UC San Diego

Computing the Heat Kernel of a Graph for a Local Clustering Algorithm

Olivia Simpson and Fan Chung

We present an efficient local clustering algorithm that finds cuts in large graphs by performing a sweep over a heat kernel vector. We show that for a subset S of Cheeger ratio ϕ, many vertices in S may serve as seeds for a heat kernel random walk which will find a cut of conductance $O(\sqrt{\phi})$. Further, the random walk process is performed in time sublinear in the size of the graph.

Local Clustering in Graphs

The goal of a local clustering algorithm is to compute a cluster in a graph near a specified vertex with size and volume constraints. For example, in **social networks** local clustering algorithms are used to identify a small and dense community around a particular member. In **protein networks** local clustering is used to isolate a group of interacting proteins to analyze a component of a biological system.

A good cluster, *S*, will be a subset of nodes in the graph with small *Cheeger ratio:*

$$\partial(S) = \{u \sim v \;:\; u \in S, v \notin S\}$$

$$\mathrm{vol}(S) = \sum_{v \in S} d_v$$

$$\Phi(S) = \frac{|\partial(S)|}{\min(\mathrm{vol}(S), \mathrm{vol}(V \setminus S))}$$

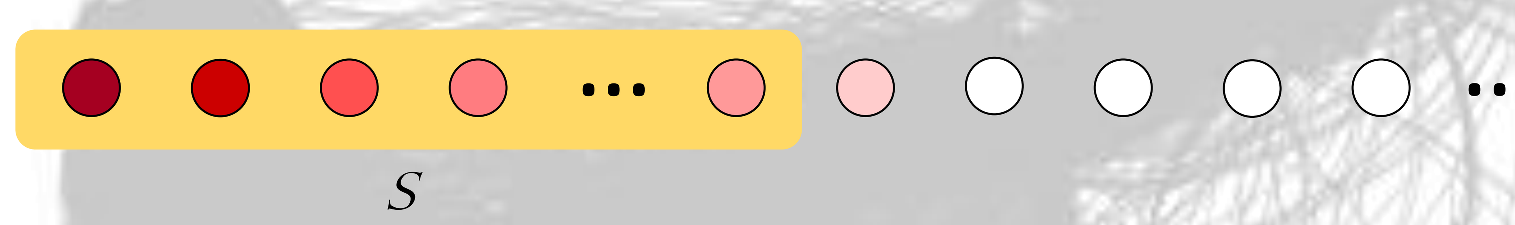
Single Sweep Algorithms

Consider a probabilistic function $f: V \to \mathbb{R}$ and order the vertices by decreasing probability-per-degree

$$\frac{f(v_1)}{d_{v_1}} \geq \frac{f(v_2)}{d_{v_2}} \geq \cdots \geq \frac{f(v_n)}{d_{v_n}}$$

For $i = 1$ to n:

- Let S be the set of the first i nodes in the ordering
- If $\Phi(S)$, $\mathrm{vol}(S)$ are within desired constraints, output S

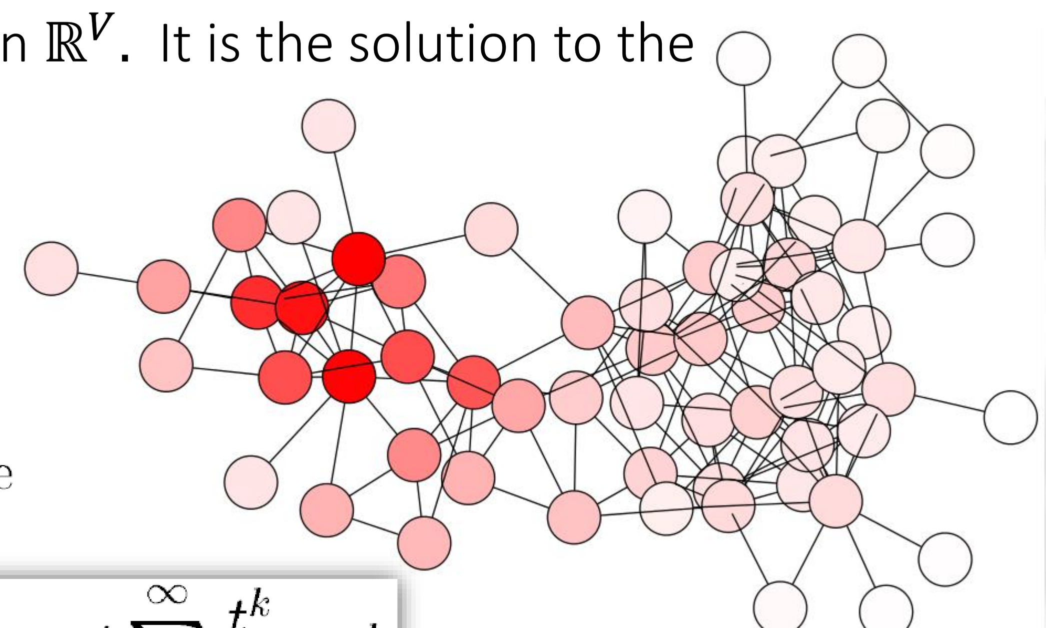


The Heat Kernel of a Graph

The heat kernel is a vector in $\mathbb{R}^V$. It is the solution to the heat equation

$$\frac{\partial}{\partial t}\rho_{t,f} = -\rho_{t,f}(I - P)$$

$$(P)_{uv} = \begin{cases} 1/d_u & \text{if } u \sim v \\ 0 & \text{otherwise} \end{cases}$$



$$\rho_{t,f} = e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!} f P^k$$

When viewed as a function, a single sweep of the heat kernel finds a cluster with Cheeger ratio $O(\sqrt{\phi})$ where ϕ is a constraint.

Approximating the Heat Kernel in Sublinear Time

- Let $f: V \to \mathbb{R}$ be a probabilistic function with all probability on a single vertex, u
- Let X be the random variable that takes on the distribution after k random walk steps starting from u with probability $p_k = e^{-t}\frac{t^k}{k!}$
- Then the expected value of X is exactly $\rho_{t,f}$

ApproxHKPRseed(G, t, u, ϵ)

input: a graph G, $t \in \mathbb{R}^+$, seed vertex $u \in V$, error parameter $0 < \epsilon < 1$.
output: ρ, an ϵ-approximation of $\rho_{t,u}$.

initialize a 0-vector ρ of dimension n, where $n = |V|$
$r \leftarrow \frac{16}{\epsilon^3}\log n$
$K \leftarrow c \cdot \frac{\log(\epsilon^{-1})}{\log\log(\epsilon^{-1})}$ for some choice of contant c

for r iterations **do**
Start
simulate a P random walk from vertex u where k steps are taken with probability $e^{-t}\frac{t^k}{k!}$ and $k \leq K$
let v be the last vertex visited in the walk
$\rho[v] \leftarrow \rho[v] + 1$
End
end for

return $1/r \cdot \rho$

The algorithm returns an ϵ-approximate vector of $\rho_{t,f}$ in time

$$O\left(\frac{\log(\epsilon^{-1})\log n}{\epsilon^3 \log\log(\epsilon^{-1})}\right)$$

Definition 1. *Let G be a graph on n vertices, and let $f: V \to \mathbb{R}$ be a vector over the vertices of G. Let $\rho_{t,f}$ be the heat kernel pagerank vector according to f and t. Then we say that $\nu \in \mathbb{R}^n$ is an* ϵ-approximate vector *of $\rho_{t,f}$ if*

1. for every vertex $v \in V$ in the support of ν,

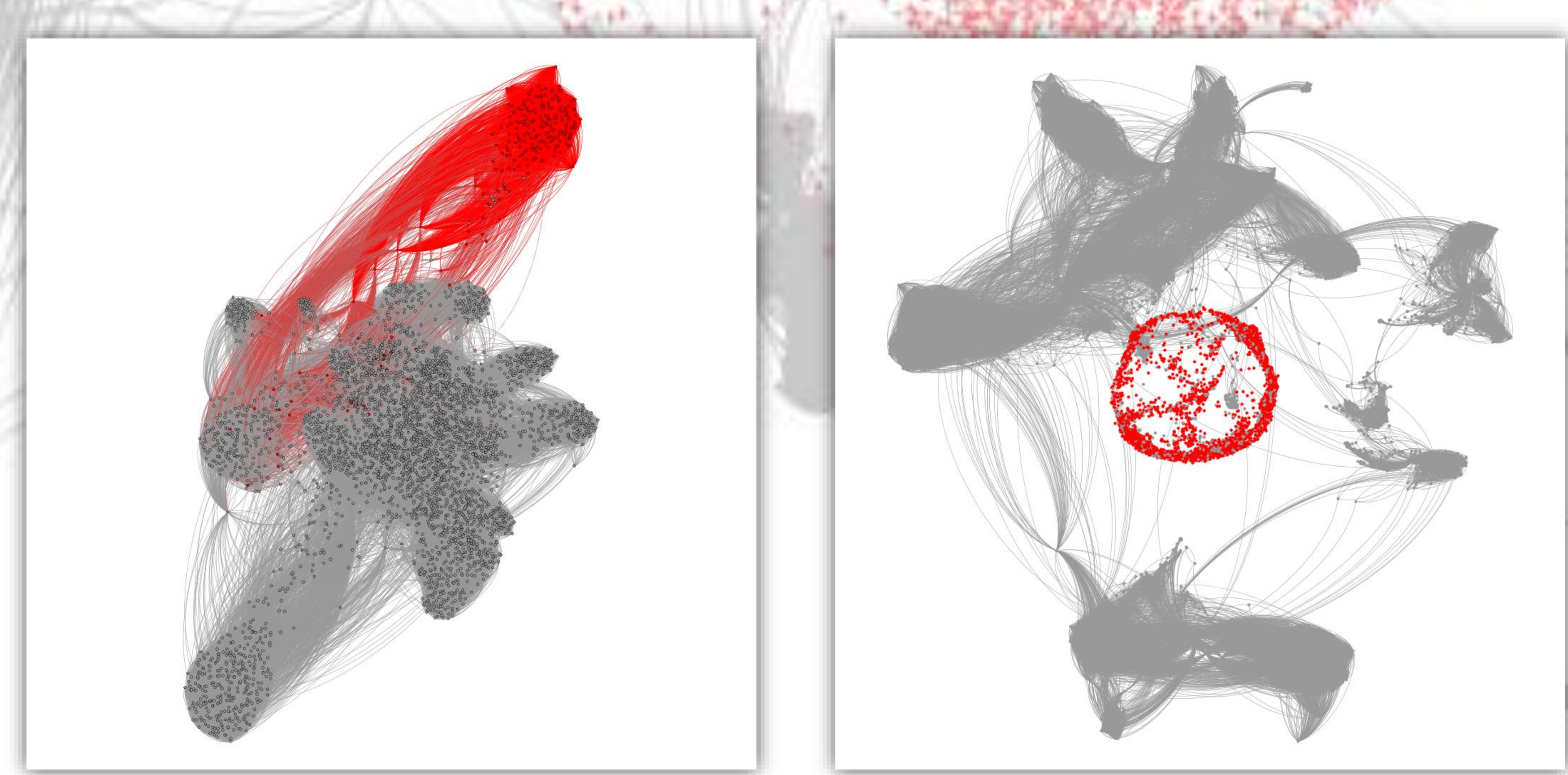
$$(1-\epsilon)\rho_{t,f}(v) - \epsilon \leq \nu(v) \leq (1+\epsilon)\rho_{t,f}(v),$$

2. for every vertex with $\nu(v) = 0$, it must be that $\rho_{t,f}(v) \leq \epsilon$.

Local Clusters with Heat Kernel Sweeps

Network	Size	Constraints	PR	HKPR	ϵHKPR
dolphins	$\lvert V\rvert$ = 62, $\lvert E\rvert$ = 159	ϕ = 0.08, s = 20, vol = 100	ϕ = 0.226, s = 23, vol = 106	ϕ = 0.164, s = 24, vol = 110	ϕ = 0.083, s = 20, vol = 96
polbooks	$\lvert V\rvert$ = 105, $\lvert E\rvert$ = 441	ϕ = 0.05, s = 30, vol = 270	ϕ = 0.08, s = 48, vol = 415	ϕ = 0.246, s = 49, vol = 403	ϕ = 0.052, s = 50, vol = 422
power	$\lvert V\rvert$ = 4941, $\lvert E\rvert$ = 6594	ϕ = 0.05, s = 200, vol = 600	ϕ = 0.375, s = 6, vol = 16	ϕ = 0.003, s = 1564, vol = 4342	ϕ = 0.347, s = 85, vol = 300
facebook	$\lvert V\rvert$ = 4039, $\lvert E\rvert$ = 88234	ϕ = 0.05, s = 200, vol = 2800	ϕ = 0.419, s = 3063, vol = 88140	ϕ = 0.001, s = 1094, vol = 67326	ϕ = 0.057, s = 258, vol = 35266
enron	$\lvert V\rvert$ = 36692, $\lvert E\rvert$ = 183831	ϕ = 0.05, s = 100, vol = 1000	ϕ = 0.488, vol = 183612	- -	ϕ = 0.037, vol = 3579

The outputs of a single sweep algorithm using three different vectors: a Personalized PageRank vector (PR), an exact heat kernel vector (HKPR) and an ϵ-approximate heat kernel vector (ϵHKPR).



Local clusters detected in the Facebook ego network (from the SNAP dataset). The image on the left is a local cluster detected with a sweep of an ϵ–approximate heat kernel vector, while the image on the right is a local cluster detected with a sweep of a Personalized PageRank vector. Both clusters are colored red.

Appeared in proceedings of IWOCA'14

osimpson@ucsd.edu