**流式图实时并行计算引擎**

**研究背景**

由于图数据结构能够很好表达数据之间的复杂依赖关系和聚集情况，图计算引擎被广泛应用于大规模图的处理与分析，如社交网络意见领袖挖掘、商品购买推荐、舆情检测、信用卡欺诈检测等。当前主流的批量式图计算引擎如Google Pregel, Spark GraphX，GraphLab等能够对静态图进行高效并行计算。然而，现实世界中的图数据是动态变化、实时更新的，当前的批式图计算引擎无法对流式动态图进行实时并行计算。学术界提出的流式图计算方法，如对大图进行抽样计算，精确率不高。目前，还没有能够对流式图进行实时并行计算的引擎。

**研究问题**

如何构造一个可以同时达到高精度和高性能的实时流式图计算引擎？

**研究内容**

研究流式图数据并行计算引擎，解决图算法在流式图数据上难以进行并行计算、计算精度难以保障的问题。引擎主要包括：

* 流式图数据模型，也就是如何对动态变化图进行抽象表达，包括点、边、以及属性等的表达。与现有图数据模型的不同是，这里的模型需要考虑如何抽象表达图的变化情况、以及如何与静态图进行结合。
* 流式图并行计算模型，也就是如何构造一个通用的面向流式图的并行计算模型。给定图处理算法，模型不仅可以将图算法并行化，而且可以在图数据更新时，进行实时计算，得到较为精确的更新结果。
* 流式图计算编程模型，也就是如何为用户提供统一的编程接口和编程方法，使得用户将现有静态图算法，改写为流式图算法，被流式图并行计算模型实时计算。

**达到的效果：**

支持高精度实时社交网络意见领袖挖掘、商品购买推荐、舆情检测、信用卡欺诈检测等应用，引擎具有

* 实时性，现有在静态图上运行的图算法能够扩展到流式图数据上，在图数据不断更新时，计算结果同时更新。
* 高性能：可以处理上亿节点和边的大图。
* 高精度：比现有抽样等算法精度有50%的提高。

本项目研究的核心问题是如何设计和实现可以同时达到高精度和高性能的实时流式图计算引擎，具体研究内容包括：

* **设计新的基于路径传播更新的动态图计算模型**，解决如何对动态图进行分布并行计算的问题，包括如何通过控制状态更新传播范围来减少计算，如何通过选择状态更新路径来加速结果收敛，如何对状态更新并行化来提高模型的扩展性，如何解决状态更新冲突来提高计算正确率。
* **设计新的基于分块矩阵的动态图数据模型**，解决如何将新加入的图与与原始静态图结合表达问题。具体包括如何利用矩阵分块来抽象表达原始图内部关系、新加入顶点与原始顶点的有向边连接关系以及新增顶点间自身连接关系。
* **设计以状态更新顶点为中心的动态图编程模型**，相比传统模型，新的编程模型既可以进行增量计算，也可以减少非状态更新顶点的计算量。具体研究设计状态初始化函数解决原始图状态保存问题， 设计状态更新函数解决新加入顶点和边带来的增量计算问题， 设计消息传播函数解决更新状态迭代传播问题，以及设计状态冲突解决函数解决顶点接收到多个更新消息的处理问题。
* **设计新的流式图算法库：**研究如何将传统批处理算法中满足流式图编程模型的算法转换为流式图算法，构建在流式图计算引擎上。