Approximate Integration of streaming data

Michel de Rougemont*,** Guillaume Vimont*,**

*University Paris II
mdr@irif.fr
**IRIF-CNRS
vimontguillaume@gmail.com

Abstract. We approximate analytic queries on streaming data with a weighted reservoir sampling. For a stream of tuples of a Datawarehouse we show how to approximate some OLAP queries. For a stream of graph edges from a Social Network, we approximate the communities as the large connected components of the edges in the reservoir. We show that for a model of random graphs which follow a power law degree distribution, the community detection algorithm is a good approximation. Given two streams of graph edges from two Sources, we define the *Community Correlation* as the fraction of the nodes in communities in both streams. Although we do not store the edges of the streams, we can approximate the Community Correlation and define the *Integration of two streams*. We illustrate this approach with Twitter streams, associated with TV programs.

1 Introduction

The integration of several Sources of data is also called the composition problem, in particular when the Sources do not follow the same schema. It can be asked for two distinct Datawarehouses, two Social networks, or one Social network and one Datawarehouse. We specifically study the case of two streams of labeled graphs from a Social network and develop several tools using randomized streaming algorithms. We define several correlations between two streaming graphs built from sequences of edges and study how to approximate them.

The basis of our approach is the approximation of analytical queries, in particular when we deal with streaming data. In the case of a Datawarehouse, we may have a stream of tuples t following an OLAP schema, where each tuple has a measure, and we may want to approximate OLAP queries. In the case of a Social network such as Twitter, we have a stream of tweets which generate edges of an evolving graph, and we want to approximate the evolution of the communities as a function of time.

The main randomized technique used is a *k-weighted reservoir sampling* which maps an arbitrarly large stream of tuples t of a Datawarehouse to k tuples whose weight is the measure $t.M$ of the tuple. It also maps a stream of edges u of a graph, to k edges and in this case the measure of the edges is 1. We will show how we can approximate some OLAP queries and