面向复杂网络可视化数据挖掘与分析的拓扑过滤

Faraz Zaidi CNRS UMR 5800 LaBRI & INRIA
Bordeaux - Sud Ouest 351, cours de la Lib´eration
33405 Talence cedex, FRANCE 电子邮件:
faraz.zaidi@labri.fr

arXiv:2304.01896v1 [cs.SI] 4 Apr 2023

摘要——**许多现实世界网络中小世界和无标度特性的发现，彻底改变了我们研究、分析、建模和处理网络的方式。分析这些复杂网络的一种重要方法，是使用图布局算法对其进行可视化。由于规模庞大且连通结构复杂，很难从这些网络的可视化表示中得出结论。本文提出了一种基于拓扑过滤、网络度量指标和可视化的大型图交互式分析方法。我们分析了若干现实世界网络，并使用所提出的方法得出了有趣的结论。**

I. 引言

近年来，计算机存储、处理和可视化大型网络的能力不断提升 [1]。术语网络对不同领域的人有不同含义，常用于表示社会网络 [2], [3], [4]、在线社交媒体 [5]、网页图 [6]、网络 [3], [4]、航空运输 [7], 、代谢通路 [8]和万维网 [9]等系统。研究表明，尽管这些领域看似多种多样，但它们在方法论基础上有很强的共通性，并且采用相似的方法来分析、建模、理解和组织这些网络。研究人员提出了不同的度量指标来对这些网络进行分类 [10]。当网络呈现小世界 [11] 和无标度 [12] 特征时，两种分类方式尤其受到关注。

纵观现有网络分析方法的分类，一种途径是开发度量指标和测度方法，以获取有关这些网络的定量信息。这些测度可分为元素级、群组级和网络级 [10]。节点的度是元素级测度的一个典型例子，它指节点与网络中其他节点相连的数量。另一方面，网络级度量指标则考察整个网络，以计算定量测度。中介中心性 [13]是网络级度量指标的一个典型例子。

另一种方法是利用可视化来理解这些网络。这些方法可分为三类：数据的可视化探索、数据挖掘结果的可视化和数据挖掘与可视化的紧密结合 [14], [15]。可视化数据探索是通过交互式探索数据的图形表示来发现知识。数据挖掘结果的可视化，是以图形方式呈现数据挖掘过程的结果，使人们能够轻松解读所得结果。将这两种方法结合起来的一种混合形式，是把可视化集成到数据挖掘算法中，并由用户或领域专家通过交互方式引导数据挖掘算法。可视化探索是一种发现隐藏知识并提取数据中有趣模式的有效方法 [16]。它已有效应用于多个不同领域，这些领域中的图具有小世界和无标度特性 [17]。

如今，大多数现实世界网络都包含数十万个节点和边。计算开销高昂的度量指标已不再适用于现实世界网络。即使计算能力不断提升，我们仍需要更快速的方法、启发式方法和分解方法，以便在合理的时间内处理这些网络。此外，对于这类大型网络，中心性和网络结构组织等测度不再主要由单个元素决定，而是更多地取决于群组和子集。这些子集之间复杂的相互作用对网络的整体行为和演化发挥着重要作用，也因此给研究人员带来了一个颇具挑战性的问题。

解决这一问题的一个有趣方法是将网络分解为子组件或子网络，然后对这些子网络应用度量或可视化方法，从而深入了解整个网络。从度量的角度来看，分解只需考虑节点的子集，而非整个网络，因此可以加快计算速度。从可视化的角度来看，与整个网络相比，节点较少的网络更易于进行视觉分析。分解技术对可视化方法至关重要，因为即使采用新型先进的可视化工具和交互技术，在计算机屏幕上可视化这些大型网络仍然几乎不可能 [18]。

本研究的贡献不在于我们对不同网络得出的结果或发现。论文中列出的所有发现都已广为人知（并且均有引用），但它们是通过若干不同的方法发现的，而这些方法往往需要复杂的数学和统计公式及推导。本研究的贡献在于所提出的方法简单易用，使来自不同领域的领域专家无需具备任何特殊技能也能从中受益。正如案例研究所示，该方法具有通用性，可应用于多种网络，包括社会网络、生物网络和技术网络。此外，以这种方式结合网络度量指标、过滤和可视化方法来发现这些结论，此前尚无先例。我们的目标是采用另一种方法重新发现已知事实和结论；这种方法适合数学和计算机科学背景非专业的领域专家，用以理

解多种现实世界网络复杂的相互关联行为。

本文介绍一种可视化数据探索方法，该方法通过基于拓扑的节点筛选对大型图进行分解，利用网络度量指标进行定量分析，并通过可视化进行宏观层面分析和交互式探索。这种拓扑分解源于真实网络的两个重要特征：第一，这些网络具有幂律度分布；第二，对这些网络进行可视化会生成高度纠缠、难以阅读的图形 [19]。我们的思路是依据度连接性创建子网络，然后将这些子网络可视化，以便理解、分析并从中提取信息。

本文其余部分安排如下：第三节介绍本文实验中使用的若干网络。第三节介绍所提出的方法，并说明如何使用该方法分析各种数据集。最后，第五节给出结论和未来的研究方向。

II. 相关工作

分解并研究复杂网络的思想并非新鲜事。Batagelj 和 Zaversnik 提出了一种基于顶点连通性的分解方法，称为 k-核分解 [20]。该方法包括识别网络中称为 k-核的子集。这些子集通过递归移除所有度小于 k 的顶点来获得，直到所有剩余顶点的度均大于或等于 k。例如，要获得网络的一个 $2-core$，我们会移除网络中所有度小于 2 的顶点，也就是度为 1 的节点。移除后，一些先前度较高的顶点现在可能变成度为 1；重复这一移除过程，直到网络中不再有度为 1 的顶点。移除后剩余的所有顶点都成为 $2-core$ 的一部分，然后可以针对更高的 k 值重新开始这一过程。

k 值较大的核对应于由高度节点构成、且仅与高度节点相连的顶点集合。因此，k 值较大的核在网络结构中处于更中心的位置 [21]。这种方法已用于多个领域，以分析网络及顶点的连通性，例如蛋白质相互作用网络的分析 [22]。除用于分析外，该方法还用于可视化大规模网络，因为它会将网络分解为中心性逐渐增高的顶点子集。它还可以帮助聚焦网络中的某些感兴趣区域 [21]。

我们提出的方法与k-核有显著区别，尽管 k-核和所提出的拓扑分解都基于顶点的度，并会创建子集。拓扑分解侧重于研究边在高度和低度节点中的分布情况。 k-核则侧重于递归识别中心节点，其目标显然不同。随着我们进一步说明该方法的细节，这些差异将更加明显。

网络还有许多其他分解方法。最广泛使用的方法是聚类 [23]，，它将相似节点归为一组。聚类常用于分解大型网络，以便更好地理解和进行可视化挖掘 [16]。聚类侧重于将相似节点归为一组，而我们的目标是分解网络，并简化节点复杂的连通性行为。节点的拓扑相似性可用作相似度指标，将节点归为一组，因此，所提出的方法可视为一种基于拓扑的聚类方法。为简便起见，我们将该方法称为分解方法，而非聚类方法。

其他用于网络探索和挖掘的方法包括 Abello 等人的工作 [24]。他们提出了高效的网络可视化技术，使用户能够通过交互式导航，发现有关参与者及其关系的有用信息。他们采用从概览到局部特写的多个抽象层级，并基于图切片的概念进行导航。Henry 等人[19] 提出了 NodeTrix，它是一种节点—链接图与邻接矩阵的混合表示，旨在更好地支持社区结构分析。该表示结合多种交互技术，有助于用户创建有用的可视化结果。焦点 +上下文技术 [25]是减少节点杂乱、进行可视化分析与挖掘的另一种方法。这些技术使用户能够同时查看概览信息并聚焦于感兴趣的群组。Kreuseler 和 Schumann[26] 采用了一种将分析方法与可视化方法相结合的交互式数据探索方案。他们提出了一个可视化数据挖掘框架，旨在帮助人们更好地理解信息空间。

III. 数据集

为进行实验，我们选取了三个数据集。由于所提出的方法仅适用于简单图，因此在需要时，我们将这些图转换为简单、无向且无权的图。对于原本不连通的图，我们只考虑每个数据集中最大的连通分量。**Geometry 网络** 是一个作者合作网络；如果两位作者曾共同撰写过一篇文章，则两者之间由一条边相连。该网络包含 3621 个节点和 9461 条边（http://vlado.fmf.uni-lj.si/pub/networks/data/. **Dblp2008 网络**是另一个从 DBLP 数据库构建的合著网络，数据库地址为 http://www.informatik.uni-trier.de/~ley/db/，数据对应 2008年，包含 93498 个节点和 260152 条边。**AirTransport 网络** 是一个城市间的航空交通网络 [7] ，包含 1540 个节点和 16523 条边。**Opte 网络** 是一个互联网层析成像网络，由从一台测试主机到互联网其他网络的路由路径组成。该数据库可从 Opte Project 网站（http://opte.org/）公开获取。该网络包含 35836 个节点和 42387 条边。**Protein 网络** 是一个蛋白质—蛋白质相互作用网络 [27]。数据可从 http://dip.doe-mbi.ucla.edu/dip 获取，包含 1246 个节点和 3142 条边。

IV. 拓扑分解

本节介绍度诱导子图 （DIS）的概念。这种子图通过对节点度施加约束而创建。我们定义了两种此类图：

定义 1：最大$_d$-DIS 是 G 的一个诱导子图，其顶点集为 V'，使得 V' 中的节点在

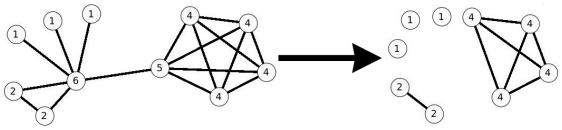

图 1。计算 Max_d−DIS 前后的示例。

G。对于图 $G(V,\ E)$，其中有 V=个节点和E=条边，最大$_d$-DIS 定义为诱导子图$G'(V',\ E')$，使得 $V' \subseteq V$且 $\forall u \in V',\ deg_G(u) \leq d$，其中 d 的取值范围为 0 到所考虑网络可能的最大节点度。

定义 2：Min$_d$-DIS 是 G 的一个诱导子图，其顶点集为 V'，其中的节点在G中的度为最小值 d。对于一个图 $G(V,\ E)$，其中包含 V=个节点和E=条边，$Min_d - DIS$被定义为诱导子图 $G'(V',\ E')$，满足 $V' \subseteq V$和 $\forall u \in V',\ deg_G(u)$。其中，$d$ 的取值范围为 0 到所研究网络可能具有的最大节点度。

A. Max$_d$-DIS：深入观察

考虑图 1（左）所示的图，并计算图中的 Max$_4$-DIS（右）。节点标注的是其在整个图中的度。计算 Max$_4$-DIS 时，度为 5 和 6 的节点被移除，并从剩余节点中得到一个诱导子图。在此示例中，很明显，在 Max$_4$-DIS 中，节点分裂成了非连通分量。因此可以简单推断，整个网络是由高度节点连接起来的；一旦移除这些节点，网络就分裂成了更小的连通分量。另一个有趣的观察是，我们可以研究边如何分布在低度节点之间，因为根据定义，对于参数 d取较小值的情况，Max$_d$-DIS 只考虑低度节点。研究无标度网络时，一个重要关注点是枢纽节点 [12] 的作用，其中度非常高的节点在网络中十分突出。理解 Max$_d$-DIS 的一种方式是：我们希望避开枢纽节点，并观察在没有它们的情况下节点如何彼此连接。

让我们通过绘制多个 Max$_d$-DIS 图，来看看现实世界图的另一个示例，即 Geometry 网络。这些图使用力导向算法 FMMM[28] 绘制。该算法会在布局中将彼此连接紧密的节点放置得更近，并将彼此不相连的节点推至较远的位置。这些算法非常适合用于直观检测网络中的社区结构。但对于具有无标度特性的网络而言，由于大量节点只连接到少数几个节点，因此很难通过视觉识别这些社区的存在。这是因为连通性非常高的节点会被放置在布局中心，从而在中心形成难以阅读且相互缠绕的图形。度较低的节点则被放置在外围，并且彼此之间的连接较为稀疏。在图 2(a) 中，我们可以看到使用 FMMM 算法绘制的整个网络。该网络包含超过 3500 个节点，因此很难从中看出任何有趣的内容。

图2(b)显示了 Max$_5$-DIS 子图，其中包含2757个节点，占整个图中节点总数的76%以上。这表明大多数节点位于该网络的度较低。大多数时候，我们关注的是在复杂网络分析中脱颖而出的高度节点。但大多数节点的节点度较低，我们强调，分析方法也应关注这些节点，因为它们占多数，并且在很大程度上影响网络的整体行为。

该子图中有 1481 条边，因此平均节点度为 0.53，而整体值为 2.6；在当前语境下，两者存在巨大差异。整个网络的最大节点度为 102。这意味着，这些低度节点中的大多数倾向于连接到度“更高”的节点。请注意，我们强调的是相对度，并使用“更高”而不是“最高”，这是为了说明：由于我们分析的是度上限为 5 的 Max$_5$-DIS，这些节点最终可能会连接到节点度为 6,7, 8 的节点，而不一定连接到节点度接近最高值（102）的节点。我们将在分析 Max$_{10}$-DIS 和 Max$_{15}$-DIS 图时对此进行考察。

另一个有趣的观察是，连通分量的数量为 1537，并且在 Max$_5$-DIS 中存在大量度为 0 的节点。请记住，整个网络是一个单独的连通分量；随着高度节点的出现，这些较小的连通分量会合并，形成一个大的单一连通分量。最后，一个重要的观察是，这些低度节点彼此连接时所形成的结构。我们提到 Georg Simmel [29] 的研究，他提出了三元组这一社会网络的基本结构概念。此后，人们在许多现实世界网络中都观察到了三元组的形成。

图2(b) 证实了三元组的存在，因为它们清晰可见。三元组不仅存在于包含三个以上节点的连通分量中，也存在于恰好包含三个节点的连通分量中。此外，我们还观察到了大小为 4 和 5 的团。根据该网络的构造方式，我们预期其中会存在团。由于该数据集涉及文章，而一篇文章通常会有 3 位、4 位或更多作者，因此，在将网络可视化时，我们显然会发现这些团。为避免产生任何错误的推论，我们想再次提及图 1：请注意其中包含 5 个节点的团；当构建 Max$_d$-DIS 时，它将不再保持相同大小的团。因此，在分析图2(b) 这样的子图时，我们不应得出网络中只有大小为 3 或 4 的团这一结论；实际上，整个网络中肯定还存在更大规模的团。

图2(c)显示了同一网络的 Max$_{10}$-DIS。从图中可以立即看出，图2(b)中的许多小型连通分量现在已经连接起来了。其中两个连通分量非常明显，此外还有一个规模相当大的连通分量。在数学中，这种现象称为巨型分量的出现 [30]。随着某些连边的引入，网络发生了巨大变化，并变成一个单一的连通分量。在这种情况下，这些连边是由度较高的节点引入网络的，它们负责连接所有这些较小的分量。Max$_{10}$-DIS 中的连通分量数量为 942，而 Max$_5$-DIS 中为 1537；这是一个显著的减少，因为我们没有度非常高的节点。

请注意，我们仍然没有出现度非常高的节点。观察图 4 所示的 Geometry 网络的度分布，我们可以看到，长尾从度约为

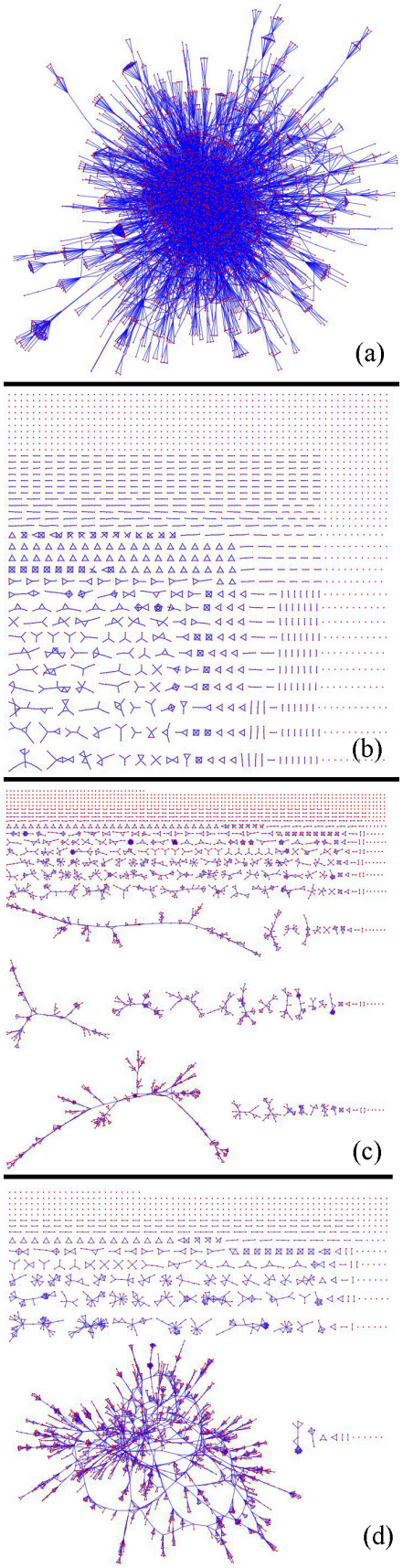

图2. Geometry 网络的 Max_d-DIS 图可视化。(a) 整个网络 (b) Max_5-DIS (c) Max_{10}-DIS (d) Max_{15}-DIS。

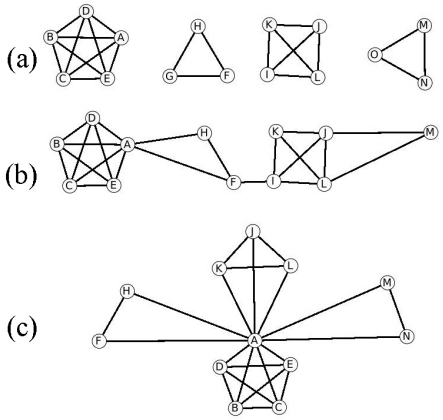

图3. (a) 合著网络中代表不同文章的四个团 (b) 这些团组合在一起，形成平均路径长度较大的网络，其中某些节点的度较高 (c) 这些团组合在一起，形成平均路径长度较小的网络，其中节点 A 是一个非常高度的节点。

10 时开始形成。从聚类的角度来看，这一点相当重要。假设我们想要将相似的节点分组，在没有高度节点的情况下，这个网络会分裂成更小的连通分量；通过观察子图，我们可以看出，整个网络中有两个较大的节点簇和一个较小的节点簇。随着高度节点逐渐加入网络，网络会变成一个大的连通分量，也就难以识别各个簇。关于度分布，有一个观察结果是：节点出现频率最高的度数范围内，度小于 4 的节点占总节点数的 60%。

最后，我们来看图 2(d)，它展示了 Max_{15}-DIS。该图的一个显著变化是形成了一个包含 2133 个节点和 3523 条边的大型连通分量。请注意，与整个网络中最高节点度为 102 相比，该图仅包含最大节点度为 15 的节点。这表明，约 59% 的节点连接在同一个连通分量中，且其中没有度非常高的节点。

一个显而易见的问题是，我们如何定义“高度”或“非常高”的度？虽然没有具体定义，但可以使用启发式方法进行较为粗略的估算。如果将 15 与 102 进行比较，两者差距很大，很容易看出数值 15 并不能真正代表“高”的度值。另一方面，如果将 15 与平均度 2.6 进行比较，这个数值就相当高，可以视为较高的节点度。从百分比来看，Max_{15}-DIS 包含 3413 个节点，占节点总数的 94%。如果将度最高的 5% 节点视为高度节点，那么 Max_{15}-DIS 表明，即使网络中没有高度节点，仍有 59% 的节点保持连通。因此，我们认为 Max_d-DIS 可用于研究 Albert等人 [31]所研究的无标度网络的一项重要特征，即无标度网络能够抵御节点随机丢失，但容易受到针对枢纽节点的针对性最坏情况攻击。

观察该子图中的平均度，整个网络中 94% 的节点平均度为 1.33，而整个网络的平均值为 2.6。因此，高度节点对节点平均度的影响很大。

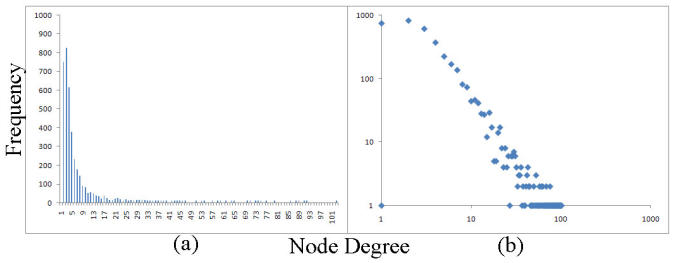

图4。（a）Geometry 网络度分布的直方图和（b）双对数散点图。

整个网络。显然，这是度分布中长尾的直接结果。尾部越长，低度节点的平均节点度与整个网络的平均节点度之间的差异就越大。

关于图 2(c) 中节点的连通性以及三个较大的连通分量，可以观察到一个有关其平均路径长度的有趣现象。这三个分量的数值分别为 12.9、12.4 和 9.7，与总体数值 5.31 相比相当高。即使是图 2(d) 中最大的连通分量，其平均路径长度也高达 12.1。随着高度节点出现在网络中，网络的平均路径长度显著下降。仔细观察图 2(a、b、c)，我们可以理解这些网络的一个重要连通性原理。当节点通过“较高”度的节点彼此连接时，由于连接形成了较长的路径，平均路径长度会稍高；而当节点通过“非常高”度的节点连接时，平均路径长度则会降低。

我们进一步通过一个合著网络的示例来解释这一点。考虑图3(a)中的四个团，它们分别代表四篇不同的文章。这四个团的作者人数分别为 5、3、4 和 3，每个节点都用字母 A 到 O 进行标记。这四个团可以通过两种原则相互连接。第一种方式如图3(b)所示，其特点是不存在度非常高的节点。假设某个人分别与两组不同的人合作撰写了两篇文章，那么代表此人的节点就会位于两个团之间。从图中的示例来看，我们假设 A 和 G 是同一个人，因此，如图3(b)所示，它们将两个团连接起来。连接两个团的另一种方式是在撰写不同文章的人之间添加边。在这个示例中，我们在 F 和 I 之间添加一条边，这表明这些人此前曾与不同的人合作，但现在他们合作共同撰写文章。同样，J 和 O 是同一个人，而 L 和 N 也是同一个人。这种连接模式的形成是由于不存在度非常高的节点，因为在这个连通网络中，没有任何节点具有非常高的度。其结果是，我们得到了一长串彼此连接的团，这与我们在图2(c,d)中看到的情况一致。另一方面，连接这些团的另一种方式如图3(c)所示，即某一个人与其他人共同撰写了许多文章。在这种情况下，该人物被表示为节点 A，他与所有其他作者都有合作关系。这使得某个节点具有非常高的节点度，并在很大程度上降低了整体平均路径长度。正如我们在图3(c,d)中看到的长路径以及图3(a)中的短路径所示，这两种连接模式都存在于 Geometry 网络中。

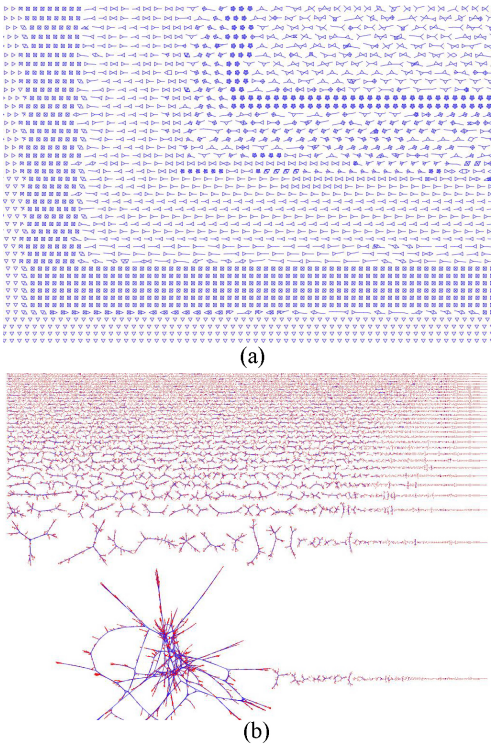

图 5. Dblp2008 网络的 Max_d-DIS 图的可视化。(a) 部分 Max_5-DIS (b) 部分 Max_{10}-DIS。

让我们总结一下本节所阐述的内容。我们尝试使用 Max_d-DIS 分解来分析 Geometry 网络，并发现了若干有趣的现象，例如：

1) 研究边在低度节点中的分布情况 2) 观察网络结构，例如较大尺寸三元组和团的形成 3) 分析枢纽节点缺失时节点的连通性，并观察网络是否容易受到定向攻击 4) 缺少高度节点时，节点会分裂成若干互不连通的分量，这启发我们考虑将连通分量分组成簇 5) 当 d取值较高时，即使没有度非常高的节点，较小的分量也开始合并为一个连通分量 6) 在缺少度非常高的节点时，DIS 子图中最大连通分量的平均路径长度显著增加，这表明度非常高的节点也有助于降低整体平均路径长度 7) 我们观察到两种连接模式：一种出现在存在度较高的节点时，另一种出现在度非常高的节点出现时。这些模式会影响连通分量的平均路径长度

以上所有发现都曾通过其他方法进行研究，而且在大多数情况下，研究不同属性所用的方法也各不相同。我们的贡献不在于提出的方法能发现什么，而在于提供一种统一且简洁的

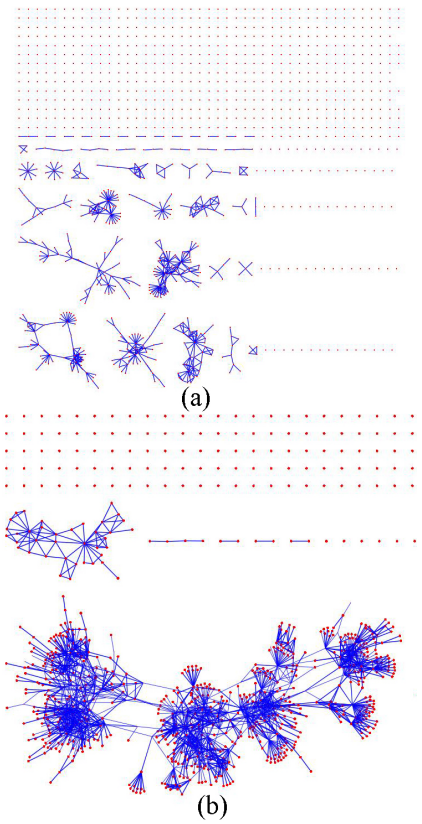

图 6. AirTransport 网络的 Max$_d$-DIS 图可视化。(a) Max$_{25}$-DIS (b) Max$_{50}$-DIS 的一部分。

方法，帮助领域专家独立开展分析并提取知识，而无需计算机科学专家的协助。通过对 Geometry 网络的详细分析，我们将进一步考察一些其他现实世界网络，看看是否也能在其中观察到类似发现。

接下来，我们考察另一个合著网络：Dblp2008 网络。图 5(a) 和 (b) 展示了该网络的 Max$_5$- DIS 和 Max$_{10}$-DIS。由于 Max$_5$-DIS 包含 60000 多个节点，Max$_{10}$-DIS 包含 80000 多个节点，因此这两幅图都只展示了整个子图的一部分。其分析与 Geometry 网络的分析颇为相似：在 Max$_5$-DIS 中，我们可以看到许多较小的连通分量，它们形成了大小在 1 到 5 之间的团。图 5(b) 清晰地显示，这些较小的分量开始合并为一个大的连通分量。这表明，即使度值为 10，较小的连通分量也开始合并为一个连通分量。

我们来看另一个例子：AirTransport 网络。图 6(a) 展示了 Max$_{25}$-DIS，图 6(b) 展示了 Part of Max$_{50}$-DIS 的一部分。在这个例子中，我们选择了相对较高的 d 值，因为在较低的 d 值下，节点或边的数量不足以观察到有趣的现象。从图 6(a) 可以看出，与前面的例子一样，该网络中存在许多非连通分量。请注意，这意味着这些网络需要度较高的节点才能彼此连接。在网络结构中，我们可以在图 6(a) 中看到三元组和星形结构。如果考察整体聚类系数，三元组的存在

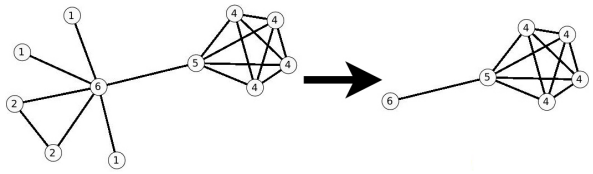

图 7. 计算 Min$_4$−DIS 前后 Min$_d$−DIS 的示例。

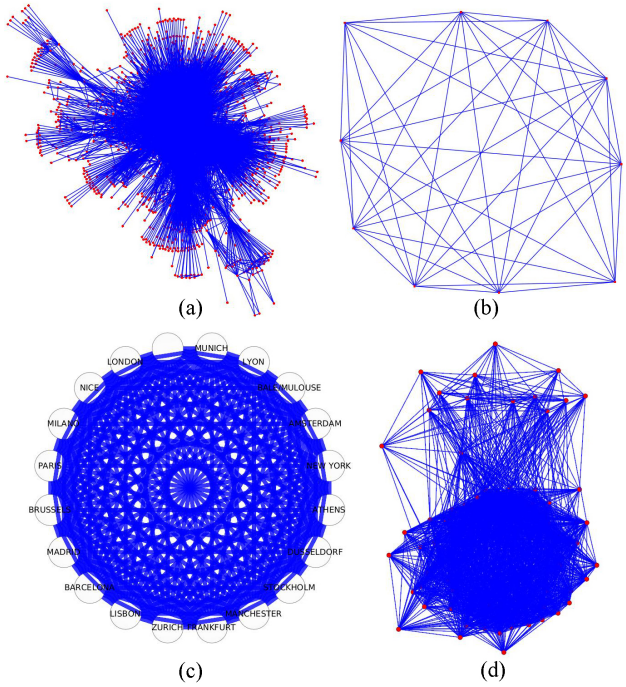

图 8. AirTransport网络的 Min$_d$-DIS 图可视化 (a) 整个网络 (b) 包含度最高的 10 个节点的 Min$_{250}$-DIS (c) 包含度最高的 20 个节点的 Min$_{185}$-DIS (d) 包含度排名前 5% 节点的 Min$_{94}$-DIS。

这一点可从相对较高的 0.49 值中得到证实。仔细观察图 6(b) 中采用 Max$_{50}$-DIS 得到的最大连通分量，可以发现其中有许多星状结构，也就是说，许多度为 1 的节点连接到同一个节点上。这是一个非常有趣的研究案例，因为我们看到，这个网络很好地融合了我们所识别出的两种基本结构：团和星。此外，Max$_{50}$-DIS 中的平均节点度为 1.49，与总体平均度 10.7 相比非常低。结合最高节点度 487 来看，这些数值表明，高度节点与网络中的其他节点连接得非常紧密，从而将平均节点度推高至 10.7。图 6(b) 中最大的连通分量包含 832 个节点和 1941 条边，占节点总数的 54%。该分量中节点的平均路径长度为 5.9，而总体平均路径长度为 2.9；两者差异显著，这再次表明，度极高的节点使整个网络的总体平均路径长度降低。

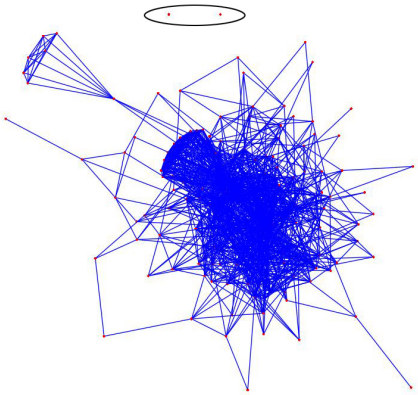

图 9. 具有 5% 最高度节点的 Geometry 网络的 Min_{17}-DIS。

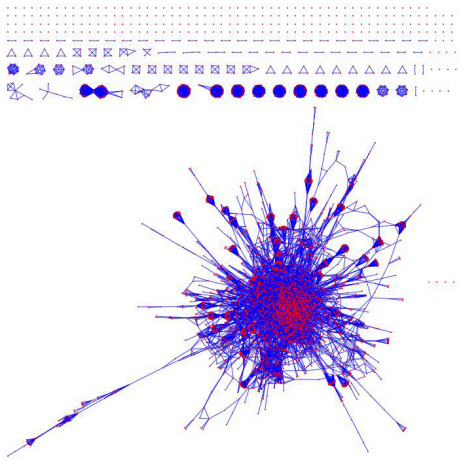

图 10. Dblp2008 网络中度最高的 5% 节点的 Min_{17}-DIS。

B. Min_d-DIS：深入了解

本节将详细介绍 Min_d-DIS 分解，以及它如何帮助我们分析网络。考虑左侧图 7 中的图及右侧对应的 Min_4-DIS。子图中仅保留度至少为 5 的节点，并移除低度节点。根据定义，该方法只考虑高度节点，因此有助于我们考察网络中高度节点之间的边是如何分布的。在分析现实世界网络时，我们以 AirTransport网络为例，使用 Min_d-DIS 进行详细分析。图 8(a) 展示了整个网络，其中包含 1540 个节点和 16523 条边。

图 8(b) 展示了网络的 $\text{Min}_{250}-$DIS，其中显示了网络中度最高的前 10 个节点，它们通过 44 条边相连，仅差一条边即可构成一个团。这表明，世界上连接最广泛的机场彼此之间都有直达航班，只有一种情况例外。城市的高度指的是它与其他城市有许多不同的连接，而非交通流量大，也不代表世界上最繁忙的机场。图 8(c) 采用环形布局绘制了世界上连接最广泛的前 20 个机场。我们还为节点标注了城市名称。这些节点通过 189 条边相连，仅缺少一条边即可构成一个团。如此高的连通性高度节点本质上源于航空公司的设计策略以及世界各地任意两座城市之间的连通性。由于目标是尽可能减少从一个地方到另一个地方所需的跳数，网络中的所有枢纽节点都彼此直接相连。图 8(d) 展示了网络中度最高的前 5% 节点。其中 77 个节点通过 1822 条边相连，平均度为 23。在这个子图中，可以用平均路径长度来衡量这种高连通性，其值为 1.3；相比之下，整个图的平均路径长度为 2.93。这一差异在此背景下非常显著。请注意，该网络中没有任何节点与其他节点断开连接，因此，如前所述，通过提高枢纽节点的连通性来提升交通运输效率，其标准显然是尽可能减少跳数。

此示例展示了 Min_d-DIS 如何帮助分析现实世界网络中枢纽节点的连通性。从上面的示例中，我们了解到这些枢纽节点彼此之间的连接非常紧密。我们使用平均路径长度来量化节点之间的接近程度。下面总结了使用 Min_d-DIS 可以执行的分析类型：

- 研究高度节点（枢纽节点）中边的分布。
- 观察高度节点的连接模式，了解它们在不同网络中如何彼此连接。我们使用平均路径长度来量化枢纽节点之间的连接紧密程度。

接下来，我们考察 Geometry 和 Dblp2008 网络。对于这两个网络，我们选取了约 5% 度最高的节点构成的 Min_d-DIS。这些网络的结果如图9和图10所示。

Geometry 网络包含 179 个节点和 1384 条边。如图 9 所示，只有两个节点未连接到最大的连通分量。该子图的平均度为 7.7，而整个网络的平均度为 2.6，这表明这些高度节点往往彼此连接。最大连通分量的平均路径长度为 2.4，而整个网络的平均路径长度为 5.3，差异同样相当显著。在这个网络中，显然跳数并非衡量效率的标准，这与 AirTransport 网络的情况不同。在这里，研究群体的社会联系对于保持网络良好的整合性发挥着重要作用；研究人员与许多不同的人合作，使该子图的平均路径长度较低。我们回到对该网络及其构建方式的描述。该数据集取自计算几何数据库，其中包含在这一领域工作的人员的引文记录。节点之间较高的连通性，以及这些高度节点较低的平均路径长度，可以用这样一个事实来解释：在同一科学领域工作的人彼此互动的可能性更高。我们可以将这个网络的高平均度和低平均路径长度特征与 AirTransport 网络的相应特征视为相似，尽管这两种特征形成背后的语义和原因截然不同。另一个合著网络是 Dblp2008，它与 Geometry 网络有显著差异。图 10 显示了

具有 3872 个节点和 20828 条边的网络的 Min_{17}-DIS。共有 326 个连通分量，这与此前使用 Min_d-DIS 分析的两个网络截然不同。虽然存在一个较大的连通分量，但也有许多非连通分量，它们内部的节点彼此连接良好，形成团。我们将这种现象称为局部峰值，因为这些人从事不同的科学领域并发表了大量文章。他们的领域不一定相互交叉，因此他们是高产作者，只与各自受限的研究群体互动，而不与其他领域的人交流。

Min_d-DIS 揭示了现实世界网络有趣的结构行为，这在分析和提高这些网络的效率时颇有用处。我们发现了关于高度节点彼此连接方式的两个有趣现象。

- 高度数的节点彼此连接紧密，形成一个大的连通分量，其平均路径长度较短（与各自整个网络的平均路径长度相比）或
- 高 度 节点会分裂成若干个连通分量，并通过较低 度 节点彼此相连

我们研究了值为 $d =$（5，10、15）的 Max_d-DIS 图，以及针对排名前 5% 节点的 Min_d-DIS 图。这些值是基于我们对不同真实世界图的分析而选取的。一种选择这些值的定量方法是依据度分布和长尾行为。如前文所述，这些图的杂乱布局是由高度节点造成的，通过查看低频值或长尾大致开始的位置即可检测到这些节点。该值不必十分精确，因为近似值已足以取得良好效果，案例研究也证明了这一点。为了研究边如何分布在低度节点中（Max_d- DIS），可以使用低于长尾起始点的值；而为了研究边如何分布在高度节点之间（Min_d-DIS），则可以使用高于该起始点的值。

整体方法的时间复杂度取决于该方法每个步骤所采用的具体技术，无论是分析还是可视化。分解本身非常高效；当在小型子图上应用平均路径长度等计算较慢的度量指标时，时间复杂度会呈指数级下降，因此整体方法相当高效。至于所提方法的可扩展性，我们研究了约有 100000 个节点和 250000 条边的图（DBLP2008 网络），使用 Tulip 软件库在标准英特尔酷睿 i3 计算机上生成所有可能的 Max_d-DIS 和 Min_d-DIS，耗时不到 1 秒，且未生成可视化表示。可视化表示取决于所使用的布局算法，而 FMMM 是已知最快的布局算法之一，能够生成清晰易读的布局图。

V. 结论与未来研究展望

本文介绍了一种基于网络拓扑分解与可视化的网络分析方法。该方法最重要的贡献在于，它能够让我们将网络可视化，并观察节点和边在不同网络中的连接方式。结合度量指标与可视化技术分析大型复杂网络，事实证明颇有助益，因为随着网络从一个领域转向另一个领域，该方法能帮助我们理解不同的连通行为。

我们发现，边在高阶和低阶节点中的分布方式呈现出一些有趣的特征。根据这种分布，我们实际上可以看到，稠密子图可能存在于高阶节点和/或低阶节点中。我们发现或重新发现了网络结构方面的一些有趣结果，例如三元组、高度数团以及星状结构的存在，这表明所提方法是有效的。

所提出的拓扑分解为可视化数据挖掘领域开辟了新的维度，因为这种方法能够简化复杂网络。我们可以研究节点连通性如何随网络节点度的变化而变化，并借此发现边在网络中的分布方式。我们认为，这种方法推动了人们更好地理解复杂网络，也为分析网络提供了重要工具，可用于网络搜索、发现有趣的模式、研究网络中节点的连通性行为，以及开发局部和群体层级指标。

参考文献

[1] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete 和 D. Fellner，“大型图的可视分析：最新进展与未来研究挑战”，《计算机图形论坛》，第30卷，第6期，第1719–1749页，2011年。[2] S. Wasserman 和 K. Faust，《社会网络分析：方法与应用。剑桥：剑桥大学出版社，1994年。[3] R. Bourqui、F. Gilbert、P. Simonetto、F. Zaidi、U. Sharan 和 F. Jourdan，“检测动态社会网络中的结构变化与指挥层级”，载于《社会网络分析与挖掘，进展国际会议》。美国加利福尼亚州洛斯阿拉米托斯：IEEE计算机学会，2009年，第83–88页。[4] R. Bourqui、F. Zaidi、F. Gilbert、U. Sharan 和 P. Simonetto，“VAST 2008挑战赛：利用手机通话模式分析社会网络动态”，载于*IEEE*可视分析科学与技术研讨会，2008年。[5] P. Comar、P.-N. Tan 和 A. Jain，“异构网络数据上的分类与社区检测同步进行”，《数据挖掘与知识发现》，第25卷，第420–449页，2012年。[6] F. Zaidi、A. Sallaberry 和 G. Melanc¸on，“揭示复杂网络中隐藏的社区结构并识别桥接节点：将网页内容分析应用于浏览”，载于《*WI-IAT '09*：*2009*年*IEEE/WIC/ACM*网页智能与智能代理技术国际会议论文集》，2009年，第198–205页。[7] C. Rozenblat、G. Melanc¸on 和 P.-Y. Koenig，“世界航空运输多层级方法中的洲际一体化（2000—2004年）”，《网络与空间经济学》，2008年。[8] S. Bilke 和 C. Peterson，“引文网络与代谢网络的拓扑性质”，《物理评论*E*》，第64卷，第036106页，2001年。[9] L. A. Adamic，“小世界网络”，载于《第三届欧洲数字图书馆研究与先进技术会议论文集》，《计算机科学讲义》丛书，第1696卷。Springer-Verlag，1999年，第443–452页。[10] U. Brandes 和 T. Erlebach，《网络分析：方法论基础（计算机科学讲义）》。Springer，2005年3月。[11] D. J. Watts 和 S. H. Strogatz，“‘小世界’网络的集体动态”，《自然》，第393卷，第440–442页，1998年6月。[12] A. L. Barab´asi 和 R. Albert，“随机网络中标度的涌现”，《科学》，第286卷，第5439期，第509–512页，1999年。[13] L. C. Freeman，“基于中介性的中心性测度集”，《社会计量学》，第40卷，第35–41页，1977年。

[14] M. C. de Oliveira 和 H. Levkowitz，“从可视化数据探索到可视化数据挖掘：综述”，《可视化与计算机图形学汇刊》，第9卷，第3期，第378–394页，2003年。[15] F. Zaidi，“复杂网络的分析、结构与组织”，博士论文，波尔多第一大学数学与计算机科学博士生院，法国，2010年。[16] D. A. Keim，“信息可视化与可视化数据挖掘”，IEEE《可视化与计算机图形学汇刊》，第8卷（第1期），第1–8页，2002年。[17] Y. Jia、M. Garland 和 J. C. Hart，“使用分层边束进行社会网络聚类与可视化”，《计算机图形论坛》，第30卷，第8期，第2314–2327页，2011年。[18] I. Herman、G. Melanc¸on 和 M. S. Marshall，“信息可视化中的图可视化与导航：综述”，IEEE《可视化与计算机图形学汇刊》，第6卷，第1期，第24–43页，2000年。[19] N. Henry、J.-D. Fekete 和 M. J. McGuffin，“Nodetrix：社会网络的混合式可视化”，IEEE《可视化与计算机图形学汇刊》，第13卷，第1302–1309页，2007年11月。[20] V. Batagelj 和 M. Zaversnik，“广义核心”，载于 CoRR，2002年。[21] J. I. Alvarez-Hamelin、L. Dall’Asta、A. Barrat 和 A. Vespignani，“k-核心分解：大型网络可视化工具”，《神经信息处理系统进展》，第18卷，第41–50页，2006年。[22] S. Wuchty 和 E. Almaas，“剥离酵母蛋白质网络。”《蛋白质组学》，第5卷，第2期，第444–449页，2005年2月。[23] S. E. Schaeffer，“图聚类”，《计算机科学评论》，第1卷，第1期，第27–64页，2007年8月。[24] J. Abello 和 J. Korn，“Mgv：大规模多重有向图可视化系统”，IEEE《可视化与计算机图形学汇刊》，第8卷，第1期，第21–38页，2002年。[25] F. Boutin、J. Thievre 和 M. Hasco¨et，“针对具有小世界现象的幂律网络的焦点式过滤聚类技术”，载于 VDA’06：《可视化数据分析——SPIE-IS&T电子成像》，第6060卷。SPIE P.，美国，2006年，第001–012页。[26] M. Kreuseler 和 H. Schumann，“一种灵活的可视化数据挖掘方法”，第8卷（第1期），第39–51页，2002年。[27] A.-C. Gavin、M. Bosche、R. Krause、P. Grandi、M. Marzioch、A. Bauer、J. Schultz、J. M. Rick、A.-M. Michon、C.-M. Cruciat、M. Remor、C. Hofert、M. Schelder、M. Brajenovic、H. Ruffner、A. Merino、K. Klein、M. Hudak、D. Dickson、T. Rudi、V. Gnau、A. Bauch、S. Bastuck、B. Huhse、C. Leutwein、M.-A. Heurtier、R. R. Copley、A. Edelmann、E. Querfurth、V. Rybin、G. Drewes、M. Raida、T. Bouwmeester、P. Bork、B. Seraphin、B. Kuster、G. Neubauer 和 G. Superti-Furga，“通过系统分析蛋白质复合物揭示酵母蛋白质组的功能组织”，《自然》，第415卷，第6868期，第141–147页，2002年1月。 [在线]。网址：http://dx.doi.org/10.1038/415141a[28] S. Hachul 和 M. J¨unger，“利用势场式多层级算法绘制大型图”，《图绘制》，第285–295页，2005年。[29] G. Simmel，《乔治·齐美尔社会学》／Kurt H. Wolff 翻译、编订并撰写导言。Free Press，伊利诺伊州格伦科，1950年。[30] S. Janson、D. E. Knuth、T. Łuczak 和 B. Pittel，“巨型连通分量的诞生”，ArXiv 数学电子印本，1993年9月。[31] Albert、Jeong 和 Barabasi，“复杂网络的抗错性与抗攻击性”，《自然》，第406卷，第378–382页，2000年。