ARTICLE IN PRESS

European Journal of Combinatorics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

European Journal of Combinatorics

journal homepage: www.elsevier.com/locate/ejc

Computing heat kernel pagerank and a local clustering algorithm☆

Fan Chung, Olivia Simpson

University of California, San Diego, La Jolla, CA 92093, United States

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

Heat kernel pagerank is a variation of Personalized PageRank given in an exponential formulation. In this work, we present a sublinear time algorithm for approximating the heat kernel pagerank of a graph. The algorithm works by simulating random walks of bounded length and runs in time $O\left(\frac{\log(\epsilon^{-1})\log n}{\epsilon^3 \log\log(\epsilon^{-1})}\right)$, assuming performing a random walk step and sampling from a distribution with bounded support take constant time.

The quantitative ranking of vertices obtained with heat kernel pagerank can be used for local clustering algorithms. We present an efficient local clustering algorithm that finds cuts by performing a sweep over a heat kernel pagerank vector, using the heat kernel pagerank approximation algorithm as a subroutine. Specifically, we show that for a subset S of Cheeger ratio ϕ, many vertices in S may serve as seeds for a heat kernel pagerank vector which will find a cut of conductance $O(\sqrt{\phi})$.

1. Introduction

In large networks, many similar elements can be identified to a single, larger entity by the process of clustering. Increasing granularity in massive networks through clustering eases operations on the network. There is a large literature on the problem of identifying clusters in a graph [8,20,28,29,34,37], and the problem has found many applications. However, in a variation of the graph clustering problem we may only be interested in a single cluster near one element in the graph. For this, local clustering algorithms are of greater use.

☆ An extended abstract appeared in Chung and Simpson (2014) [12].

E-mail addresses: fan@ucsd.edu (F. Chung), osimpson@ucsd.edu (O. Simpson).

http://dx.doi.org/10.1016/j.ejc.2017.07.013
0195-6698/

Please cite this article in press as: F. Chung, O. Simpson, Computing heat kernel pagerank and a local clustering algorithm, European Journal of Combinatorics (2017), http://dx.doi.org/10.1016/j.ejc.2017.07.013.

As an example, the problem of finding a local cluster arises in protein networks. A protein–protein interaction (PPI) network has undirected edges that represent an interaction between two proteins. Given two PPI networks, the goal of the pairwise alignment problem is to identify an optimal mapping between the networks that best represents a conserved biological function. In [27], a local clustering algorithm is applied from a specified protein to identify a group similar to that protein. Such local alignments are useful for analysis of a particular component of a biological system (rather than at a system level which will call for a global alignment). Local clustering is also a common tool for identifying communities in a network. A community is loosely defined as a subset of vertices in a graph which are more strongly connected internally than to vertices outside the subset. Properties of community structure in large, real world networks have been studied in [25], for example, where local clustering algorithms are employed for identifying communities of varying quality.

The goal of a local clustering algorithm is to identify a cluster in a graph near a specified vertex. Using only local structure avoids unnecessary computation over the entire graph. An important consequence of this are running times which are often in terms of the size of the small side of the partition, rather than of the entire graph. The best performing local clustering algorithms use probability diffusion processes over the graph to determine clusters (see Section 1.1). In this paper we present a new algorithm which identifies a cut near a specified vertex with simple computations over a heat kernel pagerank vector.

The theory behind using heat kernel pagerank for computing local clusters has been considered in previous work. Here we give an efficient approximation algorithm for computing heat kernel pagerank. Note that we use a "relaxed" notion of approximation which allows us to derive a sublinear probabilistic approximation algorithm for heat kernel pagerank, while computing an exact or sharp approximation would require computation complexity of order similar to matrix multiplication. We use this sublinear approximation algorithm for efficient local clustering.

1.1. Previous work

Heat kernel and approximation of matrix exponentials. Heat kernel pagerank was first introduced in [9] as a variant of personalized PageRank [18]. While PageRank can be viewed as a geometric sum of random walks, the heat kernel pagerank is an exponential sum of random walks. An alternative interpretation of the heat kernel pagerank is related to the heat kernel of a graph as the fundamental solution to the heat equation. As such, it has connections with diffusion and mixing properties of graphs and has been incorporated into a number of graph algorithmic primitives.

Orecchia et al. use a variant of heat kernel random walks in their randomized algorithm for computing a cut in a graph with prescribed balance constraints [35]. A key subroutine in the algorithm is a procedure for computing $e^{-A}v$ for a positive semidefinite matrix A and a unit vector v in time $\tilde{O}(m)$ for graphs on n vertices and m edges. They show how this can be done with a small number of computations of the form $A^{-1}v$ and applying the Spielman–Teng linear solver [38]. Their main result is a randomized algorithm that outputs a balanced cut in time $O(m \operatorname{polylog} n)$. In a follow up paper, Sachdeva and Vishnoi [36] reduce inversion of positive semidefinite matrices to matrix exponentiation, thus proving that matrix exponentiation and matrix inversion are equivalent to polylog factors. In particular, the nearly-linear running time of the balanced separator algorithm depends upon the nearly-linear time Spielman–Teng solver.

Another method for approximating matrix exponentials is given by Kloster and Gleich in [22]. They use a Gauss–Southwell iteration to approximate the Taylor series expansion of the column vector $e^P e_c$ for transition probability matrix P and e_c a standard basis vector. The algorithm runs in sublinear time assuming the maximum degree of the network is $O(\log \log n)$.

Local clustering. Local clustering algorithms were introduced in [38], where Spielman and Teng present a nearly-linear time algorithm for finding local partitions with certain balance constraints. Let $\Phi(S)$ denote the cut ratio of a subset S that we will later define as the Cheeger ratio. Then, given a graph and a subset of vertices S such that $\Phi(S) < \phi$ and $\operatorname{vol}(S) \leq \operatorname{vol}(G)/2$, their algorithm finds a set of vertices T such that $\operatorname{vol}(T) \geq \operatorname{vol}(S)/2$ and $\Phi(T) \leq O(\phi^{1/3}\log^{O(1)} n)$ in time $O(m(\log n/\phi)^{O(1)})$. This seminal work incorporates the ideas of Lovász and Simonovitz [30,31] on isoperimetric properties

Please cite this article in press as: F. Chung, O. Simpson, Computing heat kernel pagerank and a local clustering algorithm, European Journal of Combinatorics (2017), http://dx.doi.org/10.1016/j.ejc.2017.07.013.