第 1 章
大规模图数据处理：问题与挑战

1.1 大图数据处理的背景

随着云计算等新技术的快速发展、社交网络等新型互联网应用的兴起和各种电子设备的日益普及，人类获取和存储数据的规模正以前所未有的速度爆炸式增长，与大数据相关的技术变革成为学术界和工业界的热点问题。《中华人民共和国国民经济和社会发展第十二个五年规划纲要》提出要重点研究海量信息处理及知识挖掘的理论和方法，从国家战略层面上强调了对大数据的研究。而大图是大数据研究领域的一个重要分支，具有广泛的理论研究和应用价值。作为计算机科学中最常用的一类抽象数据结构，图能够有效地表达对象之间广泛存在的联系，在结构和语义方面比线性表和树更为复杂，更具有一般性表示能力。

现实世界中的许多应用场景都需要用图结构表示，与图相关的处理和应用几乎无所不在。传统应用如最优运输路线的确定、疾病暴发路径的预测、科技文献的引用关系等；新兴应用如道路交通管理、社交网络分析、语义 Web 分析、生物信息网络分析等。如在蛋白质交互网络中，图的顶点对应着蛋白质，边对应着蛋白质之间的联系；再如化合物的分子结构就可以被抽象为无向标号图，其中图的顶点对应着原子而边则可以表示成原子间的化学键；在社交网络中，图的顶点表示具体的用户，边表示用户之间的好友或者关注关系，相关的属性可以记录在对应的点和边上。此外，信息科学中的资源描述框架（RDF）文件、XML 文件、文本检索，以及机械工程中技术图纸的基本对象建模和通信网

络中集成电路的布局布线等领域也都大量使用了图数据。针对这些快速增长、形式多样且语义丰富的图数据，如何开发有效并且高效的查询分析技术，成为具有重要应用价值的课题。

传统的图数据库和图数据管理技术通常针对彼此独立的“小图”分别进行处理，尽管图的数目可能较多，但通常不需要复杂的迭代过程，也不会产生大量的消息，算法的时间和空间开销一般较低。然而，近年来，随着互联网的普及和 Web 2.0 技术的推动，以及生物化学等科学数据采集手段的丰富，产生了众多规模巨大且结构复杂的“大图”甚至“超大图”。以互联网和社交网络为例，据报道，Google 索引的网页数目早已超过了 1 万亿（10^{12}）幅，Facebook 2012 年的活跃用户已经超过 10 亿。特别地，这种发展的势头非常迅猛，据 CNNIC 统计，2010 年中国网页规模就已经达到 600 亿，年增长率 78.6%，国内如微信、微博等社会媒体的发展也异常迅猛。而在生物信息学领域，人脑级别的图建模已经达到了 10^{14} 的规模。

真实世界中实体规模的扩张，导致相应图模型的数据规模迅速增长，动辄有数十亿个顶点和上万亿条边。以搜索引擎中常用的 PageRank 计算为例[1]，网页用图顶点表示，网页之间的链接关系用有向边表示，一个网页的 PageRank 得分根据网页之间相互的超链接关系计算而得到。假设按邻接表形式存储 100 亿个图顶点和 600 亿条边，每个顶点及出度边的存储空间占 100 字节，那么整个图的存储空间将超过 1TB。值得注意的是，庞大的顶点和边数目构成的结构信息只是大图数据规模惊人的冰山一角，复杂应用中的图数据为了表达复杂的语义，因此在顶点和边上往往会附带各类属性信息，这些属性信息内容丰富，需要大量的空间开销。此外，除了静态的结构和属性信息，相比于基于属性的简单查询和搜索，大图上的统计分析算法往往需要基于图的结构进行循环和递归操作，直至达到收敛条件，因此需要频繁地处理并行迭代过程中由于通信交互产生的消息数据等中间结果。面对如此大规模的静态和动态数据，对其存储、索引、查找和分析等处理的时间开销和空间开销远远超出了传统集中式图数据管理的承受能力。对大规模图数据的高效管理和计算，已经成为急需解决的问题，也是一项极具挑战性的工作。

1.2　图数据的表示

作为数学的一个重要分支，图论以图作为研究对象，在简单图的基础上衍生出超图理论、极图理论、拓扑图论等，使图可以从多方面表达现实世界。当前大规模图数据管理，采用的数据模型有多种，按照图中节点的复杂程度分为简单节点图模

型和复杂节点图模型，按照一条边可以连接的顶点数目分为简单图模型和超图模型。不论是简单图模型、超图模型、简单节点模型还是复杂节点模型，它们的顶点和边都可以带有属性。

1. 简单图模型

这里所说的简单图，并不是图论中的简单图，是相对于超图而言的，一条边只能连接两个顶点，可以存在环路。其形式化表示，即 $G=(V, E)$。其中，$V=\{v_1, v_2, \cdots, v_n\}$，$E=\{e_1, e_2, \cdots, e_n\}=\{\{v_1, v_2\}, \{v_3, v_4\}, \cdots, \{v_{n-1}, v_n\}\}$。简单图的存储和处理都比较容易，对于一般的应用，简单图的表达能力完全可以胜任，如 PageRank 计算、最短路径查询等。简单图模型的常用组织存储结构包括邻接矩阵、邻接表、十字链表和邻接多重表等多种方式。不同的系统根据目标不同采用不同的表示方式。

2. 超图模型

一条边可以连接任意数目的图顶点。此模型中图的边称为超边。基于这种特点，超图比上述简单图的适用性更强，保留的信息更多。形式化表示为 $G=(V, E)$，其中，$V=\{v_1, v_2, \cdots, v_n\}$，$E=\{e_1, e_2, \cdots, e_n\}=\{\{v_1, v_2, v_3\}, \{v_3, v_4, v_5\}, \cdots, \{v_{n-2}, v_{n-1}, v_n\}\}$。例如，以图顶点代表文章，每条边代表两个顶点（文章）享有同一个作者。现有三篇文章 v_1（作者 A、B），v_2（作者 A、C），v_3（作者 A、D），三篇文章的作者都有 A。图 1-1 的左图表示了简单图存储模式，边集 $E=\{e_1, e_2, e_3\}=\{\{v_1, v_2\}, \{v_1, v_3\}, \{v_2, v_3\}\}$，无法直接保留作者 A 同时是三篇文章 v_1, v_2, v_3 的作者这一信息。图 1-1 的右图代表了超图存储模式，边集 $E=\{e_1\}=\{\{v_1, v_2, v_3\}\}$，超边 e_1 中直接保留了 A 是三篇文章 v_1, v_2, v_3 的作者这一信息。对于具有复杂联系的应用，可以使用超图模型建模，例如社交网络、生物医学网络等。

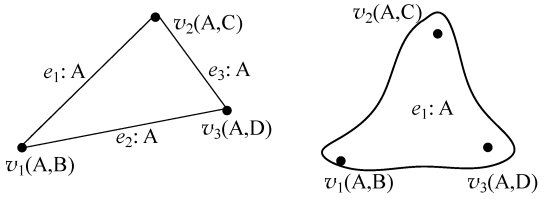
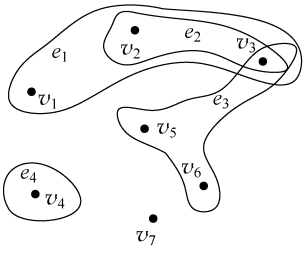


图 1-1 简单图（左）和超图（右）

超图模型的组织方式主要使用关系矩阵，它与邻接矩阵较为相似。图 1-2 展示了超图的关系矩阵表示方法。与邻接矩阵不同的是，关系矩阵的行和列分别表示图顶点编号和超边的编号，关系矩阵中，1 表示一条超边包含某个图顶点，或一个图顶点隶属于某条超边。



	e_1	e_2	e_3	e_4
v_1	1	0	0	0
v_2	1	1	0	0
v_3	1	1	1	0
v_4	0	0	0	1
v_5	0	0	1	0
v_6	0	0	1	0
v_7	0	0	0	0

图 1-2　超图的关系矩阵表示方法

1.3　传统的大图数据管理方法

传统的分布式图处理技术主要基于 NoSQL 数据库，关注的是图数据、特别是小图数据集的管理问题，查询相对简单，具有高效的索引，可以支持数据的更新。当图数据更新时，需要解决在分布式环境下的一致性控制问题，提供事务功能。

特别地，NoSQL 数据库采用的数据模型主要有文档存储(Document Store)模型、列族存储(ColumnFamily Store)模型、Key-Value 存储模型、图形模型等几大类[2]。

Key-Value 存储模型的存储模式简单，在高并发环境下可以提供高效的查询或遍历服务，能存储海量数据，非常适合通过主键进行查询或遍历，但对复杂的条件查询支持度不佳。从图处理的角度看，像 PageRank 计算等并不需要复杂查询，Key-Value 模型完全可以胜任。对于采用邻接表组织的图数据，可以将图顶点及其值作为 Key，将出边或出度顶点列表作为 Value。文献[3]结合语义 Web 和传统的 Key-Value 模型，提出 Key-Key-Value 模型。以社交网络为例，Key-Key-Value 模型将 Alice 和 Bob 之间的好友关系组织为一个三元组〈Alice，Bob，FriendShip〉。该模型存储的信息比传统的 Key-Value 模型更加丰富，可以据此进行数据迁移和合并，以提高时空局部性，使得在查询处理时能减少数据远程跨机读取的次数，因而可以提高数据读取效率。文档存储模型在存储格式方面十分灵活，比较适合存储系统日志等非结构化数据，对以邻接矩阵或邻接表组织的图数据来讲，意义不大。而且文档存储模型为支持灵活性所导致的处理效率的降低也会成为大规模图数据管理的瓶颈。列族存储模型比较适合对某一列进行随机查询处理，但是对于穷举式遍历，反而不如传统的面向行的存储模式。图模型的相关研究目前还不完善，只有少数分布式图数据库，如 Neo4j[4] 等采用这种模型存储图数据。文献[5]从管理数据的规模和模型的复杂性两个维度比较了这 4 种基本存储模型，见图 1-3。

在查询处理方面，传统的分布式图数据库和一些新型的大图在线查询系统往往只支持图的简单查询检索，返回用户感兴趣的信息，而通常不支持大图上的复杂

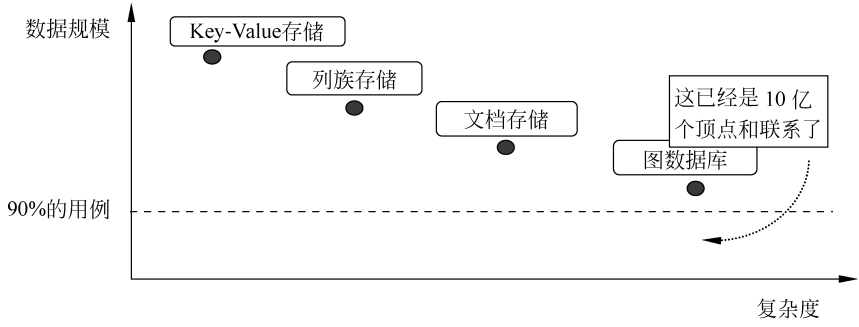


图 1-3 NoSQL 数据库的 4 种主要模型的数据规模和复杂度比较

迭代计算、分析和挖掘任务。检索过程中，对于某些应用，通过建立合适的索引、调整查询顺序和查询复用等技术，可以避免对整个图顶点的遍历，有效提高检索效率。从所处理的查询请求和优化技术方面考虑，此类图查询类似于普通的数据库查询。传统集中式数据库不仅为用户提供了良好的 SQL 查询语言接口，还通过索引和查询优化，提供高效的查询。这里简单介绍 Horton[6] 和 HyperGraphDB[7]。

Horton 是微软正在设计的一款专门针对大规模图数据提供高效在线查询服务的系统。Horton 为减少查询时的网络通信代价，对图数据进行专门分割，以提高局部性。在查询语言和处理机制方面，Horton 将用户的具体查询请求切割为若干原语，按照有限自动机组织指导底层的分布式广度优先搜索执行顺序。至于查询优化方面，在具体应用中（如好友关系和照片关注关系），研究发现，以不同的顺序执行有限自动机，查询代价相差很大。如图 1-4 所示的查询，如果按照图中边的序号进行广度优先搜索，则方案 1 的代价（即处理步骤数）为 8，而方案 2 的代价为 4，优于 1 号方案。Horton 采用基于图顶点出度统计信息的预处理优化技术，能够以较低的预处理代价获得较优的自动机执行顺序。在高并发环境下，Horton 还可以复用不同查询请求的原语。此外，Horton 可以根据查询的历史统计信息，在常用查询顶点之间建立“衍生边”并持久化存储，以空间换时间，可以减少后续查询过程中图顶点的遍历开销。如图 1-4 所示，对李四朋友关注的照片建立“衍生边”，在 3 号方案中，可以直接定位找到查询结果，代价比 2 号方案更优。

HypergraphDB 是一个分布式图数据库，支持超图模型，底层以 Key-Value 格式存储，可以建立 B-tree 索引，为用户提供查询原语接口以表达具体的查询请求，如图顶点和边条件限定、比较操作、连接操作等，可以通过广度优先搜索或深度优先搜索返回符合条件的信息集合。

总体来看，图数据库和大图在线查询系统所支持的查询处理都比较简单，并不考虑迭代计算的优化问题，无法完成类似最短路径计算等基本计算需求，也无法支持复杂的分析和挖掘任务。本书主要关注的是那些面向复杂任务的大图计算系

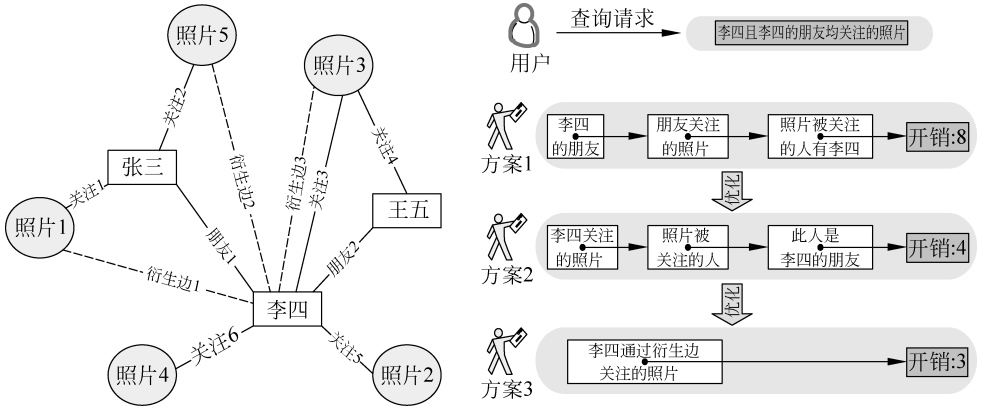


图 1-4　微软 Horton 系统查询处理示意图

统，特别是离线分析批处理系统的研究进展和关键技术，而不重点研究传统的图数据管理系统和面向简单任务的在线图查询系统。

1.4　云计算环境处理大图数据的优势

近些年来，云计算以其在大规模数据处理方面的诸多优势，得到了学术界和产业界的广泛关注。云计算是网格计算、分布式计算、并行计算、效用计算、网络存储、虚拟化等先进计算机技术和网络技术发展融合的产物，具有普遍适用性。云计算技术的发展，一直与大规模数据处理密切相关。因此，依靠云计算环境对大规模图数据进行高效处理，是一个非常有发展潜力的方向[8]，其主要优势表现在：

（1）海量图数据的存储和维护　大规模图的数据量可达几百 GB 甚至 PB 级别，难以在传统文件系统或数据库中存储，而云计算环境提供分布式存储模式，可以汇聚成百上千普通计算机的存储能力和计算能力，提供高容量的存储服务，完全能够存放和处理大规模的图数据。云计算环境下的并发控制、一致性维护、数据备份和可靠性等控制策略，可以为大规模图数据的维护提供保障。

（2）分布式并行处理　利用云计算分布平行处理的特点，可以将一个大图分割成若干子图，把针对一个大图的处理分割为若干针对子图的处理任务。云计算分布式并行运算能力，能够显著提高对大规模图的处理能力。

（3）良好的可扩展性和灵活性　从技术角度和经济角度讲，云计算环境具有良好的可伸缩性和灵活性，非常适合处理数据量弹性变化的大规模图问题。云计算环境通常由廉价的普通计算机构成。随着图数据规模的不断增大，可以向云中动态添加节点来扩展存储容量和计算资源，而无须传统并行机模式的巨大投资。

总的来说，除了云计算本身具有的经济收益，大图数据处理主要是利用云环境

下基于集群的分布式存储和并行处理策略，系统应具有可扩展性和高容错性的特征，能够通过扩展廉价的计算节点支撑不断增长的图数据规模。在以云计算为代表的分布式处理环境下，侧重大图离线分析的新一代大图处理系统的研发目标已经发生变化。从传统的分布式图数据库系统转变为分布式并行大图计算系统，这类系统往往提供的只是计算平台，从外部的文件系统或数据库加载数据，特别地，磁盘驻留的分布式计算系统还支持为特定任务定制的高效图数据和消息数据的索引，从而优化 I/O 的读取。这些面向大数据环境下基于集群的计算任务的优化目标与传统数据库管理系统的数据维护和简单属性检索目标有着很大的区别，面临着许多新的挑战。

1.5 新型大图计算系统面临的挑战

虽然云计算环境对于大规模图数据的管理有诸多优势，但是由于云计算只是一个通用的处理框架，而且其本身也正处于发展阶段，采用云计算为代表的分布式环境对大规模图数据进行处理，仍有很多关键技术难题需要解决。图计算及其分布式并行处理通常涉及复杂的处理过程，需要大量的迭代和数据通信，针对联机事务处理和传统数据仓库等应用的技术很难直接应用到图数据处理中，需要研究新型的大图计算系统。总体上，面向复杂分析任务的大规模图数据分布式处理主要面临以下典型挑战。

(1) 图数据的强耦合性和动态性　由于图顶点数据之间的关联性，其并行计算过程具有强耦合性，增大了并行处理过程中消息通信的开销。同时，这也增加了图拓扑结构的复杂性、随机性，使数据访问的局部性差，加剧了并行处理过程中的“水桶效应”，即整个查询的完成，依赖于执行最慢的子任务完成。这就需要合理划分和组织数据，优化消息处理，适应性地调整数据分布，以提高局部性，维护负载均衡。此外，数据随着时间动态变化是大数据时代的又一典型挑战。这种动态性首先体现在新数据的不断增加，例如 Facebook 每月新增用户达到 2000 万，每 20 分钟新增 197 万好友关系；更复杂地，还涉及对存在联系的更新和删除而导致图结构的逐渐变化，例如社交网络上的好友关系会不断地增删，粉丝对名人的关注也会随着名人的流行度迁移和演化。因此我们需要处理的不是静态的图数据，而是不断扩充且动态演化的动态图。特别地，面向图数据流的内存实时更新和处理框架，并不适合针对超大图的复杂离线迭代分析任务。为了更好地管理动态图以高效支持迭代任务，需要探索低成本高时效的批处理更新方法；需要设计可高效增量维护的划分、存储、索引和概要等关键技术，避免重新组织数据；还需要在针对连续查询的任务时，研究增量计算方法，避免重复处理。

(2) 图查询处理的复杂迭代性　复杂的图算法通常采用递归调用方式。在实

现上，图的计算需要多次迭代处理，即从起始顶点出发进行迭代，直到搜索到所有顶点，然后收敛到最终结果。期间，图的状态也会随之发生变化。相邻迭代之间，会产生大量中间结果，导致消息通信和磁盘操作代价高昂，而迭代的累积，也会放大“水桶效应”。特别地，一个处理任务的迭代频次通常跟图数据的规模有关，对于超大图，并行迭代处理任务往往需要很多个迭代周期才能收敛，例如对一个 10 亿顶点和 100 亿边的超大图进行单源最短路径查询，迭代处理次数通常在 100 次以上，对于直径较大的稀疏图，这一数字还会成倍增加。磁盘驻留环境下，这种高频迭代性导致的反复 I/O 读取和消息通信，将会极大地影响查询和分析的效率；因此，需要针对迭代过程进行全面优化，提出减少迭代和加快收敛的本地计算与消息通信机制。此外，不同的迭代任务在不同迭代周期的激活顶点数目和变化规律不同，具有明显的状态差异，考虑到不同顶点在不同阶段状态的迭代周期有着不均衡的计算和消息负载，需要有效地建模和监控迭代周期的非线性变化，有针对性地调整本地计算和数据交互的策略。因此，应充分利用迭代过程中图的状态转换，进行查询优化；通过选取合适的图计算模型，避免迭代过程中反复启动任务，降低中间结果的规模；同时，需进行有效的同步控制和消息通信优化，减少通信开销，减轻“水桶效应”。

（3）图计算任务的可调可控性　对大规模图处理，需要相对较长的时间来完成计算任务，执行过程中需消耗大量的时间和资源。随着集群数目的增加，系统的平均故障时间将会显著较少，系统的可用性和可靠性成为关键的问题。这对云计算环境的执行保障机制，提出了严格要求，需要研究富有弹性的资源管理机制和高效的容错控制机制，以便能够在图计算处理过程中，动态调整优化集群的资源分配，使整个集群实现动态负载均衡，并降低故障探测和故障恢复开销。此外，针对高复杂度的计算任务，如何设计高效的采样、概要、压缩以及提前终止迭代等近似处理方法，同时满足高精度的准确性保证，具有重要的意义，也面临着巨大的挑战。如何根据图数据的结构、处理任务的性质、用户的服务质量需求动态地优化系统性能，实现可调整、可增量、可近似的自适应任务管控机制，为用户提供多样化、高性能、可订制的服务，也是重要的研究课题。

1.6　关键技术问题

面向复杂迭代计算任务的大图数据处理是一项系统性的工作，涉及诸多关键的技术。我们从以下几方面进行说明。

1. 计算方法

云环境下的大数据分布式处理为了满足不断增长的数据量，需要系统有很高

的扩展性，设计有效的计算框架是基础的问题，决定分布式并行执行方式，是进行解耦处理和提高可靠性的基础。此外，依赖于云环境的大规模图处理，任务负载重，执行时间长，因此执行框架必须提供高可靠性、高灵活性、高效率和高伸缩性的执行机制，包括消息通信、同步控制、容错管理、任务调度、扩展性保证等关键技术为作业的高效、顺利运行提供支持。本书第 2 章将对该部分内容进行详细介绍。

2. 数据组织

决定了计算框架，最重要的基础性问题即为如何针对复杂的图计算任务合理地组织数据，从而提高整个处理过程的执行效率，主要包括图数据的划分、图数据存储和索引。首先实现低耦合的划分是实现大图分布式处理的基本操作，是保证负载均衡、减轻水桶效应的基础，特别是在处理过程中根据实时的负载变化进行动态的重划分，面临着诸多技术挑战；此外，面向复杂迭代分析任务的图数据存储和索引主要研究分布式磁盘处理系统在内存受限环境下，大图数据计算过程中的数据组织和 I/O 读写优化，包括对消息等中间结果数据的高效管理，从而减少对图和消息数据的访问次数与随机磁盘访问，这与面向简单查询任务的传统分布式数据库中的数据存储和索引技术的研究目标并不相同。本书第 3 章和第 4 章将分别介绍大图数据的划分技术以及面向分布式大图计算的存储和索引技术。

3. 特定处理任务的优化

以上讨论的计算方法和数据组织往往针对通用模型，当给定具体的查询分析任务时，可能需要研究定制的实现和优化方案。目前作为基准的大图计算任务主要有最短路径查询、Pagerank、k-means 聚类、SimRank 等，大多数系统都是针对这些算法进行评测和优化。此外，有些文献开始关注近似的分布式计算方法[9]和增量的动态图计算技术[10]。本书第 5 章到第 9 章依次介绍了几种典型的大图复杂查询、分析和挖掘算法的优化技术，包括三角形查询、最大 k 边连通子图查询、最小生成树搜索、频繁子图挖掘和重叠社区发现，以此说明如何将特定算法有效地融合到分布式计算框架中。这些计算任务有着非常重要的应用，针对它们进行的分布式计算优化方法的研究还有待进一步开展。

4. 系统实现技术

以上关键技术需要以系统的形式作为载体呈现给用户，合理设计系统的定位，选取适合的软硬件环境，提高程序开发的质量，才能将核心的关键技术高效、完整地呈现出来。近些年来，针对大图数据的磁盘系统和分布式计算系统层出不穷，采用了不同的计算模型、数据组织和优化技术，功能和适应性也不尽相同，对它们进行充分的研究可以更好地理解相关技术的实现方法。特别是大量高质量的开源图

处理系统的出现，推动了大图分布式数据处理学术研究和产业的快速发展，以它们为平台实现具体算法的优化和二次开发成为可能。本书第 10 章对现有的主要分布式大图计算系统进行了综述和分析，并对大图分布式处理的典型应用进行总结。

综上，在大数据时代，图数据的规模、形式以及图处理任务都发生了巨大的变化，传统的分布式图数据查询技术不能满足新的图计算需求。随着云计算技术的发展，利用集群进行可扩展的分布式并行大图处理具有巨大的潜力，但也面临着诸多技术挑战。本书将对分布式图处理的基础技术、典型算法和主要系统的研究现状进行总结，并着重介绍我们在大图数据分布式处理方面的研究成果。