Supplementary material for: p-Norm Flow Diffusion for Local Graph Clustering

Kimon Fountoulakis [*] **Di Wang** [*] **Shenghao Yang** [*]

A. Missing Proofs of Section 4

Proof of Lemma 1. For $p < \infty$, the optimal routing will never push mass out of a node u unless u's sink is saturated, i.e. $f^*(u, v) > 0$ for u, v only if $(B^T f^* + \Delta)(u) = \deg(u)$. To see this, take the optimal primal solution f^*, and consider the decomposition of f^* into flow paths, i.e., the path that the diffusion solution used to send each unit of mass from its source node to the sink at which it settled. If any node other than the last node on this path has remaining sink capacity, we can truncate the path at that node, and strictly reduce the total cost of the diffusion solution. As each unit of mass is associated with the sink of one node, the total amount of mass $\delta \cdot \mathrm{vol}(S)$ upper-bounds the total volume of the saturated nodes since it takes $\deg(v)$ amount of mass to saturate the node v. This observation proves the first claim.

The dual variable $x^*(u)$ corresponds to the primal constraint $(B^T f^* + \Delta)(u) \le \deg(u)$, and it is easy the check the third claim of the lemma is just complementary slackness. The second claim follows from the first and the third claim. □

Proof of Lemma 3. By Assumption 1 there is least $\Delta(C) \ge 2\mathrm{vol}(C)$ amount of source mass trapped in C at the beginning. Since the sinks of nodes in C can settle $\mathrm{vol}(C)$ amount of mass, the remaining at least $\mathrm{vol}(C)$ amount of excess needs to get out of C using the k cut edges. We focus on the cost of f^* restricted to these edges alone. Since $p > 1$, the cost is the smallest if we distribute the routing load evenly on the k edges, and it is simple to see this incurs cost

$$\left(k \cdot \left(\frac{\mathrm{vol}(C)}{k}\right)^p\right)^{1/p} = \mathrm{vol}(C)/k^{(p-1)/p}.$$

The total cost of f^* must be at least the cost incurred just by routing the excess out of C. □

Proof of Lemma 4.

$$\sum_{e=(u,v)} |x^*(u) - x^*(v)| \cdot \tilde{l}(e)^{q-1} \le \sum_{e=(u,v)} \tilde{l}(e)^q = \sum_{e:l(e)=\tilde{l}(e)} l(e)^q + \sum_{e:l(e)<\tilde{l}(e)} \frac{1}{\mathrm{vol}(C)} \le 1 + \frac{2}{\beta}$$

The second to last equality follows from the fact that our perturbation only increases the lengths to $\frac{1}{\mathrm{vol}(C)^{1/q}}$. The last inequality follows from that we only increase the length of an edge when its original edge is positive, which means at at least one of its endpoints has positive dual variable value. From Assumption 1 that $\delta = 2/\alpha$ we know that the total amount of mass is at most $\frac{2}{\alpha}\mathrm{vol}(S)$. Together with the conditions in Theorem 2 we get $\frac{2}{\alpha}\mathrm{vol}(S) \le \frac{2}{\beta}\mathrm{vol}(C)$. This upper-bounds $\mathrm{vol}_G(\mathrm{supp}(x^*))$ by Lemma 1. Thus, the number of edges with positive $l(e)$ is also at most $\frac{2}{\beta}\mathrm{vol}(C)$. □

Proof of Claim 1. Both claims follow from changing the order of summation to get

$$(\Delta - \vec{\boldsymbol{d}})^T x^* = \int_{h=0}^{\infty} \left(\Delta(S_h) - \mathrm{vol}(S_h)\right) dh,$$

and

$$\sum_{e=(u,v)} |x^*(u) - x^*(v)| \cdot \tilde{l}(e)^{q-1} = \int_{h=0}^{\infty} \sum_{e \in E(S_h, V \setminus S_h)} \tilde{l}(e)^{q-1} dh,$$

and then invoke Lemma 3 and Lemma 4 respectively. To see first claim, pick any node v in the first equation. v contributes $(\Delta(v) - \deg(v)) \cdot x^*(v)$ to the left hand side, and the same amount to the right hand side as v is in the level cuts for all $h \in (0, x^*(v)]$.

For the second claim, pick any edge $e = (u, v)$, the edge will cross the level cuts $E\left(S_h, V \setminus S_h\right)$ for all $h \in (x^*(v), x^*(u)]$ (assuming wlog $x^*(u) \geq x^*(v)$), so the contribution from any edge will be the same to both sides of the equation. □

B. Constrained and penalized problems from Section 5

We are interested in solving the q-norm cut problem

$$\begin{aligned} \max\ & (\Delta - \vec{\boldsymbol{d}})^T x \\ \text{s.t.}\ & \|Bx\|_q \leq 1 \\ & x \geq 0. \end{aligned} \tag{B.1}$$

We turn the constrained formulation into a regularized problem and show their equivalence,

$$\min_{x \geq 0}\ F(x) := (\vec{\boldsymbol{d}} - \Delta)^T x + \tfrac{1}{q}\|Bx\|_q^q. \tag{B.2}$$

Lemma B.1 (Equivalence)**.** *The solution sets of problems* (B.1) *and* (B.1) *are scaled versions of each other.*

Proof. First note that constant non-zero vectors cannot be solutions to any of the problems (B.1) and (B.2). This is because we pick Δ such that $\sum_i \Delta(i) \leq \sum_i \deg(i)$ and an all-zero vector is feasible and has better objective function value than any constant non-zero vector. Let x^* denote a non-constant solution of (B.2). Then x^* satisfies the optimality conditions of (B.2)

- $-(\Delta - \vec{\boldsymbol{d}}) - y + \frac{1}{q}\nabla\|Bx\|_q^q = 0$
- $y(i)x(i) = 0\ \forall i \in V$
- $x, y \geq 0,$

for some optimal dual variables y^*. The optimality conditions of problem (B.1) are

- $-(\Delta - \vec{\boldsymbol{d}}) - y + \lambda\nabla\|Bx\|_q^q = 0$
- $y(i)x(i) = 0\ \forall i \in V$
- $\lambda(1 - \|Bx\|_q^q) = 0$
- $\|Bx\|_q^q \leq 1$
- $x, y \geq 0, \lambda \geq 0,$

where y are the dual variables for the constraint $x \geq 0$, and λ is the dual variable for $\|Bx\|_q^q \leq 1$. Observe that by setting $x := x^*/\|Bx^*\|_q^q$, $\lambda := \frac{1}{q}(\|Bx^*\|_q^q)^{q-1}$ and $y := y^*$, then the triplet x, λ, y satisfies the latter optimality conditions.

Let us now prove the reverse. Note that because we pick Δ such that there exists at least one negative component in $-(\Delta - \vec{\boldsymbol{d}})$, then $\lambda = 0$ can never be an optimal dual variable. Thus $\lambda > 0$. Let $\hat{x}, \hat{\lambda}, \hat{y}$ be a solution to the latter optimality conditions. Observe that $x := (q\hat{\lambda})^{\frac{1}{q-1}}\hat{x}$ and $y := \hat{y}$ satisfy the former optimality conditions. This means that the solution sets of the two problems are scaled versions of each other. □

Lemma B.1 is important because the output of the Sweep Cut procedure is equal for both solutions. This is because the output of Sweep Cut depends only on the ordering of dual variables and not on their magnitude. Therefore, we can use the solution of any of these problems for the local graph clustering problem.

Following Lemma B.1, it is easy to see that if one of (B.1) or (B.2) is unbounded, the other must also be unbounded. Moreover, one can easily verify that the q-norm regularized problem (B.2) is the dual of an equivalent p-norm flow problem

$$\begin{aligned} &\min \tfrac{1}{p}\|f\|_p^p \\ &\text{s.t. } B^T f + \Delta \leq \vec{\boldsymbol{d}} \end{aligned} \tag{B.3}$$

and that strong duality holds, as any $x > 0$ is a Slater point for (B.2). Throughout the discussions in this section, we assume that $|\Delta| \leq \text{vol}(G)$, so (B.3) is feasible, and hence (B.2) is bounded.

Although the two formulations (B.1) and (B.2) are inherently equivalent, the computational advantage of (B.2) permits the use of many off the shelf first order optimization methods, which is crucial for obtaining strongly local algorithms. In subsequent discussions, our goal is prove the running time guarantee of Algorithm 1 in the main text that finds ϵ accurate solution to (B.2). For convenience we have copied the algorithmic steps from the main text to Algorithm B.1.

Algorithm B.1 Coordinate solver for smoothed q-norm cut problem

```
Initialize: $x_0 = 0$
For $k = 0, 1, 2, \ldots,$ do
    Set $S_k = \{i \in V \mid \nabla_i F_\mu(x_k) < 0\}$.
    Pick $i_k \in S_k$ uniformly at random.
    Update $x_{k+1} = x_k - \frac{\mu^{2-q}}{\deg(i_k)} \nabla_{i_k} F_\mu(x_k) e_{i_k}$.
    If $S_k = \emptyset$ then return $x_k$.
```

C. Smoothed objective function from Section 5

In this section we smooth the original objective function $F(x)$ and we prove some important properties of the smoothed problem that will be used to obtain a local running time result for Algorithm B.1.

The objective function $F(x)$ of the regularized q-norm cut problem (B.2) has non-Lipschitz gradient for any $q < 2$, therefore we smooth it by perturbing the q-norm term around zero. Consider the following *globally* smoothed problem

$$\min_{x \geq 0} \ F_\mu(x) := \frac{1}{q} \sum_{(i,j)\in E} ((x(i) - x(j))^2 + \mu^2)^{q/2} - x^T(\Delta - \vec{\boldsymbol{d}}), \tag{C.1}$$

where $\mu > 0$ is a smoothing parameter.

Let x_μ^* denote an optimal solution of (C.1). We define a *locally* smoothed problem

$$\min_{x \geq 0} \ F_\mu^l(x), \tag{C.2}$$

where $F_\mu^l(x)$ is obtained by perturbing the q-norm term around zero on the edges defined by $\text{supp}(Bx_\mu^*) \subseteq E$,

$$F_\mu^l(x) \ := \ \frac{1}{q} \sum_{(i,j)\in \text{supp}(Bx_\mu^*)} ((x(i) - x(j))^2 + \mu^2)^{q/2} \ + \ \frac{1}{q} \sum_{(i,j)\notin \text{supp}(Bx_\mu^*)} |x(i) - x(j)|^q \ - \ x^T(\Delta - \vec{\boldsymbol{d}}).$$

The conceptual construction of $F_\mu^l(x)$ is useful because we can bound, for all x, the maximum gap between the values of $F(x)$ and $F_\mu^l(x)$ by a quantity that depends only on the cardinality of $\text{supp}(Bx_\mu^*)$ and not the dimension of the ambient space. This is critical for establishing the local running time result claimed in Theorem 6 in the main text. Therefore, a key step in our analysis is demonstrating the equivalence between the globally smoothed problem (C.1) and the localized version (C.2). In particular, we prove in Theorem E.1 that the two objectives $F_\mu(x)$ and $F_\mu^l(x)$ share the same unique minimizer,

$$\operatorname*{argmin}_{x \geq 0} F_\mu(x) = x_\mu^* = \operatorname*{argmin}_{x \geq 0} F_\mu^l(x).$$

We start by establishing lower and upper bounds for the locally smoothed objective function $F_\mu^l(x)$ with respect to the original objective function $F(x)$. The following lemma gives an upper bound on the cardinality of $\text{supp}(Bx_\mu^*)$.

Lemma C.1 (Locality)**.** *The number of edges defined by* $\text{supp}(Bx^*_\mu)$ *is bounded by the total amount of initial mass, i.e.,*

$$|\text{supp}(Bx^*_\mu)| < |\Delta|.$$

Proof. By the first-order optimality condition of (C.1), for all $i \in \text{supp}(x^*_\mu)$,

$$\nabla_i F_\mu(x^*_\mu) = \sum_{j\sim i} ((x^*_\mu(i) - x^*_\mu(j))^2 + \mu^2)^{q/2-1}(x^*_\mu(i) - x^*_\mu(j)) - \Delta_i + \deg(i) = 0.$$

Hence,

$$\begin{aligned} |\text{supp}(Bx^*_\mu)| &\le \text{vol}_G(\text{supp}(x^*_\mu)) \\ &= \sum_{i\in\text{supp}(x^*_\mu)} \deg(i) \\ &= \sum_{i\in\text{supp}(x^*_\mu)} \Delta(i) - \sum_{i\in\text{supp}(x^*_\mu)} \sum_{j\sim i} ((x^*_\mu(i) - x^*_\mu(j))^2 + \mu^2)^{q/2-1}(x^*_\mu(i) - x^*_\mu(j)) \\ &= \sum_{i\in\text{supp}(x^*_\mu)} \Delta(i) - \underbrace{\sum_{\substack{i,j\in\text{supp}(x^*_\mu)\\ i\sim j}} ((x^*_\mu(i) - x^*_\mu(j))^2 + \mu^2)^{q/2-1}(x^*_\mu(i) - x^*_\mu(j))}_{=0} \\ &\qquad - \underbrace{\sum_{\substack{i\in\text{supp}(x^*_\mu)\\ j\notin\text{supp}(x^*_\mu)\\ i\sim j}} ((x^*_\mu(i) - x^*_\mu(j))^2 + \mu^2)^{q/2-1}(x^*_\mu(i) - x^*_\mu(j))}_{>0} \\ &\le |\Delta|. \end{aligned}$$

□

Lemma C.2 (Local smooth approximation)**.** *For any x, we have that*

$$F(x) \le F^l_\mu(x) \le F(x) + \tfrac{1}{q}\mu^q|\Delta|.$$

Proof. It suffices to show

$$\|Bx\|_q^q \quad \le \sum_{(i,j)\in\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2+\mu^2)^{q/2} \ + \sum_{(i,j)\notin\text{supp}(Bx^*_\mu)} |x(i)-x(j)|^q \quad \le \quad \|Bx\|_q^q + \mu^q|\Delta|.$$

First of all, it is straightforward to see that

$$\|Bx\|_q^q = \sum_{(i,j)\in E} ((x(i)-x(j))^2)^{q/2} \le \sum_{(i,j)\in\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2+\mu^2)^{q/2} \ + \sum_{(i,j)\notin\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2)^{q/2}.$$

On the other hand, since $q \in (1, 2]$, the function $h : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ given by $h(s) = s^{q/2}$ is concave and $h(0) = 0$. So h is sub-additive. Thus

$$\begin{aligned} &\sum_{(i,j)\in\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2+\mu^2)^{q/2} \ + \sum_{(i,j)\notin\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2)^{q/2} \\ \le &\sum_{(i,j)\in\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2)^{q/2} \ + \sum_{(i,j)\in\text{supp}(Bx^*_\mu)} (\mu^2)^{q/2} \ + \sum_{(i,j)\notin\text{supp}(Bx^*_\mu)} ((x(i)-x(j))^2)^{q/2} \\ = &\sum_{(i,j)\in E} ((x(i)-x(j))^2)^{q/2} + \sum_{(i,j)\in\text{supp}(Bx^*_\mu)} \mu^q \\ = &\|Bx\|_q^q + \mu^q|\text{supp}(Bx^*_\mu)| \\ \le &\|Bx\|_q^q + \mu^q|\Delta|, \end{aligned}$$

where the last inequality is due to Lemma C.1. □

The approximation bound in Lemma C.2 means that, by setting out μ inversely proportional to $|\Delta|$, a minimizer of $F^l_\mu(x)$ can be easily turned into an approximate minimizer of $F(x)$, up to any desired ϵ accuracy. Therefore, in order to obtain an ϵ accurate solution to the regularized q-norm cut problem (B.2), it suffices to minimize $F^l_\mu(x)$. In Theorem E.1 we show that, solving the globally smoothed problem (C.1) effectively minimizes $F^l_\mu(x)$.

In the following discussions we analyze important properties of $F_\mu(x)$ which also extend to $F^l_\mu(x)$ on $\text{supp}(x^*_\mu)$. Lemmas C.3, C.4 and C.5 provide some technical results which will be used to prove Theorem E.1.

Lemma C.3 (Lipschitz continuity)**.** *$\nabla F_\mu(x)$ is Lipschitz continuous with coordinate Lipschitz constant $L_i = \deg(i)\mu^{q-2}$.*

Proof. The first and second order derivatives of $F_\mu(x)$ are

$$\begin{aligned}\nabla F_\mu(x) &= B^T C(x) B x - \Delta + \vec{\boldsymbol{d}},\\ \nabla^2 F_\mu(x) &= B^T \tilde{C}(x) B,\end{aligned}$$

where both $C(x)$ and $\tilde{C}(x)$ are diagonal matrices of size $|E| \times |E|$ whose diagonal entries correspond to edges $(i,j) \in E$,

$$\begin{aligned}[C(x)]_{(i,j),(i,j)} &= \left((x(i)-x(j))^2+\mu^2\right)^{\frac{q}{2}-1},\\ [\tilde{C}(x)]_{(i,j),(i,j)} &= \left((x(j)-x(j))^2+\mu^2\right)^{\frac{q}{2}-2}((q-1)(x(i)-x(j))^2+\mu^2).\end{aligned}$$

In order to obtain coordinate Lipschitz constant L_i of $\nabla_i F_\mu(x)$, we upper bound $[\tilde{C}(x)]_{(i,j),(i,j)}$ by

$$\begin{aligned}[\tilde{C}(x)]_{(i,j),(i,j)} &= \left((x(i)-x(j))^2+\mu^2\right)^{\frac{q}{2}-2}((q-1)(x(i)-x(j))^2+\mu^2)\\ &\le \left((x(i)-x(j))^2+\mu^2\right)^{\frac{q}{2}-1}\frac{(q-1)(x(i)-x(j))^2+\mu^2}{(x(i)-x(j))^2+\mu^2}.\end{aligned}$$

Since $q \in (1,2]$, we get that

$$\frac{(q-1)(x(i)-x(j))^2+\mu^2}{(x(i)-x(j))^2+\mu^2} \le 1,$$

and thus

$$[\tilde{C}(x)]_{(i,j),(i,j)} \le \left((x(i)-x(j))^2+\mu^2\right)^{\frac{q}{2}-1} \le \mu^{q-2}.$$

This means that

$$\nabla^2 F_\mu(x) = B^T \tilde{C}(x) B \preceq \mu^{q-2} B^T B.$$

Therefore

$$L_i = \mu^{q-2} e_i^T B^T B e_i = \deg(i)\mu^{q-2}.$$

□

Lemma C.4 (Monotone gradients)**.** *For any x, the following hold*

- *$h_{ii}(t) := \nabla_i F_\mu(x + te_i)$ is strictly monotonically increasing in t,*
- *$h_{ij}(t) := \nabla_i F_\mu(x + te_j)$ is strictly monotonically decreasing in t if $j \sim i$ and constant if $j \not\sim i$.*

Proof. Expand $\nabla_i F_\mu(x)$ as

$$\nabla_i F_\mu(x) = \sum_{j\sim i}\left((x(i)-x(j))^2+\mu^2\right)^{q/2-1}(x(i)-x(j)) - \Delta(i) + \deg(i).$$

So

$$h_{ii}(t) = \sum_{j\sim i}\left((x(i)+t-x(j))^2+\mu^2\right)^{q/2-1}(x(i)+t-x(j)) - \Delta(i) + \deg(i).$$

Each term in the above sum is a function $g(y) := \left(y^2 + \mu^2\right)^{q/2-1} y$. The derivative of $g(y)$ is

$$g'(y) = (y^2 + \mu^2)^{q/2-2}((q-1)x^2 + \mu^2) > 0,$$

therefore, $g(y)$ is strictly monotonically increasing, and hence so it $h_{ii}(t)$. Similarly, $h_{ij}(t)$ is equal to

$$h_{ij}(t) = \sum_{j \sim i} \left((x(i) - t - x(j))^2 + \mu^2\right)^{q/2-1} (x(i) - t - x(j)) - \Delta(i) + \deg(i).$$

Using the same reasoning as above we get that function $h_{ij}(t)$ is strictly monotonically decreasing if $j \sim i$ and is constant if $j \neq i$ and $j \not\sim i$. □

Lemma C.5 (Non-positive gradients). *For any iteration k and node i in Algorithm B.1, $\nabla_i F_\mu(x_k) \leq 0$, $\forall i \in \text{supp}(x_k)$.*

Proof. Suppose for some iteration $k \geq 0$ we have that $\nabla_i F_\mu(x_k) \leq 0$ for every $i \in \text{supp}(x_k)$, and that $S_k \neq \emptyset$. Say we have picked to update some $i_k \in S_k$. This means $\nabla_{i_k} F_\mu(x_k) < 0$. It follows from Lipschitz continuity (cf. Lemma C.4) that

$$\begin{aligned}|\nabla_{i_k} F_\mu(x_{k+1}) - \nabla_{i_k} F_\mu(x_k)| = &\left|\nabla_{i_k} F_\mu(x_k - \tfrac{\mu^{2-q}}{\deg(i_k)} \nabla_{i_k} F_\mu(x_k) e_{i_k}) - \nabla_{i_k} F_\mu(x_k)\right| \\ \leq &\ L_{i_k} \left|\tfrac{\mu^{2-q}}{\deg(i_k)} \nabla_{i_k} F_\mu(x_k)\right| = |\nabla_{i_k} F_\mu(x_k)|.\end{aligned}$$

Since $\nabla_{i_k} F_\mu(x_k) < 0$, we must have $\nabla_{i_k} F_\mu(x_{k+1}) \leq 0$. Moreover, since

$$x_{k+1}(i_k) = x_k(i_k) - \tfrac{\mu^{2-q}}{\deg(i_k)} \nabla_{i_k} F_\mu(x_k) > x_k(i_k),$$

and $x_{k+1}(j) = x_k(j)$ for all $j \neq i_k$, by Lemma C.4 we know that $\nabla_j F_\mu(x_{k+1}) < \nabla_j F_\mu(x_k) \leq 0$ for every $j \sim i_k$, and $\nabla_j F_\mu(x_{k+1}) = \nabla_j F_\mu(x_k)$ for every $j \not\sim i_k$. The proof is complete by noting that $\nabla F_\mu(x_0) \leq 0$ holds trivially. □

D. Strong convexity from Section 5

In this section we prove that the smoothed objective function $F_\mu(x)$ is strongly convex on a strict subset of $\mathbb{R}^{|V|}$ that contains x^*_μ, and we provide a lower bound on the strong convexity parameter. The results in this section will be used to obtain a local running time result for Algorithm B.1.

Although not entirely necessary, we provide an equivalent, but more intuitive, combinatorial description of Algorithm B.1, as shown in Figure D.1. With this combinatorial interpretation we obtain a trivial upper bound on the gap $|x_k(i) - x_k(j)|$ for any $(i,j) \in E$ at any point in the sequence $\{x_k\}_{k=0}^\infty$ generated by Algorithm B.1. We will use it to demonstrate a strong convexity result for the smoothed objective functions.

The basic setting is as follows. Each node v maintains a height $x(v)$. For any fixed $q \in (1, 2]$, the flow of mass from node u to node v is determined by the relative heights $x(u)$ and $x(v)$ and a smoothing parameter $\mu \geq 0$,

$$\text{flow}_q(u, v; x, \mu) := ((x(u) - x(v))^2 + \mu^2)^{q/2-1}(x(u) - x(v)).$$

Therefore, given x and μ, the mass and the excess at node v are

$$\begin{aligned}m(v; x, \mu) &= \Delta(v) + \sum_{u \sim v} \text{flow}_q(u, v; x, \mu), \\ \text{ex}(v; x, \mu) &= \max\{0, m(v; x, \mu) - \deg(v)\}.\end{aligned}$$

One can easily verify that, under these definitions,

$$-\nabla_i F_\mu(x) = m(i; x, \mu) - \deg(i),$$

and that the steps we layout in Figure D.1 are indeed equivalent to the steps in Algorithm B.1.

Let $\{x_k\}_{k=0}^\infty$ be a sequence generated by Algorithm B.1. Let $k \geq 0$ be arbitrary. By Lemma C.5, we know that $\text{ex}(i; x_k, \mu) \geq 0$ for all nodes i with nonzero height $x_k(i) > 0$. This means that at iteration k when we route a flow around

```
1. Initially, all nodes have height 0, i.e., $x(v) = 0$ for all $v$.
2. While $\text{ex}(v; x, \mu) > 0$ for some node $v$:
    (a) Pick any $v$ where $\text{ex}(v; x, \mu) > 0$ uniformly at random.
    (b) Route the flow around node $v$ by raising it to a new height
            $x(v) = x(v) + \mu^{q-2} \frac{\text{ex}(v; x, \mu)}{\deg(v)}$.
3. Return $x$.
```

Figure D.1. A combinatorial description of Algorithm B.1

node i_k, we never remove more mass from i_k than its current excess $\text{ex}(i_k; x_k, \mu)$, so $m(v; x_k, \mu) \geq 0$ for all node v. On the other hand, since the directions of flow, i.e., $\text{sign}(\text{flow}_q(u, v; x_k, \mu))$, are completely determined by the ordering of current heights x_k, we know that flows cannot cycle. This is because if there is a directed cycle in the induced sub-graph on $\text{supp}(Bx_k)$, where edges are oriented according to the directions of flow, then it means that there must exist a set of heights $\{x_k(i_j)\}$ where $x_k(i_1) < x_k(i_2) < \ldots < x_k(i_1)$, which is not possible. Now, since all nodes have non-negative mass and there is no cycling of flows, the net flow on any edge $(i, j) \in E$ cannot be larger than the total amount of initial mass $|\Delta|$ minus one (one is the lowest possible degree of a node in the underlying connected graph),

$$((x_k(i) - x_k(j))^2 + \mu^2)^{q/2-1} |x_k(i) - x_k(j)| = |\text{flow}_q(i, j; x_k, \mu)| \leq |\Delta| - 1, \quad \text{for all } (i, j) \in E. \tag{D.1}$$

Lemma D.1 (Strong convexity parameter)**.** *Let* $x^*_\mu \in \text{argmin}_{x \geq 0} F_\mu(x)$. *We have that*

$$F_\mu(y) \geq F_\mu(x) + \nabla F_\mu(x)^T (y - x) + \frac{\gamma}{2} \|y - x\|_2^2, \quad \forall\, x, y \in U(x^*_\mu), \tag{D.2}$$

where

$$U(x^*_\mu) := \left\{ x \in \mathbb{R}^{|V|} \;\middle|\; \text{supp}(x) \subseteq \text{supp}(x^*_\mu) \text{ and } (x(i) - x(j))^2 + \mu^2 \leq |\Delta|^{2p-2}\ \forall (i, j) \in E \right\},$$

and the strong convexity parameter γ *satisfies*

$$\gamma > \frac{1}{(p-1)|\Delta|^p}.$$

Proof. Recall the second order derivative of $F_\mu(x)$ is

$$\nabla^2 F_\mu(x) = B^T \tilde{C}(x) B,$$

where $\tilde{C}(x)$ is a diagonal matrix of size $|E| \times |E|$ whose diagonal entries correspond to edges $(i, j) \in E$,

$$\begin{aligned} [\tilde{C}(x)]_{(i,j),(i,j)} &= \left((x(i) - x(j))^2 + \mu^2\right)^{\frac{q}{2}-2} \left((q-1)(x(i) - x(j))^2 + \mu^2\right) \\ &\geq (q-1) \left((x(i) - x(j))^2 + \mu^2\right)^{\frac{q}{2}-1}. \end{aligned}$$

Let $S := \text{supp}(x^*_\mu)$. Let B_S denote the sub-matrix of B with columns chosen such that they correspond to nodes in $\text{supp}(x^*_\mu)$. In order to lower bound γ, it suffices to lower bound the smallest eigenvalue of $B_S^T \tilde{C}(x) B_S$ for all $x \in U(x^*_\mu)$. Note that

$$\min_{x \in U(x^*_\mu)} \lambda_{\min}(B_S^T \tilde{C}(x) B_S) \;\geq\; \lambda_{\min}(B_S^T B_S) \min_{x \in U(x^*_\mu)} \min_{(i,j) \in \text{supp}(Bx^*_\mu)} [\tilde{C}(x)]_{(i,j),(i,j)}.$$

Let $s := (x(i) - x(j))^2 + \mu^2$, then by definition, $s \leq |\Delta|^{2p-2}$ for all $x \in U(x^*_\mu)$ and for all $(i, j) \in \text{supp}(Bx^*_\mu)$. Each diagonal entry in $\tilde{C}(x)$ is lower bounded by a function $h(s) = (q-1)s^{q/2-1}$, and since $h'(s) \leq 0$ for all $s \geq 0$, we know that

$$\min_{x \in U(x^*_\mu)} \min_{(i,j) \in \text{supp}(Bx^*_\mu)} [\tilde{C}(x)]_{(i,j),(i,j)} \;\geq\; \min_{0 \leq s \leq |\Delta|^{2p-2}} h(s) \;=\; (q-1) \left(|\Delta|^{2p-2}\right)^{q/2-1} \;=\; \frac{1}{(p-1)|\Delta|^{p-2}}.$$

Finally, the result follows because

$$\lambda_{\min}(B_S^T B_S) > \frac{1}{|\Delta|^2},$$

which can be easily shown by using the local Cheeger-type result in (Chung, 2007). □

Note that the same argument also apply to $F^l_\mu(x)$ as $\nabla F_\mu(x) = \nabla F^l_\mu(x)$ for all $x \in U(x^*_\mu)$. The strong convexity result implies that the minimizer of the smoothed objective is unique.

E. Convergence, iteration complexity and running time from Section 5

Theorem E.1 (Convergence)**.** *Let x^*_μ denote the optimal solution for the globally smoothed problem* (C.1)*, i.e.,*

$$x^*_\mu \ := \ \operatorname*{argmin}_{x \geq 0} F_\mu(x).$$

The iterates $\{x_k\}_{k=0}^\infty$ generated by Algorithm B.1 converge to x^*_μ. Moreover,

$$x^*_\mu \ = \ \operatorname*{argmin}_{x \geq 0} F^l_\mu(x).$$

Proof. Recall the optimality conditions of (C.1),

(a) Dual feasibility[1]: $x \geq 0$;

(b) Primal feasibility: $\nabla F_\mu(x) \geq 0$.

(c) Complementary slackness (under primal feasibility): $\nabla_i F_\mu(x) \leq 0$ for every $i \in \text{supp}(x)$.

By the definition of index set S_k in Algorithm B.1, the iterates $x_0, x_1, x_2, \ldots$ are monotone, i.e., $0 \leq x_1 \leq x_2 \leq \ldots$, so x_k satisfies item (a) for all k. By Lemma C.5, x_k also satisfies item (c), for all k. We may assume $S_k \neq \emptyset$ for all k, as otherwise Algorithm B.1 terminates with the optimal solution that satisfies item (b). So assume that $S_k \neq \emptyset$ for all k, we argue that the sequence $\{x_k\}_{k=0}^\infty$ converges. Suppose not, then by Lemma C.3,

$$x_{k+1} := x_k - \tfrac{\mu^{2-q}}{\deg(i_k)} \nabla_{i_k} F_\mu(x_k) e_{i_k} = x_k - \tfrac{1}{L_{i_k}} \nabla_{i_k} F_\mu(x_k) e_{i_k},$$

for all k, and hence by coordinate Lipschitz continuity we get

$$F_\mu(x_{k+1}) \leq F_\mu(x_k) - \tfrac{1}{2L_{i_k}} \left(\nabla_{i_k} F_\mu(x_k)\right)^2,$$

for all k. Because the sequence of iterates $\{x_k\}_{k=0}^\infty$ does not converge, the sequence of gradients $\{\nabla_{i_k} F_\mu(x_k)\}_{k=0}^\infty$ must be bounded away from zero. This means that the sequence of function values $\{F_\mu(x_k)\}_{k=0}^\infty$ does not converge, either. But then this implies that $F_\mu(x)$ is unbounded below, contradicting our assumption that an optimal solution to (C.1) exists. Therefore the sequence $\{x_k\}_{k=0}^\infty$ must converge to a limit point $\bar{x}$. Now, because $\nabla F_\mu(x)$ is continuous, and since each x_k satisfies optimality conditions (a) and (c), $\bar{x}$ must also satisfy items (a) and (c). It is easy to see that $\bar{x}$ satisfies item (b), too, because otherwise there must be some x_k where $x_k(i) > \bar{x}(i)$ for some coordinate i, which is not possible. Thus, $\bar{x}$ is a minimizer of $F_\mu(x)$, and by uniqueness we have that $\bar{x} = x^*_\mu$.

Furthermore, for all x such that $\text{supp}(x) \subseteq \text{supp}(x^*_\mu)$, we have that $F^l_\mu(x) = F_\mu(x) - \frac{1}{q}\mu^q(|E| - |\text{supp}(Bx^*_\mu)|)$, that is, $F_\mu(x)$ and $F^l_\mu(x)$ only differ by a constant. Hence $\nabla F^l_\mu(x) = \nabla F_\mu(x)$. But this means that x^*_μ must also satisfy the optimality conditions of the locally smoothed problem (C.2). This means x^*_μ is also the optimal solution of (C.2). □

A direct result of Theorem E.1 is that we can use $F_\mu(x)$ and $F^l_\mu(x)$ interchangeably in the iteration complexity and running time analysis of Algorithm B.1.

[1] We call $x \geq 0$ dual feasibility because $F_\mu(x)$ smoothes the dual problem of p-norm flow diffusion.

Corollary E.2 (Interchangeability). *Let $\{x_k\}_{k=0}^\infty$ be any sequence of iterates generated by Algorithm B.1. Let F_μ^* and F_μ^{l*} be optimal objectives values of* (C.1) *and* (C.2)*, respectively. Let $X := \{x_k\}_{k=0}^\infty \cup \{x_\mu^*\}$. Then for any $x \in X$, we have that*

$$F_\mu(x) - F_\mu^* = F_\mu^l(x) - F_\mu^{l*}.$$

Proof. It immediately follows by noting that $F_\mu(x)$ and $F_\mu^l(x)$ differ by a constant value, i.e., $F_\mu^l(x) = F_\mu(x) - \frac{1}{q}\mu^q(|E| - |\text{supp}(Bx_\mu^*)|)$, for all $x \in X$. □

The convergence of monotonic iterates generated by Algorithm B.1 also guarantees that $F_\mu(x)$ is strongly convex on the iterates.

Corollary E.3 (Strong convexity on iterates). *Let $\{x_k\}_{k=0}^\infty$ be any sequence of iterates generated by Algorithm B.1. If $\mu \leq 1$, then $F_\mu(x)$ is strongly convex on $X := \{x_k\}_{k=0}^\infty \cup \{x_\mu^*\}$.*

Proof. By Lemma D.1, we just need to verify that $X \subseteq U(x_\mu^*)$. Following Theorem E.1, the sequence $\{x_k\}_{k=0}^\infty$ is monotonically increasing and converges to x_μ^*, so $\text{supp}(x_k) \subseteq \text{supp}(x_\mu^*)$ for all k. It suffices to show that

$$(x(i) - x(j))^2 + \mu^2 \leq |\Delta|^{2p-2}, \quad \forall (i,j) \in E, \quad \forall x \in X.$$

We use the flow upper bound (D.1) and the assumption $\mu \leq 1$ to get that, for all $x \in X$ and for all $(i,j) \in E$,

$$\begin{aligned}((x(i) - x(j))^2 + \mu^2)^{(q-1)/2} &= ((x(i) - x(j))^2 + \mu^2)^{q/2-1}((x(i) - x(j))^2 + \mu^2)^{1/2} \\ &\leq ((x(i) - x(j))^2 + \mu^2)^{q/2-1}(|x(i) - x(j)| + \mu) \\ &= ((x(i) - x(j))^2 + \mu^2)^{q/2-1}|x(i) - x(j)| + ((x(i) - x(j))^2 + \mu^2)^{q/2-1}\mu \\ &\leq ((x(i) - x(j))^2 + \mu^2)^{q/2-1}|x(i) - x(j)| + (\mu^2)^{q/2-1}\mu \\ &= ((x(i) - x(j))^2 + \mu^2)^{q/2-1}|x(i) - x(j)| + \mu^{q-1} \\ &\leq |\Delta| - 1 + 1 \\ &= |\Delta|,\end{aligned}$$

and thus we have

$$(x(i) - x(j))^2 + \mu^2 \leq |\Delta|^{2/(q-1)} = |\Delta|^{2p-2},$$

as required. □

Theorem E.4 (Iteration complexity). *Let x_μ^* and F_μ^* denote the optimal solution and optimal value of* (C.1)*, respectively. For any $k \geq 1$, let x_k denote the kth iterate generated by Algorithm B.1. Let γ be the strong convexity parameter as described in Lemma D.1, and let $L_\mu = \max_{i \in \text{supp}(x_\mu^*)} L_i$, where $L_i = \deg(i)\mu^{q-2}$ be the coordinate Lipschitz constants of $F_\mu(x)$. After $K = \mathcal{O}\big(\frac{|\Delta|L_\mu}{\gamma} \log \frac{1}{\epsilon}\big)$ iterations, one has*

$$\mathbb{E}[F_\mu(x_K)] - F_\mu^* \leq \epsilon.$$

Furthermore, let F^* denote the optimal objective value of (B.2)*. If we pick $\mu = \mathcal{O}\big(\big(\frac{\epsilon}{|\Delta|}\big)^{1/q}\big)$, then after*

$$K' = \mathcal{O}\left(\frac{|\Delta|\bar{d}}{\gamma}\left(\frac{|\Delta|}{\epsilon}\right)^{2/q-1} \log \frac{1}{\epsilon}\right)$$

iterations, where $\bar{d} = \max_{i \in \text{supp}(x_\mu^*)} \deg(i)$, one has

$$\mathbb{E}[F(x_{K'})] - F^* \leq \epsilon.$$

Proof. Using coordinate Lipschitz constants given in Lemma C.3 we have that

$$F_\mu(x_{k+1}) = F_\mu(x_k - \tfrac{1}{L_i}\nabla_i F_\mu(x_k)e_i) \leq F_\mu(x_k) - \tfrac{1}{L_i}\nabla_i F_\mu(x_k)^2 + \tfrac{1}{2L_i}\nabla_i F_\mu(x_k)^2 = F_\mu(x_k) - \tfrac{1}{2L_i}\nabla_i F_\mu(x_k)^2.$$

Let $L_{S_k} := \max_{i \in S_k} L_i$, and take conditional expectation

$$\begin{aligned}\mathbb{E}\left[F_\mu(x_{k+1}) \mid x_k\right] \leq \; & F_\mu(x_k) - \tfrac{1}{2}\sum_{i \in V} \tfrac{1}{|S_k|}\tfrac{1}{L_i}\nabla_i F_\mu(x_k)^2 \\ = \; & F_\mu(x_k) - \tfrac{1}{2|S_k|}\sum_{i \in S_k} \tfrac{1}{L_i}\nabla_i F_\mu(x_k)^2 \\ \leq \; & F_\mu(x_k) - \tfrac{1}{2|S_k|L_{S_k}}\|\nabla_{S_k} F_\mu(x_k)\|_2^2 \\ \leq \; & F_\mu(x_k) - \tfrac{1}{2|S_k|L_{S_k}}\|\nabla F_\mu(x_k)\|_2^2.\end{aligned}$$

From strong convexity of $F_\mu(x)$ on $\{x_k\}_{k=0}^\infty \cup \{x_\mu^*\}$ (Lemma D.1 and Corollary E.3) we have that

$$\|\nabla F_\mu(x_k)\|_2^2 \leq 2\gamma(F_\mu(x_k) - F_\mu^*).$$

Therefore, we get

$$\mathbb{E}\left[F_\mu(x_{k+1}) \mid x_k\right] \leq F_\mu(x_k) - \tfrac{\gamma}{|S_k|L_{S_k}}(F_\mu(x_k) - F_\mu^*),$$

and so

$$\mathbb{E}\left[F_\mu(x_{k+1}) \mid x_k\right] - F_\mu^* \leq \left(1 - \tfrac{\gamma}{|S_k|L_{S_k}}\right)(F_\mu(x_k) - F_\mu^*).$$

Note that $S_k \subseteq \text{supp}(x_\mu^*)$, thus $|S_k| \leq |\text{supp}(x_\mu^*)| \leq |\text{supp}(Bx_\mu^*)| \leq |\Delta|$. Let $L_\mu := \max_{i \in \text{supp}(x_\mu^*)} L_i$, then $L_\mu \geq L_{S_k}$. Using these simple inequalities we get

$$\mathbb{E}\left[F_\mu(x_{k+1}) \mid x_k\right] - F_\mu^* \leq \left(1 - \tfrac{\gamma}{|\Delta|L_\mu}\right)(F_\mu(x_k) - F_\mu^*).$$

Take conditional expectations over all $x_{k-1}, x_{k-2}, \ldots, x_1, x_0$ we get

$$\mathbb{E}\left[F_\mu(x_{k+1})\right] - F_\mu^* \leq \left(1 - \tfrac{\gamma}{|\Delta|L_\mu}\right)^k(F_\mu(x_0) - F_\mu^*).$$

Therefore, after $K = \mathcal{O}\big(\frac{|\Delta|L_\mu}{\gamma}\log\frac{1}{\epsilon}\big)$ iterations, one has

$$\mathbb{E}[F_\mu(x_k)] - F_\mu^* \leq \tfrac{\epsilon}{2}.$$

Using Corollary E.2 we get that, after K iterations,

$$\mathbb{E}[F_\mu^l(x_k)] - F_\mu^{l*} = \mathbb{E}[F_\mu(x_k)] - F_\mu^* \leq \tfrac{\epsilon}{2},$$

where F_μ^{l*} is the optimal objective value of (C.2). It then follows from Lemma C.2 that

$$\mathbb{E}[F(x_K)] - F^* \;\leq\; \mathbb{E}[F_\mu^l(x_K)] - F_\mu^{l*} + \tfrac{1}{q}\mu^q|\Delta| \;\leq\; \tfrac{\epsilon}{2} + \tfrac{1}{q}\mu^q|\Delta|.$$

Hence setting $\mu = \mathcal{O}\big(\big(\frac{\epsilon}{|\Delta|}\big)^{1/q}\big)$ gives the required iteration complexity. □

Corollary E.5 (Running time)**.** *If we pick $\mu = \mathcal{O}\big(\big(\frac{\epsilon}{|\Delta|}\big)^{1/q}\big)$, the total running time of Algorithm B.1 to obtain an ϵ accurate solution of* (B.2) *is $\mathcal{O}\big(\frac{|\Delta|\bar{d}^2}{\gamma}\big(\frac{|\Delta|}{\epsilon}\big)^{2/q-1}\log\frac{1}{\epsilon}\big)$.*

Proof. This is straightforward by noticing that at each step in Algorithm B.1, we touch only the nodes j such that $j \sim i_k$, for updating gradient vector to $\nabla F_\mu(x_{k+1})$ and for obtaining S_{k+1}. □

F. Empirical Set-up and Results

F.1. Computing platform and implementation detail

We implemented Algorithm B.1 in Julia[2]. When $p = q = 2$, the objective function of the regularized q-norm cut problem (B.2) has coordinate Lipschitz constants $L_i = \deg(i)$, therefore we can directly solve (B.2) in linear and strongly local running time. As discussed earlier, coordinate methods enjoy a natural combinatorial interpretation as routing mass in the underlying graph. Algorithm F.1 provides a direct specialization of Algorithm B.1 to the case $q = 2$, where we describe the algorithmic steps in the equivalent combinatorial setting as diffusing excess mass in the graph.

Algorithm F.1 Coordinate solver for (B.2) when $q = 2$

```
1. Initially, $x(v) = 0$ and ex$(v) = \max\{\Delta(v) - \deg(v), 0\}$ for all $v \in V$.
2. While ex$(v) > 0$ for some node $v$:
    (a) Pick any $v$ where ex$(v) > 0$.
    (b) Apply push$(v)$.
3. Return $x$.

push$(v)$:
Make the following updates:
1. $x(v) \leftarrow x(v) + \text{ex}(v)/\deg(v)$.
2. ex$(v) \leftarrow 0$.
3. For each node $u \sim v$: ex$(u) \leftarrow$ ex$(u) + \text{ex}(v)/\deg(v)$.
```

For general $p > 2$ and $1 < q < 2$, our implementation adds an additional line-search step. More specifically, instead of using the fixed step-sizes $\mu^{2-q}/\deg(i_k)$ given in Algorithm B.1, we use binary search to find step-sizes α_k such that

$$\nabla_{i_k} F_\mu\big(x_k - \alpha_k \nabla_{i_k} F_\mu(x_k) e_{i_k}\big) = 0.$$

This leads to coordinate minimization steps that can improve practical convergence. Computing the required step-sizes α_k through binary line-search is possible because the partial gradients are monotone (cf. Lemma C.4).

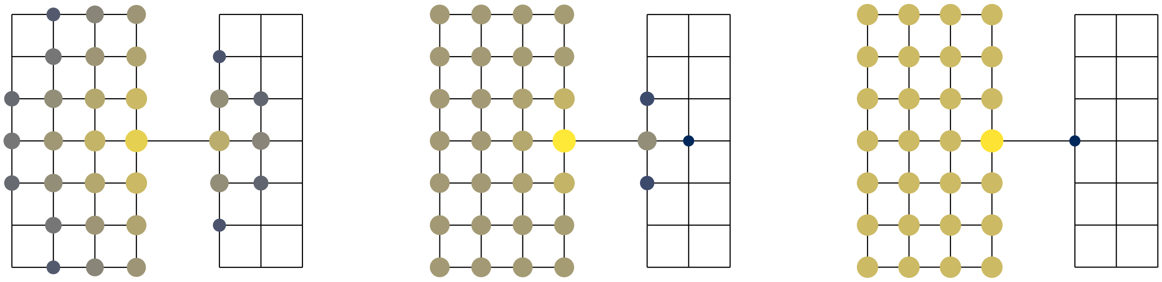
Finally, for efficient implementation that avoids iteratively sampling *with* replacement the indices for coordinate updates, we adopt a sampling *without* replacement approach that is seen in random-permutation cyclic coordinate updates (Lee & Wright, 2018). That is, every time an index set S_k is constructed, we loop over all coordinates in S_k randomly without replacement, before computing a new index set S_{k+1}.

F.2. Diffusion on a dumbbell

The best way to visualize p-norm flow diffusions for various p values is to start the diffusion processes on the same graph and the same set of seed nodes, with equal amount of initial mass. To this end, we run p-norm diffusions for $p \in \{2, 4, 8\}$ on a synthetic small scale "dumbbell" graph obtained by removing edges from a 7×7 grid graph. We pick a single seed node which locates on one side of the "bridge", and set $|\Delta| = 121$. For each p, we plot optimal dual variables in Figure F.1, where we use color intensities and circle sizes to indicate the relative magnitude of dual values, i.e., brighter colors and larger circles size represent higher dual values for a fixed p, and no circle means the corresponding node has zero dual value.

Recall that a dual variable is nonzero only if the corresponding node is saturated (i.e., the mass it holds equals its degree). Observe that 2-norm diffusion leaks a lot of mass to the other side of dumbbell, whereas 4-norm and 8-norm diffusions saturate entire left-hand side, without leaking much mass to the right. The reason that this happens is because $p = 4$ and $p = 8$ put significantly larger penalty on the flow that passes through the "bridge", making it difficult to send mass over to the other side.

[2] Our code is available at https://github.com/s-h-yang/pNormFlowDiffusion.



(a) 2-norm diffusion (b) 4-norm diffusion (c) 8-norm diffusion

Figure F.1. Diffusion on a dumbbell: color intensities and circle sizes are chosen to reflect relative magnitude of optimal dual variables

Note that, if we were to perform the sweep cut procedure described in Section 4, then 2-norm diffusion would fail to recover the "correct" cluster, while both 4-norm and 8-norm diffusions return the entire left-hand side of dumbbell, which is the best possible result in terms of low conductance. Although this example is overly simplistic, it does demonstrate that p-norm flow diffusion with higher p values are more sensitive to bottlenecks when routing mass around a seed node, and, on the other hand, it is possible to overcome such bottleneck even when p is slightly larger than 2, say $p = 4$ and $p = 8$. Indeed, our subsequent experiments show that there is already a significant improvement in the context of local graph clustering, when raising $p = 2$ to $p = 4$.

F.3. Missing plots from LFR synthetic experiments

The following experiment on an LFR synthetic graph with $\mu = 0.3$ demonstrates that, not surprisingly, the more overlap between an initial set of seed nodes and the target cluster, the better result we can get from p-norm flow diffusion. Note that the parameter $\mu = 0.3$ means that 30% of all edges of a node from some target cluster links to the outside of that cluster. Because of such noise level, the goal is not to recover the ground truth exactly, but to obtain a cluster that overlaps well with the target (e.g., has a good F1 score). We have chosen $\mu = 0.3$ because it represents reasonably noisy clusters that are not completely noise. For target clusters that have lower conductance and are connected better internally, the results of p-norm diffusion are already very good even when we start from a singe node (cf. Figure 2).

We randomly pick a set of seed nodes S from some target cluster C in the graph and vary the percentage overlap of S in C, i.e., $|S|/|C|$. Note that this results in different values of α in Theorem 2. Then for each $p \in \{2, 4, 8\}$, we run p-norm flow diffusion and use the sweep cut procedure to find a smallest conductance cluster. Figure F.2 shows the mean and the variance for the conductance and the F1 measure while varying the ratio $|S|/|C|$. As expected, as the percentage overlap $|S|/|C|$ increases, which also means α decreases, we recover clusters with lower conductances and higher F1 scores. Observe that we have a large gap in both conductance and F1 measure when $p = 2$ is increased to $p = 4$, but while there is some gain in raising from $p = 4$ to $p = 8$, but it is very marginal.

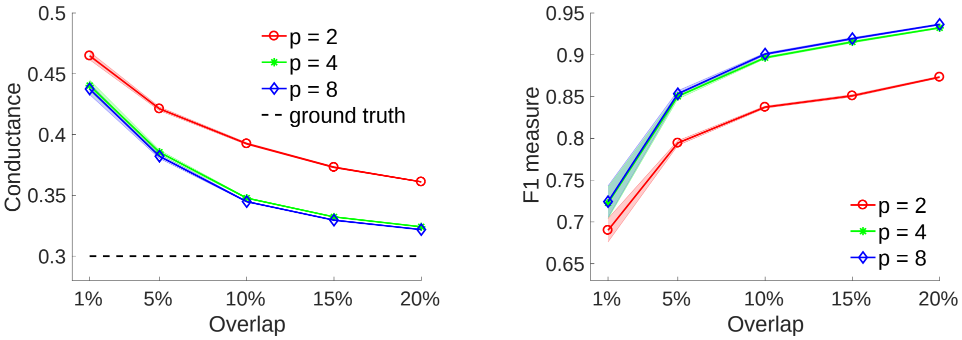


Figure F.2. The conductance and the F1 measure on varying overlaps on an LFR synthetic graph. The black dashed line show ground truth conductance. The bands show the variation over 100 trials.

F.4. Datasets

To compare the performances of different methods on graph clustering tasks, we use four real datasets consisting of both social and biological networks. The graphs that we consider are all unweighted and undirected. Table F.1 shows some basic characteristics of these graphs.

Table F.1. Summary of real-world graphs

dataset	number of nodes	number of edges	description
FB-Johns55	5157	186572	Facebook social network for Johns Hopkins University
Colgate88	3482	155043	Facebook social network for Colgate University
Sfld	232	15570	Pairwise similarities of blasted sequences of proteins
Orkut	3072441	117185083	Large-scale on-line social network

The two Facebook graphs are chosen from the Facebook100 dataset based on their assortativity values in the first column of Table A.2 in (Traud et al., 2012), where the data were first introduced and analyzed. Each of these comes with some features, e.g., gender, dorm, major index and year. We consider a set of nodes with the same feature as a ground truth cluster, and filter the "ground truth" clusters by setting a 0.6 threshold on maximum conductance. We also omit clusters whose volume is larger than one third of the volume of the entire graph, since discovering clusters whose sizes are close to the entire graph is not the purpose of local clustering.

The biology dataset Sfld contains pairwise similarities of blasted sequences of 232 proteins belonging to the amidohydrolase superfamily (Brown et al., 2006). A gold standard is provided describing families within the given superfamily. According to the gold standard the amidrohydrolase superfamily contains 29 families. We consider each family as a ground truth cluster, and filter them by setting a 0.9 threshold on the maximum conductance.

The last dataset Orkut is a free on-line social network where users form friendship each other. It can be downloaded from (Leskovec & Krevl, 2014). This network comes with 5000 ground truth communities, which we filter by setting minimum community size 50, maximum cut conductance 0.5, and minimum ratio between internal connectivity (i.e., the smallest nonzero eigenvalue of the normalized Laplacian of the subgraph defined by the cluster) and cut conductance to 0.85. This resulted in 11 reasonably noisy clusters, having conductances between 0.4 and 0.5.

We include the statistics of all ground truth clusters that we used in the experiments in Table F.2.

F.5. Methods and parameter setting

We compare the performance of p-norm flow diffusion with ℓ_1-regularized PageRank (Fountoulakis et al., 2017) and nonlinear diffusion under power transfer (Ibrahim & Gleich, 2019). Given a starting node v_s, teleportation probability α, and tolerance parameter ρ, the ℓ_1-regularized PageRank is an optimization problem whose optimal solution is an approximate personalized PageRank (APPR) vector (Andersen et al., 2006). This ℓ_1-regularized variational formulation allow us to apply coordinate method and obtain an APPR vector in linear and strongly local running time. The nonlinear diffusion model applies a nonlinear transformation to node values after each matrix-vector products between the Laplacian matrix and the current node values. Since it is demonstrated in (Ibrahim & Gleich, 2019) that a nonlinear transfer defined by the power function $u \mapsto u^{0.5}$ has the best overall performance when compared to other nonlinear functions like $\tanh$ or heat kernel, we only compare p-norm flow diffusion with nonlinear diffusion governed by this power function. We choose $p \in \{2, 4\}$ for p-norm diffusion and demonstrate the advantage of our method even when p is a small constant.

Our goal here is to compare the behaviour of different algorithms under a unified setting, and not to fine tune any particular model. Therefore, we always start p-norm diffusion from a single seed node, and set $|\Delta| = t \cdot \text{vol}(C)$ for some constant factor t and some target cluster C (recall from Assumption 1 this is WLOG). In particular, on the Facebook datasets we set $t = 3$ because the target clusters already have a large volume, and on Orkut dataset we set $t = 5$. On the other hand, because the clusters in Sfld are very noisy, we vary $t \in \{1, 2, \ldots, 10\}$ and pick the cluster with the lowest conductance. In all cases, we make sure the choice of t is such that $|\Delta|$ is less than the volume of the whole graph. For the nonlinear diffusion model with power transfer, we use the same parameter setting as what the authors suggested in (Ibrahim & Gleich, 2019). The ℓ_1-regularized PageRank is the only linear model here for comparison, so we allow it to use ground truth information to choose the teleportation parameter α giving the best conductance result. We tune α by picking $\alpha \in \{\lambda/8, \lambda/4, \lambda/2, \lambda, 2\lambda\}$, where λ is the smallest nonzero eigenvalue of the normalized Laplacian for the subgraph that corresponds to the target cluster.

Table F.2. Filtered "ground truth" clusters for real-world graphs

dataset	feature	volume	nodes	conductance
FB-Johns55	year 2006	81893	845	0.54
	year 2007	89021	842	0.49
	year 2008	82934	926	0.39
	year 2009	33059	910	0.21
	major index 217	10697	201	0.26
Colgate88	year 2004	14888	230	0.54
	year 2005	50643	501	0.50
	year 2006	62065	557	0.48
	year 2007	68382	589	0.41
	year 2008	62430	641	0.29
	year 2009	35379	641	0.11
Sfld	urease	31646	100	0.42
	AMP	3186	28	0.53
	phosphotriesterase	381	7	0.78
	adenosine	1062	10	0.83
	dihydroorotase3	494	7	0.83
	dihydroorotase2	3119	13	0.90
orkut	A	49767	383	0.42
	B	31912	202	0.45
	C	16022	141	0.45
	D	11698	113	0.46
	E	26248	194	0.47
	F	4617	64	0.47
	G	13786	128	0.47
	H	14109	107	0.48
	I	18652	195	0.49
	J	41612	318	0.50
	K	20204	223	0.50

Since the support size of APPR vector is bounded by $1/\rho$ (Fountoulakis et al., 2017), and the support size of dual variables for p-norm diffusion is bounded by $|\Delta|$, for comparison purposes, we set the tolerance parameter ρ for ℓ_1-regularized PageRank so that $\rho = 1/|\Delta|$.

F.6. Additional discussion

The clustering results are already listed in Table 1 in the main text. We comment on some observations from the results of the biological dataset Sfld. The six ground truth clusters in this dataset are very noisy, having median conductance around 0.8. Hence it is highly likely that a ground truth cluster is contained in some larger sets (but not one of the 29 true families) that have much lower conductances. This partially explains why nonlinear power diffusion returns low conductance clusters but has poor recovery performance in terms of F1 measures. Note that the nonlinear diffusion model is the only global method that we compare with. Both p-norm flow diffusions and ℓ_1-regularized PageRank are local methods, and they take advantage of locality to find local clusters that align well with the ground truth. Therefore, the results for the Sfld dataset demonstrate the advantage of local methods at recovering relatively small ground truth clusters.

References

Andersen, R., Chung, F., and Lang, K. Local graph partitioning using pagerank vectors. *FOCS '06 Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pp. 475–486, 2006.

Brown, S. D., Gerlt, J. A., Seffernick, J. L., and Babbitt, P. C. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome biology*, 7(1):R8, 2006.

Chung, F. Random walks and local cuts in graphs. *Linear Algebra and its Applications*, 423(1):22–32, 2007. ISSN 0024-3795. Special Issue devoted to papers presented at the Aveiro Workshop on Graph Spectra.

Fountoulakis, K., Roosta-Khorasani, F., Shun, J., Cheng, X., and Mahoney, M. W. Variational perspective on local graph clustering. *Mathematical Programming*, 174:553–573, 2017.

Ibrahim, R. and Gleich, D. Nonlinear diffusion for community detection and semi-supervised learning. *WWW'19: The World Wide Web Conference*, pp. 739–750, 2019.

Lee, C. and Wright, S. J. Random permutations fix a worst case for cyclic coordinate descent. *IMA Journal of Numerical Analysis*, 39(3):1246–1275, July 2018. ISSN 0272-4979.

Leskovec, J. and Krevl, A. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

Traud, A. L., Mucha, P. J., and Porter, M. A. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.