P-NORM FLOW DIFFUSION FOR LOCAL GRAPH CLUSTERING

Shenghao Yang, Di Wang, Kimon Fountoulakis

SIAM Workshop on Network Science 2020
July 9–10 · Toronto

Summary

We bridge numerical and combinatorial approaches and propose a family of convex optimization formulations for local graph clustering based on the idea of diffusion with p-norm network flow for $p \in (1, \infty)$. We characterize the optimal solutions for these problems and show their usefulness in finding low conductance cuts around input seed set. We prove quadratic approximation of conductance in the case of $p = 2$ similar to the Cheeger-type bounds of spectral methods, constant factor approximation when $p \to \infty$ similar to max-flow based methods, and a smooth transition for general p values in between. We show that the proposed problem can be solved in strongly local running time for $p \geq 2$ and we conduct empirical evaluations on both synthetic and real-world networks to illustrate our approach compares favorably with existing methods.

Introduction

Exploiting small-scale local structures inside massive graphs is of central importance in many areas of applied mathematics and machine learning, e.g. community detection in networks [13, 6, 5] and PageRank-based spectral ranking of webs [10, 4]. In this work, we consider local graph clustering as the task of finding a community-like cluster around a given set of seed nodes, where nodes in the cluster are densely connected to each other while relatively isolated to the rest of the graph.

An algorithm is called *strongly local* if it runs in time proportional to the size of the output cluster rather than the size of the whole graph. Strongly local algorithms for local graph clustering are predominantly based on the idea of diffusion, which is the generic process of spreading mass among vertices by sending mass along edges. Historically, the most popular diffusion methods are spectral methods [7, 8, 2], due to the ease of implementation, efficient running time and good performance in many contexts. However, it is also known in theory and in practice that spectral methods can spread mass too aggressively and they might fail to find the right cluster when structural heterogeneities exist. Recent diffusion methods are based on the combinatorial idea of max flow exploiting the canonical duality between flow and cut [12, 3, 9]. These methods offer improved theoretical guarantees in terms of locating local cuts, and have been shown to outperform spectral methods in practice for pathological cases. However, combinatorial methods are generally accepted to be more difficult to understand and implement due to the more complicated underlying dynamics.

In this work, we propose and study a family of primal and dual convex optimization problems for local graph clustering. We call the primal problem p-norm flow diffusion, parameterized by the L_p-norm used in the objective function, and the problem defines a natural diffusion model on graphs using network flow. We call the dual problem as the q-norm local cut problem where q defines the dual norm of p (i.e. $1/p + 1/q = 1$). The optimal solution to the q-norm local cut problem can be used to find good local clusters with provable guarantees. We note that almost all previous diffusion methods are defined with the dynamics of the underlying diffusion procedure, i.e. the step-by-step rules of how to send mass, and the analysis of these methods is based on the behaviors of the algorithm. On the other hand, our work starts with a clear optimization objective, and analyze the properties of the optimal solution independent from what method is used to solve the problem. This top-down approach is distinct in theory, and is also very valuable in practice since the de-coupling of objective and algorithm gives the users the freedom at implementation to choose the most suitable solver based on availability of infrastructure and code-base.

Diffusion as Optimization

We consider a diffusion on a graph $G = (V, E)$ as the task of spreading mass from a small set of nodes to a larger set of nodes. Given the signed incidence matrix B and two functions $\Delta, T : V \to \mathbb{R}_{\geq 0}$, we propose the following pair of convex optimization problems, which are the p-norm flow diffusion

$$\begin{aligned} &\text{minimize } \|f\|_p \\ &\text{s.t. } \; B^T f + \Delta \leq T \end{aligned} \tag{1}$$

and its dual formulation with q such that $1/q + 1/p = 1$

$$\begin{aligned} \text{maximize } & (\Delta - T)^T x \\ \text{s.t. } & \|Bx\|_q \leq 1 \\ & x \geq 0 \end{aligned} \tag{2}$$

The function Δ will specify the amount of initial mass starting at each node, and the function T will give the sink capacity of each node. We denote the *density* (of mass) at a node v as the ratio of the amount of mass at v over its degree. Naturally in a diffusion, we start with Δ having small support and high density, and the goal is to reach a state with bounded density enforced by the sink function. This gives a clean physical interpretation where paint (i.e. mass) spills from the source nodes and spreads over the graph and there is a sink at each node where up to a certain amount of paint can settle.

The solution to the dual problem $x^* \in \mathbb{R}^{|V|}_{\geq 0}$ gives an embedding of the nodes on the (non-negative) real line. This embedding is what we actually compute in the context of local clustering. Our first main result is a novel theoretical analysis that gives worst case guarantees on the cluster obtained from x^* using a standard sweep cut procedure. In particular, suppose there exists a cluster B with conductance $\phi(B)$, and we are given a set of seed nodes that overlaps reasonably with B. Then x^* can be used to find a cluster A with conductance at most $\mathcal{O}(\phi(B)^{1/q})$. For $p = 2$, this result resembles the Cheeger-type quadratic guarantees that are well-known in spectral-based local graph clustering literature [11, 1]. When $p \to \infty$, our conductance guarantee approaches a constant factor approximation similar to combinatorial diffusions, while achieving a smooth transition for general p values in between. We observe in practice that our optimization formulation can achieve the best of both worlds in terms of speed and robustness to noise when p lies in the range of small constants, e.g. $p = 4$.

Strongly Local Algorithm

We design a randomized coordinate descent variant that obtains an ϵ accurate solution of (2) for $p \geq 2$ in strongly local running time $\mathcal{O}\big(\frac{\hat{d}^2}{\gamma}\big(\frac{|\Delta|}{\epsilon}\big)^{2/q-1}\log\frac{1}{\epsilon}\big)$, where $\hat{d}$ is the maximum degree of nodes defined on the support of x^*, γ is a strong convexity parameter bounded away from zero, and $|\Delta|$ is the amount of initial mass. Our analysis illustrates a natural trade-off as a function of p between robustness to noise and the running time for solving (2). In particular, for $p = 2$ the diffusion problem can be solved in time linear in the size of the local cluster, but may have quadratic approximation error $\mathcal{O}(\sqrt{\phi(B)})$. On the other hand, the approximation error guarantee improves when p increases, but it also takes longer to converge to the optimal solution. We demonstrate empirically that the regime of p being small constants offer the best trade-offs in general.

References

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

[2] F. Chung. The heat kernel as the PageRank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735–19740, 2007.

[3] K. Fountoulakis, D. F. Gleich, and M. W. Mahoney. An optimization approach to locally-biased graph algorithms. *Proceedings of the IEEE*, 105(2):256–272, 2017.

[4] D. F. Gleich. PageRank beyond the web. *SIAM Review*, 57(3):321–363, 2015.

[5] L. G. S. Jeub, P. Balachandran, M. A. Porter, P. J. Mucha, and M. W. Mahoney. Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. *Physical Review E*, 91:012821, 2015.

[6] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[7] L. Lovász and M. Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 346–354, 1990.

[8] L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Struct. Algorithms*, 4(4):359–412, 1993.

[9] L. Orecchia and Z. A. Zhu. Flow-based algorithms for local graph clustering. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1267–1286, 2014.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[11] D. A. Spielman and S. H. Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Scientific Computing*, 42(1):1–26, 2013.

[12] D. Wang, K. Fountoulakis, M. Henzinger, M. W. Mahoney, and S. Rao. Capacity releasing diffusion for speed and locality. *Proceedings of the 34th International Conference on Machine Learning*, 70:3607–2017, 2017.

[13] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In *SDM '05: Proceedings of the 5th SIAM International Conference on Data Mining*, pages 76–84, 2005.