# Graph sampling with applications to estimating the number of pattern embeddings and the parameters of a statistical relational model

基于Fürer和Kasiviswanathan(2014)提出的理论采样思想,用于在图中近似计算pattern数量

将Fürer和Kasiviswanathan(2014)的方法推广到标记和有向图

从样本中估算计算统计量的准确性

## 文章主要内容

处理对象：有向，有label，数据规模大的图

1. 发现有用的pattern（出现次数频繁，但又不是特别频繁的）

2. 估计pattern在graph中的数量。

何为有用的pattern？文章中的定义是：

We vary the pattern size from 4 to 10 nodes.

### 文章综述/相关资料

比较重要的文献：

1. Frequent Subgraph Mining on a Single Large Graph Using Sampling Techniques
   1. 利用抽样方法研究了在单个大型图上进行频繁子图挖掘的任务，发现抽样是一种可行的方法。
   2. 评估了不同的抽样方法，并提出了一种新的抽样方法，称为“随机区域选择抽样”
   3. 比目前所有具有自定义参数的图抽样方法产生了更好的结果
2. Sampling from Large Graphs
   1. 几种抽样方法，提出了检验抽样优劣的新方法，并发展了一套描述原始和样本性质之间关系的标度法则
   2. 总体而言，基于随机游动和“森林火灾”的方法表现最好;它们非常精确地匹配静态和进化图模式，样本大小减少到原始图的15%左右

### 文章思路

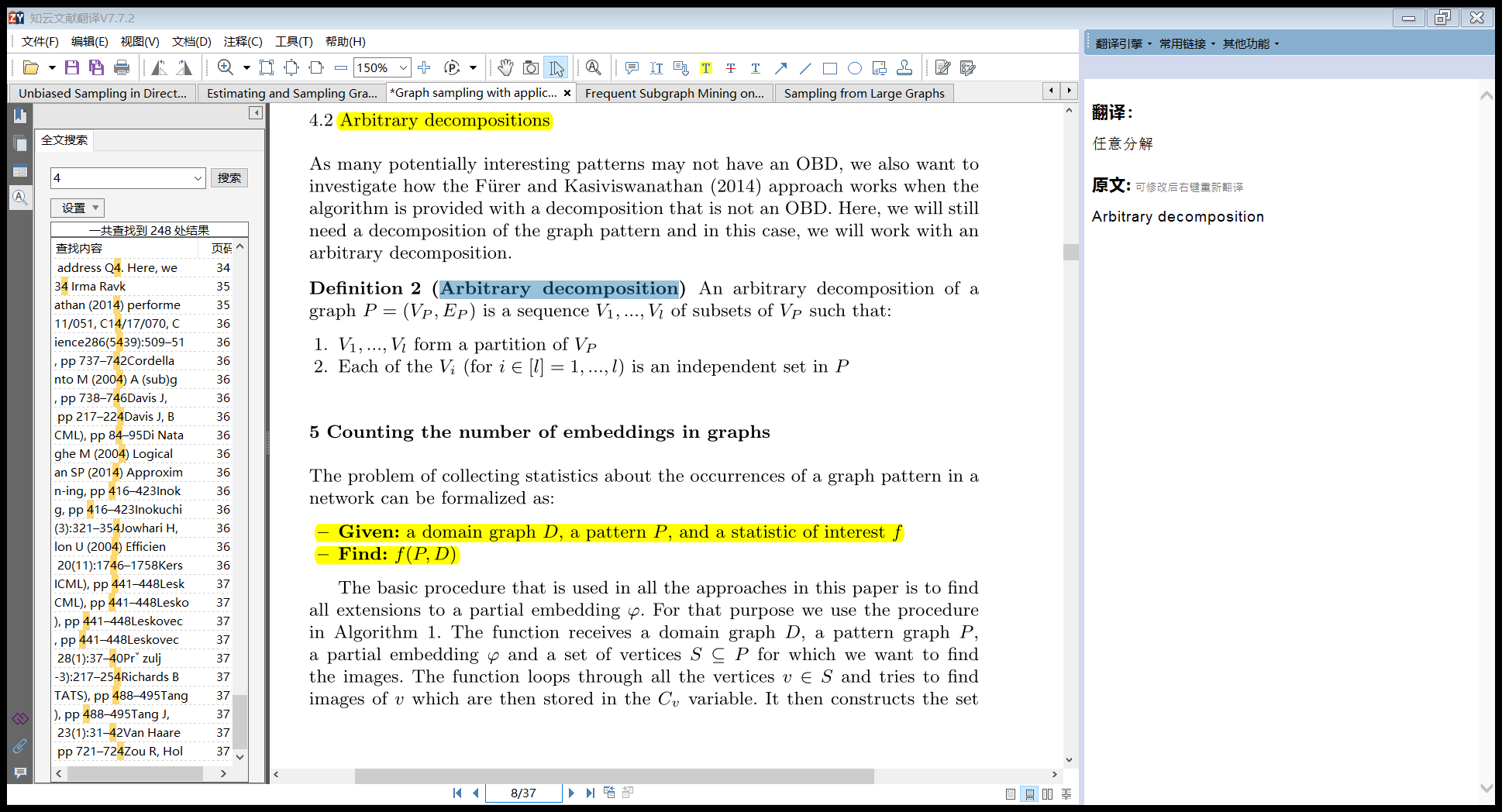
1. 图的分解

用分治的策略来获取图中pattern数量。首先将复杂图分解为多个分区，同时匹配单个分区中的顶点（这一步已可以通过分布式实现，而在每个单机上需要进一步细分）

1. Ordered bipartite decompositions

将图分成两部分，每个部分里面的点之间不会相连

1. Arbitrary decomposition

将图任意分成若干个部分，每个部分保持独立，互不相干

1. 计算图中pattern的数量
2. 循环遍历所有的顶点，找pattern，将数量存储到变量中返回
3. 