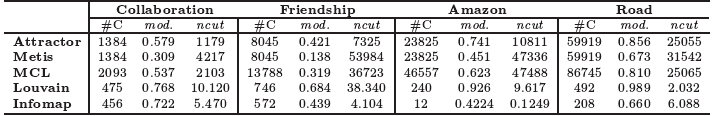
**Figure 9: Attractor on political book network.**

## 4.2现实世界数据(Real World Data)

在本节中，我们评估在UCI网络数据存储库（https://networkdata.ics.uci.edu/index.php）和斯坦福大型网络公司可以公开获得的现实世界网络上的不同社区检测算法的性能数据集合（http://snap.stanford.edu/data/）。七个网络的统计资料总结在表1中。

表2：标记的现实世界网络上不同的图形聚类算法的性能。

 表3：没有类信息的大型现实世界网络上不同算法的性能。



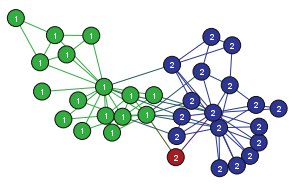


图7：空手道俱乐部网络吸引力。节点颜色表示不同的检测到的社区。

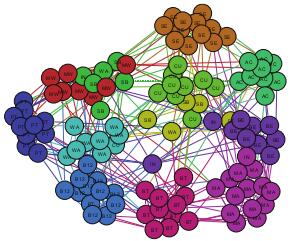


图8：美式足球网络吸引力。

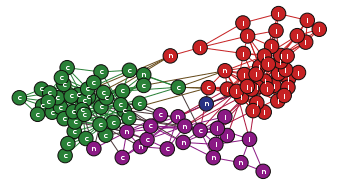


图9：政治书籍网络的吸引力。

**(1). Networks with class information**

We first investigate the networks for which the ground truth of community structure is already known. The external measures such as NMI, ARI and purity are reported.

**Zachary’s karate club network:** The famous network, derived by Zachary’s observation about a karate club, reflects the friend relationship among these members. Specially, the network could be divided into two communities, which reflects the disagreement between the administrator and the instructor. Fig. 7 shows that Attractor identifies the communities with a high degree of success (with high values of NMI, ARI and Purity), and outperforms other comparing algorithms (Table 2). Specifically, two communities are successfully found, except one member is viewed as noise (node‘10’). It is also interesting to observe that this member is located between two communities, and links with the hub nodes of two communities, respectively. In real-world scenario, it is also diﬃcult to determine its community belonging to. Actually, it is more likely to assign this node to both communities, which is the overlapping clustering that we will not discuss in this study. For comparing algorithms of MCL and Ncut, they also achieve a good performance, and most members are correctly grouped. However, for the algorithms of Modularity, Louvain and Informap, many members are wrongly grouped, which result in a relatively low values of NMI. The performance of the different algorithms is summarized in Table 2.

**American college football:** The network derived from the American football games of the schedule of Division I during regular season Fall 2000, where 115 vertices in the graph represent teams, and 613 edges represent regular- season games between the two teams they connect. The teams are divided into 12 conferences containing around 8-12 teams each, and thereby the real community structure is already known. Fig. 8 plots the communities which are detected by Attractor. It is interesting to note that Attractor automatically finds 12 communities with high quality (NMI= 0.923, ARI = 0.897, Purity = 93.0%). From this Figure, we can observe that most of teams are correctly assigned into corresponding communities. Ncut and MCL find the similar community structure as Attractor. For Metis, Modularity, Louvain and Infomap, however, they are difficult to discover the natural community structure (Table 2).

**Books about US politics:** This network, derived from the politic books about US politics published around the time of the 2004 presidential election, consists of 105 nodes and 441 edges. Nodes represent books sold by the online bookseller Amazon.com. Edges represent frequent co-purchasing of books by the same buyers. Each book is labeled with ‘l’, ‘n’, or ‘c’ to indicate whether they are “liberal”,

“neutral”, or “conservative”, based on Newman’s reading of the descriptions and reviews of the books posted on Amazon. Attractor allows a good grouping these books into there categories, where two clusters well represent the corresponding liberal and conservative books, respectively (Fig. 9 and Table 2). For algorithms of Modularity, Metis and Ncut, they yield comparable results, while MCL and Infomap produce a relatively bad grouping on this network.

**Amazon network:** This network consists of 334,863 nodes and 925,872 edges, and each node represents a product on the Amazon website. Each product is categorized to corresponding community based on its category provided by Amazon, and the top 5,000 communities with highest quality were investigated in [20]. Due to the high time and space complexity of eigenvalue decomposition in Ncut and Modularity, they cannot handle this network. Relying on the partial ground-truth communities (top 5,000 communities), Attractor obtains the best community quality comparing to other algorithms with high measures (NMI = 0.931, ARI = 0.580, Purity = 0.998) (Table 2). MCL algorithm allows producing a reasonable result (NMI = 0.902). However, for Metis, Louvain and Infomap, they tend to fail, especially for the algorithm Informap (NMI = 0.209). Moreover, for comprehensive evaluation, all results of the five algorithms are also evaluated by the internal criteria of modularity and ncut (see Table 3).

### 具有类信息的网络(. Networks with class information)

我们首先调查社区结构的基本真相已经知道的网络。诸如NMI、ARI和纯度等外部指标被报道。

**扎卡里的空手道俱乐部网络：**由扎克里对扎伊尔空手道俱乐部的观察而得名的网络反映了这些会员之间的友谊关系。特别地，网络可以分为两个社区，这反映了管理员和教练之间的不一致。图7显示吸引力识别出高度成功的社区（具有较高的NMI，ARI和Purity），并且优于其他比较算法（表2）。具体来说，成功找到了两个社区，除了一个成员被视为噪声（node'10'）。还有一点很有意思的是，这个成员位于两个社区之间，分别与两个社区的中心节点连接。在现实情况下，确定其社区属性也是不利的。实际上，更有可能将这个节点分配给两个社区，这是我们在本研究中不会讨论的重叠聚类。为了比较MCL和Ncut的算法，它们也实现了良好的性能，大多数成员正确分组。然而，对于模块化，Louvain和Informap的算法，许多成员被错误地分组，这导致NMI的值相对较低。表2总结了不同算法的性能。

**美国大学足球：**来自美国足球赛的网络是在常规赛2000赛季的第一赛季，其中图中的115个顶点代表队，613个边代表了他们连接的两队之间的常规赛。 团队分为12个会，每个会约8-12个团队，因此，真正的社区结构已经是已知的。 图8绘制Attractor检测到的社区。有趣的是，Attractor自动发现12个社区的质量（NMI = 0.923，ARI = 0.897，Purity = 93.0％）。从该图可以看出，大多数团队都被正确分配到相应的社区。 Ncut和MCL找到与Attractor类似的社区结构。 然而，对于Metis，Modularity，Louvain和Infomap来说，他们很难发现自然社区结构（表2）。

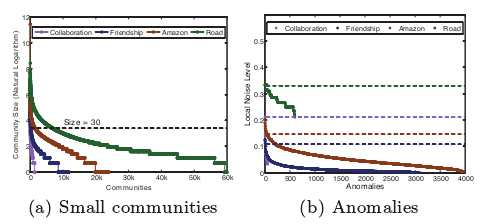
**关于美国政治的书籍：**在2004年总统选举之前出版的关于美国政治的政治书籍的这个网络由105个节点和441个边组成。节点代表网络书商Amazon.com销售的图书。边代表同一买家频繁地共同购书。根据纽曼阅读亚马逊书籍的描述和评论，每本书都标有“l”，“n”或“c”，以表明他们是“自由主义”，“中立”还是“保守”。吸引力允许将这些书分类到这些类别，其中两个群体分别很好地代表相应的自由和保守的书籍（图9和表2）。对于Modularity，Metis和Ncut的算法，它们产生可比较的结果，而MCL和Infomap在该网络上产生相对较差的分组。

**亚马逊网络：**该网络由334,863个节点和925,872个边组成，每个节点代表亚马逊网站上的一个产品。 每个产品根据亚马逊提供的类别分类到相应的社区，在[20]中调查了最高质量的前5000名社区。 由于Ncut和Modularity中特征值分解的时间和空间复杂度高，因此无法处理该网络。依靠部分地面真相社区（前5 000个社区），与其他具有较高措施的算法（NMI = 0.931，ARI = 0.580，Purity = 0.998）相比，Attractor获得了最佳的社区质量（表2）。 MCL算法可以产生合理的结果（NMI = 0.902）。 然而，对于Metis，Louvain和Infomap，它们往往会失败，特别是对于算法Informap（NMI = 0.209）。此外，为综合评估，五种算法的所有结果也通过模块化和ncut的内部标准进行评估（见表3）。

**(2). Networks without class information**

In this section, due to the time complexity of Ncut and Modularity, we limit the comparison to the clustering algorithms Metis, MCL, Louvain and Infomap on large-scale networks without class ground truth (Table 3). As there exist no convincing measures for the unlabeled network, we use Modularity and Ncut to evaluate these algorithms in an informative way.

**Hepth collaboration network:** The network is a collaboration network of 9,875 authors working on the theory of high energy physics. Attractor identifies 1384 communities, which results in modularity = 0.579 and ncut = 1179. On the data set, Metis (K = 1384) and MCL also yield a good partitioning while its performance is worse than Attractor in terms of the two measures (Table 3). If we just look at modularity and ncut, Louvain and Infomap seem much better than Attractor, MCL and Metis. However, the reason is that the two algorithms only yield few communities, naturally lead to better values of modularity and ncut, based on the definitions. If we regress out the effect of different number of communities, Attractor actually obtains better performance.

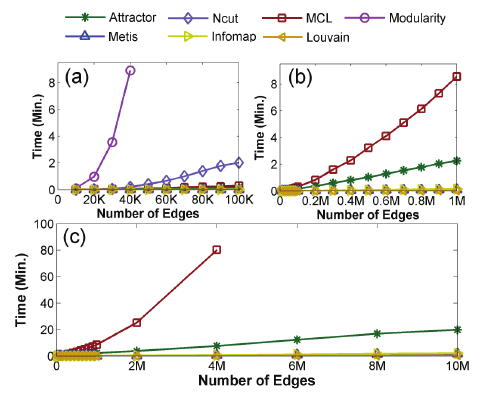


**Figure 10: Evaluation of small communities and anomalies.**

**Brightkite friendship network:** The graph is a location-based friendship network consisting of 58,228 nodes and 214,078 undirected edges. Attractor finds 8045 communities and shows a clear advantage over two other algorithms based on the two measures. For Metis (K=8045), many friends seem to be incorrectly grouped, which result in a low value of modularity = 0.138. As usual, Louvain and Infomap, the two algorithms tend to produce small number of communities, and many small-sized communities cannot be detected. The reason is might due to the “resolution limit”.

**Pennsylvania road network:** This network reflects road structure of Pennsylvania, where nodes represent intersections or the endpoints and edges represent the roads connecting these intersections or endpoints. Here, we set the parameter λ = 0.6 for Attractor and i = 1.4 for MCL as the default values of the two algorithms cannot result in a good results due to the very sparsity of the network. Attractor finally identifies 59,919 clusters with modularity = 0.856 and ncut = 25055. MCL achieves the comparable performance and is better than the algorithm Metis. Louvain and Infomap only find small number of equal-size communities (492 and 208, respectively).

In total, the experiments on all real-world networks demonstrate that Attractor not only allows extracting meaningful communities in networks with class label (with highest performance in terms of all measures), but also scales up large-scale networks and yields a good graph partitioning in terms of the internal (modularity and ncut ) and external measures (NMI, ARI and Purity) (see Table 2, Table 3).



**Figure 11: The runtime of the different algorithms.**

### (2)无类信息的网络(. Networks without class information)

在本节中，由于Ncut和Modularity的时间复杂度，我们限制了与没有类地面实况的大规模网络上的聚类算法Metis，MCL，Louvain和Infomap的比较（表3）。由于对未标记的网络没有说服力的措施，我们使用Modularity和Ncut以信息方式评估这些算法。

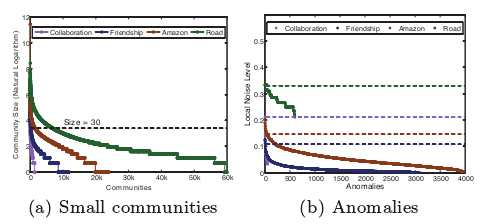
**Hepth协作网络：**该网络是一个9,875名作者，致力于高能物理学理论的协作网络。吸引力识别1384个社区，其结果是模块化= 0.579，ncut = 1179.在数据集上，Metis（K = 1384）和MCL也产生良好的分区，而在两个度量方面，性能差于Attractor（表3） ）。 如果我们只是看模块化和ncut，Louvain和Infomap似乎比Attractor，MCL和Metis好得多。然而，原因在于，这两种算法只能产生很少的社区，根据这些定义自然会导致更好的模块化和ncut值。 如果我们回归不同社区的影响，吸引者实际上会获得更好的表现。

图10：评估小社区和异常情况。

**Brightkite友情网络：**该图是一个基于位置的友谊网络，包括58,228个节点和214,078个无向边。 吸引力发现8045个社区，并且基于两个措施显示出优于另外两个算法的明显优势。对于Metis（K = 8045），许多朋友似乎被错误地分组，这导致模块化的低值= 0.138。像Louvain和Infomap一样，这两种算法往往会产生少量的社区，许多小型社区也无法被发现。 原因可能是由于“分辨率限制”。

**宾夕法尼亚道路网络：**该网络反映了宾夕法尼亚州的路面结构，其中节点代表交叉路口或端点，边代表连接这些交叉点或端点的道路。 这里，我们为Attractor设置参数λ= 0.6，对于MCL设置i = 1.4，因为两个算法的默认值由于网络的稀疏性而不能导致良好的结果。 吸引力最终识别59,919个群集，其中moduleity = 0.856，ncut = 25055。MCL实现了可比性能，比Metis算法更好。 Louvain和Infomap只发现少数同等社区（分别为492和208）。

总的来说，所有现实世界网络的实验表明，吸引力不仅允许在具有类标签的网络中提取有意义的社区（在所有措施方面具有最高性能），而且可以扩展大型网络并产生良好的图形分区 在内部（模块化和ncut）和外部措施（NMI，ARI和纯度）方面（见表2，表3）。

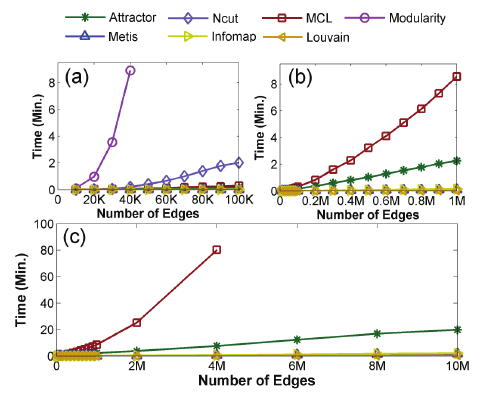


图11：不同算法的运行时间。

**4.3 Small Community and Anomaly Detection**

In this section, we evaluate whether Attractor allows identifying small meaningful communities and anomalies. Fig.10(a) plots the distribution of community size for the four large real-world networks, and we can see that Attractor can find many small communities with different size. To demonstrate the potential high-quality of the communities, we further examine the quality of the resulting small communities (size≤30) on Amazon network as it has the ground truth for the top 5,000 communities. It is interesting to note that the1458 small communities (size≤30) result in high values of NMI = 0.941, ARI = 0.637 and Purity = 0.989, which shows the desirable property of small community detection.

Moreover, to check whether the detected anomalies are the potential noisy/unusual nodes, we evaluate the local noise level of each node, which is defined as the fraction of node degree over the number of all links of its neighbors. Fig. 10(b) depicts the local noise level for all resulting anomalies compared to the average noise level (indicated by dashed lines) by Attractor on the four real-world networks, providing a potential evidence for the effective anomaly detection.

**4.4 Runtime**

To assess the scalability of Attractor with respect to network size, we generate several benchmark networks [10] with different edge sizes ranging from ten thousand to ten million by fixing the average node degree k = 20. Fig. 11 shows the running time for different graph clustering algorithms. We can observe that Attractor is faster than Modularity, Ncut and MCL since its time complexity is linear against to |E|. However, Attractor is a bit slower than scalable community detection algorithms Metis, Louvain and Infomap. Although the three algorithms are much faster than Attractor, they suffer in the quality of resulting communities.

## 4.3小社区和异常检测( Small Community and Anomaly Detection)

在本节中，我们评估吸引力是否允许识别有意义的小社区和异常。 图10（a）绘制了四个大型现实世界网络社区规模的分布情况，我们可以看到，吸引者可以找到许多不同规模的小社区。 为了展示社区的潜在高品质，我们进一步检查亚马逊网络上由此产生的小社区（大小≤30）的质量，因为它拥有前五千个社区的实质。有趣的是，1458个小社区（大小≤30）导致NMI的高值= 0.941，ARI = 0.637，Purity = 0.989，这表明小社区检测的理想性质。

此外，为了检查检测到的异常是否是潜在的噪声/异常节点，我们评估每个节点的局部噪声水平，其被定义为节点度数除以其邻居的所有链路的数量的分数。图10（b）描绘了吸引力对四个真实世界网络的平均噪声水平（由虚线表示）的所有结果异常的局部噪声电平，为有效的异常检测提供了潜在的证据。

## 4.4 运行(Runtime)

为了评估Attractor在网络尺寸方面的可扩展性，我们通过固定平均节点度k = 20，生成不同边大小范围从一万到一千万的几个基准网络[10]。 图11显示了不同图形聚类算法的运行时间。 我们可以观察到，吸引力比模块化，Ncut和MCL更快，因为它的时间复杂度是线性的| E |。 然而，Attractor比可扩展的社区检测算法Metis，Louvain和Infomap有点慢。 虽然这三种算法比吸引力要快得多，但它们受到社区质量的影响。

**5. CONCLUSIONS**

In this paper, we introduce a new community detection algorithm, called Attractor, to automatically uncover community structure in networks based on distance dynamics. Extensive experiments further demonstrate that Attractor allows finding communities in small to large size networks with high quality, and also shows attractive benefits compared to several state-of-the-art methods. In future work, we plan to focus on exploring large network abstraction and visualization based on the intuitive dynamic interaction model.

**6. ACKNOWLEDGMENTS**

The research was supported partially by the National Natural Science Foundation of China (Grant No. 61403062, 61433014), Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J053, ZYGX2014J091) and the Postdoctoral Science Foundation of China (Grant No. 2014M552344).

# 5.结论(Conclusions)

在本文中，我们引入了一种称为吸引子的新型社区检测算法，可以根据距离动态自动发现网络中的社区结构。广泛的实验进一步证明，吸引力允许在高质量的小型到大型网络中查找社区，并且与几种最先进的方法相比，也显示出有吸引力的好处。在未来的工作中，我们计划着眼于直观的动态交互模型探索大型网络抽象和可视化。

# 6.致谢(Acknowledges)

研究部分由中国国家自然科学基金（拨款号：61403062,61433014），中央大学基础研究基金（批准号：ZYGX2014J053，ZYGX2014J091）和中国博士后科学基金（批准号：2014M552344）。

# 7.参考文献(REFERENCES)

[1] Nir Ailon, Yudong Chen, and Huan Xu. Breaking the small cluster barrier of graph clustering. In ICML, pages 995–1003, 2013.

[2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008, 2008.

[3] Santo Fortunato. Community detection in graphs. Physics Reports, 486(3):75–174, 2010.

[4] Santo Fortunato and Marc Barth´elemy. Resolution limit in community detection. PNAS, 104(1):36–41, 2007.

[5] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):7821–7826, 2002.

[6] Christian Hennig and Bernhard Hausdorf. Design of dissimilarity measures: A new dissimilarity between species distribution areas. In Data Science an Classification, pages 29–37. Springer, 2006.

[7] Paul James, Yaso Nadarajah, Karen Haive, and Victoria Stead. Sustainable communities, sustainable development: Other paths for papua new guinea. Hawaiian Journal of History, 48, 2014.

[8] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on scientific Computing, 20(1):359–392, 1998.

[9] George Karypis and Vipin Kumar. Multilevelk-wa partitioning scheme for irregular graphs. Journal of Parallel and Distributed computing, 48(1):96–129, 1998.

[10] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. Physical review E, 78(4):046110, 2008.

[11] Mark EJ Newman. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23):8577–8582, 2006.

[12] William M Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336):846–850, 1971.

[13] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. PNAS, 105(4):1118–1123, 2008.

[14] Satu Elisa Schaeﬀer. Graph clustering. Computer Science Review, 1(1):27–64, 2007.

[15] Junming Shao, Xiao He, Qinli Yang, Claudia Plant, and Christian Boehm. Robust synchronization-based graph clustering. In PAKDD, pages 249–260, 2013.

[16] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. IEEE TPAMI, 22(8):888–905, 2000.

[17] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research, 3:583–617, 2003.

[18] Stijn Van Dongen. A cluster algorithm for graphs. Report-Information systems, (10):1–40, 2000.

[19] Zhenyu Wu and Richard Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. IEEE TPAMI, 15(11):1101–1113, 1993.

[20] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In IEEE ICDM, pages 745–754, 2012.