Alternative Symmetrizations of Hitting Times in Graphs

Timothy Chu
Carnegie Mellon University
tzchu@andrew.cmu.edu

Gary Miller
Carnegie Mellon University
glmiller@cs.cmu.edu

November 8, 2018

1 Problem with this approach

The hitting time metric I described does not have Rayleigh property, which seems very important. If A and B are connected by many paths, and A and C are not, but then I decide to hang a ton of other paths onto B leading away from A, then A and B become less connected.

Normalizing by degree doesn't necessarily help (or does it?). Does this concept of hitting times work for d-regular graphs?

Maybe this works fine for d regular graphs. That is, hitting times to a destination may measure how many paths there are to that destination. But even then, I think I can build counterexamples. For instance, if I hang one paths onto B that lead to big chunks of the graph far from A, that's makes B far from A. Meanwhile, if I hung one path leading to a small tightly connected cluster away from A, that makes B closer to A. Is that a bad property, or maybe that's good?

But for non d-regular graphs, I'm at a loss. It seems like I'll want to do something other than hitting time, or even degree-normalized hitting time.

Abstract

Hitting times of random walks are widely studied asymmetric measures in combinatorics and data mining. The most common symmetrization of hitting times are known as commute times, proportional to the effective resistance. It is known that commute times have properties that are considered undesirable in large scale machine learning. For example, a ball of radius r in the commute-time metric can be disconnected.

In this paper, we propose an alternative symmetrization of hitting times with the property that any ball in this symmetrization is identical to some ball in the original hitting time measure. This ensures that balls form connected components, and captures the desirable property that points A and C are considered closer than A and B if A and C are in the same dense cluster. This symmterized notion of hitting times has desirable properties when computing Voronoi cells, including the property that any Voronoi diagram with two sites is connected.

It is known that for large-scale random geometric graphs, the hitting times capture only local properties of the graph: $H(a,b) \sim \frac{1}{d_b}$ as the number of vertices get large. Critically, our symmetrization of hitting times avoids this problem, and captures the global structure of the graph. Our analysis and construction is not restricted to only hitting time measures, and we will present criteria for when any measure can be symmetrized in this fashion.

Expected hitting times are widely used measures in combinatorics and data mining [CLT08]. The work of Von Luxburg et. al. [VLRH14] showed that hitting times are largely meaningless measures on large-scale, random geometric graphs such as the k-NN graphs, ε-graphs, and Gaussian similarity graphs. They proved that hitting time from a to b captures only local information of a graph, as it converges to $\frac{1}{d_b}$.

Since hitting times are asymmetric, often it can be more useful to deal with a symmetric measure. Commute times are the most popular symmetrization of hitting times, defined as the expected time to travel from vertex a to b and back again. Commute times are known to be scaled versions of effective resistance, and also capture few global properties of a random geometric graph: commute times converge to $\frac{1}{d_b}+\frac{1}{d_a}$ in ε-graphs and k-NN graphs [VLRH14]. Additionally, balls of radius r measured in commute time are not necessarily connected, which makes commute times an unsuitable similarity measure in some machine learning settings.

The key result of this paper is to generate a new symmetrization of hitting times that captures the global properties of the graph, avoiding Von Luxburg's problem. Namely, we build a symmetric function f such that $f(a,b) > f(c,b)$ if and only if the hitting time from a to b is larger than the hitting time from c to b. This way, f captures the *partial ordering* of hitting times given a fixed destination, rather than concern itself with the actual value of the hitting time. In this paper, we show:

1. Function f exists,

2. f avoids von Luxburg's general problem with hitting time, and computes global properties of underlying graph G

3. Balls centered at point p with radius r (where this ball is defined as $\{x \in V(G) : f(x,p) \leq r\}$) are connected, and

4. The Voronoi cells of f are the same for any function f respecting the partial ordering of hitting times, and show that when there are two Voronoi sites then each Voronoi cell is connected.

The last property may make our symmetrized hitting time f suitable for bisecting k-means, as Voronoi cell computation is an important primitive in k-means clustering. These Voronoi cells can be computed without explicit access to f. This symmetric hitting time f captures the additional desirable feature that two points A and B are considered closer if there are many paths from A to B.

TIMOTHY: More things: Explicitly construct f rather than implicitly TIMOTHY: Maybe explicit f will be nice TIMOTHY: Random walks with resets? Other kind sof random walks? TIMOTHY: Half live?

1.1 Preliminaries

Definition 1.1. *The hitting time $H_G(a,b)$ from vertex a to b in graph G is the expected time it takes for a random walk from a to hit b. It can be alternatively defined as follows:*

$$H_G(a,b) \overset{\text{def}}{=} (\chi_a - \chi_b)^T L_G^{\dagger}(\overline{d} - \chi_b).$$

Here, χ_a is the indicator vector for vertex a, and $\overline{d} \in \mathbb{R}^{|V(G)|}$ is the vector with $\overline{d}_i$ equal to the degree of vertex i.

2 Symmetrization

Theorem 2.1. *There exists a function f satisfying:*

1. *f respects the ordering of hitting times to a single destination: $H(B,A) < H(C,A) \Rightarrow f(B,A) < f(C,A)$.*
2. *$H(B,A) = H(C,A) \Rightarrow f(B,A) = f(C,A)$.*
3. *$f(B,A) = f(A,B)$ (symmetry).*

In other words, there exists a symmetric function f that respects the poset formed by hitting times to a single destination.

Our proof is non-constructive. Proof TBA.

3 Voronoi diagrams with Hitting Times

Let $A_1, A_2, \ldots A_k$ be a collection of Voronoi sites, and suppose we want to build a Voronoi diagram with respect to f. In this section, we claim that:

The Voronoi diagram generated by f (for any function f respecting the hitting time ordering) can be generated through H by assigning point p to site:

$$\arg\min_i H(A_i, p).$$

Theorem 3.1. *Any Voronoi diagram generated through any f respecting the hitting time ordering, must be the Voronoi diagram listed above.*

Proof. You would want point p to b assigned to site $\arg\min_i f(A_i,p)$, which is the same as $\arg\min_i H(A_i,p)$ by definition of f. □

Thus, our Voronoi cells are well defined.

Theorem 3.2. *If there are only two Voronoi sites, then the Voronoi cells are connected.*

Theorem 3.2 has counterexamples when there are three site. **TIMOTHY: Draw it in**.

Lemma 3.3.

$$H(A,p) - H(B,p) = \frac{1}{d_p} \sum_{q \in N(p)} H(A,q) - H(B,q)$$

for all $p \neq A, B$.

Proof. (of Theorem 3.2, using Lemma 3.3)

Suppose there is a set of vertices disconnected from A_1, that are closer to A_1 than any other vertex. Lemma 3.3 tells us that $H(A_1, p) - H(A_2, p) < 0$, so take the minimum p on this set of disconnected vertices. Since it's minimal, all its neighbors must have the same value of that difference, and all of those neighbors must too, and so forth. However, since the graph is connected, this is impossible since the value of this difference must be positive at A_2. □

Proof. (of Lemma 3.3)

Recall

$$\begin{aligned} H(A,p) - H(B,p) &= (\chi_A - \chi_B)^T L^\dagger(\overline{d} - 2m\chi_p) \\ H(A,q) - H(B,q) &= (\chi_A - \chi_B)^T L^\dagger(\overline{d} - 2m\chi_q) \end{aligned}$$

So

$$\frac{1}{d_p} \sum_{q \in N(p)} H(A,q) - H(B,q) = (\chi_A - \chi_B)^T L^\dagger(\overline{d} - 2m \sum_{q \in N(p)} \frac{1}{d_p}\chi_q). \tag{1}$$

Now,

$$\frac{1}{d_p} \sum_{q \in N(p)} \chi_q = AD^{-1}\chi_p.$$

Therefore, we can simplify Equation 1 to:

$$\frac{1}{d_p} \sum_{q \in N(p)} H(A,q) - H(B,q) \tag{2}$$

$$= (\chi_A - \chi_B)^T L^\dagger(\overline{d} - 2m \sum_{q \in N(p)} \frac{1}{d_p}\chi_q). \tag{3}$$

$$= (\chi_A - \chi_B)^T L^\dagger \overline{d} - 2m(\chi_A - \chi_B)^T (L^\dagger A D^{-1}\chi_p) \tag{4}$$

However, note

$$L^\dagger(D - A)D^{-1} = I_{\perp\mathbf{1}} D^{-1},$$

or

$$L^\dagger A D^{-1} = L^\dagger - I_{\perp\mathbf{1}} D^{-1},$$

so Equation 4 equals

$$(\chi_A - \chi_B)^T L^\dagger \overline{d} - 2m(\chi_A - \chi_B)^T (L^\dagger - I_{\perp\mathbf{1}} D^{-1})\chi_p$$

or

$$(\chi_A - \chi_B)^T L^\dagger \overline{d} - 2m(\chi_A - \chi_B)^T L^\dagger \chi_p + 2m(\chi_A - \chi_B)^T I_{\perp\mathbf{1}} D^{-1})\chi_p$$

but note that the last term in the summand is 0, since $(\chi_A - \chi_B)^T I_{\perp\mathbf{1}} = (\chi_a - \chi_B)^T$, which is orthogonal to $D^{-1}\chi_p$, so

$$\frac{1}{d_p} \sum_{q \in N(p)} H(A,q) - H(B,q) = H(A,p) - H(B,p)$$

as desired. □

References

[CLT08] Mo Chen, Jianzhuang Liu, and Xiaoou Tang. Clustering via random walk hitting time on directed graphs. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 616–621. AAAI Press, 2008.

[VLRH14] Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *J. Mach. Learn. Res.*, 15(1):1751–1798, January 2014.