From random walks to distances on unweighted graphs

Tatsunori B. Hashimoto[1], Yi Sun[2], Tommi S. Jaakkola[1]

[1]MIT CSAIL [2]MIT Mathematics

Introduction and results

Motivation: What is a natural measure of latent distance on a graph?

Our answer: The Laplace transform of the hitting time.

Our contributions:

1. A new technique for analyzing random walks on graphs through continuum limits.
2. Extension of von Luxburg-Radl-Hein 2014: the expected hitting time is a degenerate measure of distance.
3. A principled estimator — **Laplace transformed hitting time**:
   - Consistent for recovering a latent distance metric
   - Respects underlying density and cluster structure
   - Robust to model misspecification.

The spatial graph model

Summary: We define a random network model with edges determined by a latent metric.

The spatial graph model depends on the following latent quantities:

- $p(x)$: latent probability density in compact connected smooth domain $D \subset \mathbb{R}^d$;
- $\mathcal{X}_n = \{x_1, \dots, x_n\}$: coordinate points drawn i.i.d. from $p(x)$;
- $\varepsilon_n(x_i) : \mathcal{X}_n \to \mathbb{R}_{>0}$: local scale function (may depend on $\mathcal{X}_n$);
- $h : \mathbb{R}_{\geq 0} \to [0,1]$: connectivity kernel with $h(x) = 0$ for $x > 1$, $h(1) > 0$, and h left-continuous at 1.

Definition 1: Spatial graph

The **spatial graph** G_n corresponding to ε_n and h is the random graph with:

- vertex set $\mathcal{X}_n$;
- a directed edge from x_i to x_j with probability $h(|x_i - x_j|\varepsilon_n(x_i)^{-1})$.

Examples:

1. **Directed k-NN:** $h(x) = 1_{x \in [0,1]}$; $\varepsilon_n(x)$ = distance to k^{th} nearest neighbor.
2. **(Truncated) Gaussian kernel:** $h(x) = e^{-\frac{x^2}{\sigma^2}} 1_{x \in [0,1]}$; $\varepsilon_n(x)$ = bandwidth.

Continuum limits of random walks on the graph

Summary: We characterize the continuum limit of the simple random walk on a spatial graph as an Itô drift-diffusion process.

As the graph grows large ($n \to \infty$), suppose there exist scaling constants g_n and a deterministic continuous function $\overline{\varepsilon} : D \to \mathbb{R}_{>0}$ so that

$$n \to \infty, \qquad g_n \to 0, \qquad g_n n^{\frac{1}{d+2}} \log(n)^{-\frac{1}{d+2}} \to \infty, \qquad \varepsilon_n(x) g_n^{-1} \to \overline{\varepsilon}(x) \text{ for } x \in \mathcal{X}_n.$$

Let X_t^n be the **simple random walk** on the spatial graph G_n. A key observation:

For ε_n small and t large, the random variable $X_t^n - X_0^n$ is the sum of many small independent (but not identically distributed) increments.

Theorem 1: Continuum limit of the walk (Hashimoto-Sun-Jaakkola 2015)

The simple random walk X_t^n converges uniformly in Skorokhod space $D([0,\infty), \overline{D})$ after a time scaling $\hat{t} = t g_n^2$ to the Itô process $Y_{\hat{t}}$ valued in the space of continuous functions $C([0,\infty), \overline{D})$ with reflecting boundary conditions on D defined by

$$dY_{\hat{t}} = \nabla \log(p(Y_{\hat{t}}))\overline{\varepsilon}(Y_{\hat{t}})^2/3 d\hat{t} + \overline{\varepsilon}(Y_{\hat{t}})/\sqrt{3} dW_{\hat{t}}.$$

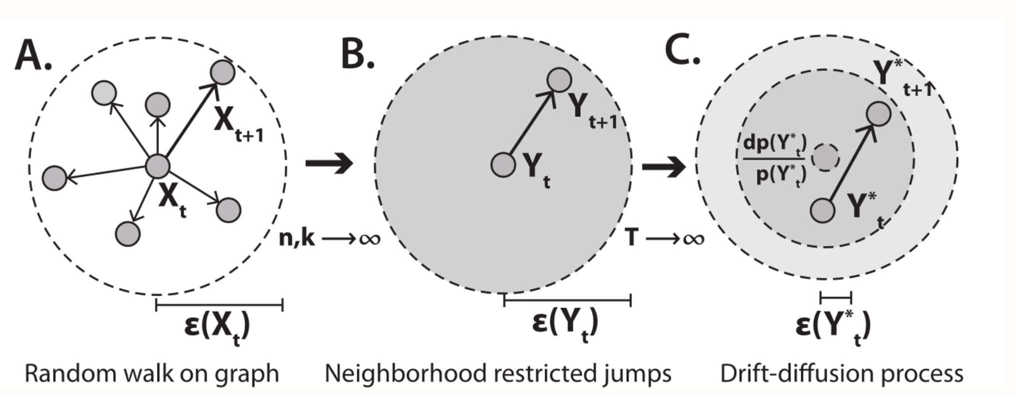

Figure 1 : An heuristic explanation of the scaling limit of Theorem 1.

Key idea: Attributes of stochastic processes (stationary distribution, hitting time) are defined for both $Y_{\hat{t}}$ and X_t^n. Often Theorem 1 implies **the continuous attribute is a rescaled limit of the discrete one.**

Application: Degeneracy of expected hitting time

Summary: Although expected hitting time between two vertices is a commonly used measure of distance, we show it is degenerate for spatial graphs and recovers only the inverse of the stationary density.

Notation: $T_{x_j,n}^{x_i}$ = the hitting time of X_t^n started at x_i to x_j.

A common measure of distance is:

$$\text{Expected hitting time} = \mathbb{E}[T_{x_j,n}^{x_i}].$$

We show that this measure is unrelated to distance.

Theorem 2: Degeneracy of expected hitting time

For $d \geq 2$ and any i, j, we have

$$\frac{\mathbb{E}[T_{x_j,n}^{x_i}]}{n} \xrightarrow{a.s.} \frac{1}{\hat{\pi}(x_j)}.$$

Note: Generalizes surprising result of von Luxburg-Radl-Hein 2014 in the undirected case.

Proof intuition:

1. Compare the hitting time of the simple random walk to its Itô process equivalent;
2. Compare the Itô process with Brownian motion and show by transience that it is unlikely to hit quickly;
3. Conditioned on slow hitting, the random walk mixes before it hits, yielding the stationary distribution.

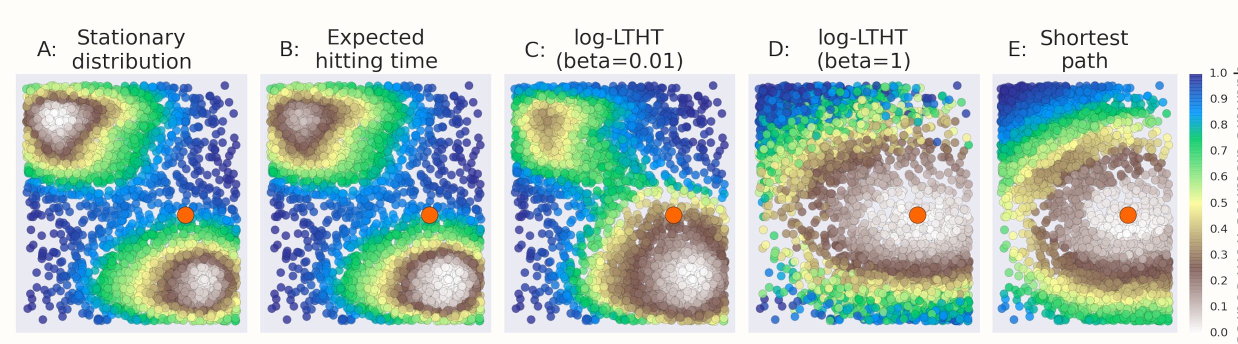

Figure 2 : Estimated distance from orange starting point on a k-nearest neighbor graph constructed on two clusters. A and B show degeneracy of hitting times (Theorem 2). C, D, and E show that log-LTHT interpolate between hitting time and shortest path.

Simulating Brownian motion on the latent space

Summary: We give a method to modify the transition probabilities of the random walk on a spatial graph so trajectories converge to Brownian motion on the latent metric space.

Notation: $q_t(x, x_i)$ = the probability of transitioning to x started at x_i in t time.

We make a regularity assumption on transitions of the simple random walk:

$$\text{For } t = \Theta(g_n^{-2}),\ n q_t(x, x_i) \text{ is a.s. eventually uniformly equicontinuous.} \qquad (\star)$$

Hashimoto-Sun-Jaakkola 2015: under $(\star)$, can consistently estimate $p(x)$ and $\overline{\varepsilon}(x)$.

Theorem 3: Simulating Brownian motion on the latent space

Let $\hat{p}$ and $\hat{\varepsilon}$ be consistent estimators of the density and local scale and A be the adjacency matrix. Then the random walk $\hat{X}_t^n$ with transitions defined below converges to Brownian motion.

$$\mathbb{P}(\hat{X}_{t+1}^n = x_j \mid \hat{X}_t^n = x_i) = \begin{cases} \frac{A_{i,j}\hat{p}(x_j)^{-1}}{\sum_k A_{i,k}\hat{p}(x_k)^{-1}} \hat{\varepsilon}(x_i)^{-2} & i \neq j \\ 1 - \hat{\varepsilon}(x_i)^{-2} & i = j \end{cases}$$

- Attributes of reweighted walk $\to$ attributes of Brownian motion
- Brownian motion is simpler to analyze than general Itô process

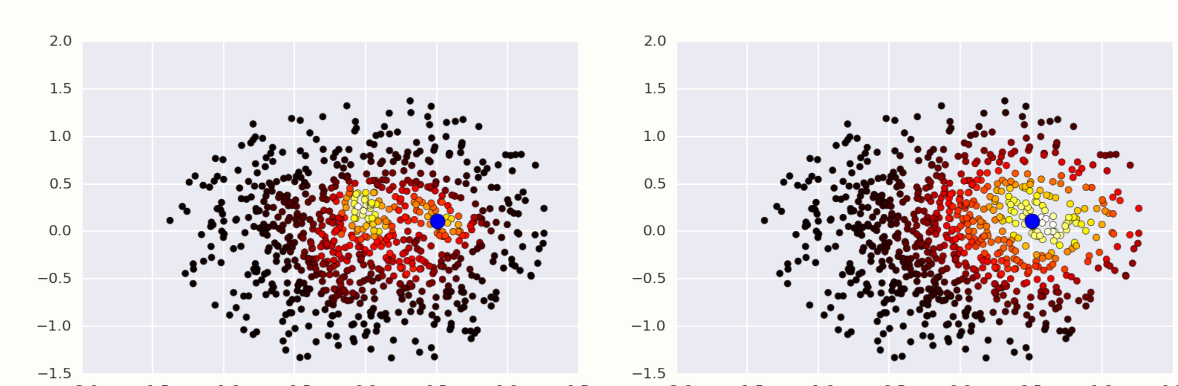

Figure 3 : Distributions of 40-step random walks on a k-nn graph with original and Brownian weighting. Points drawn from Gaussian restricted to a disk.

Laplace transformed hitting time (LTHT)

Summary: We propose the Laplace transform of the hitting time as a metric estimator.

Our analysis of expected hitting time reveals the following drawbacks:

1. The expectation is dominated by long paths;
2. Long paths depend on regions of the graph unrelated to the geodesic.

Resolution: Consider the **Laplace transform** of the hitting time.

Definition 2: The scaled log-Laplace transformed hitting time

Let β_n be a scale parameter. The scaled log **Laplace transformed hitting time (LTHT)** is defined as:

$$-\log(\mathbb{E}[e^{-\beta_n T_{x_j,n}^{x_i}}])/\sqrt{2\beta_n} g_n.$$

and can be computed via the matrix inverse for any transition matrix W

$$-\log(\mathbb{E}[\exp(-\beta T_{x_j,n}^{x_i})])/\sqrt{2\beta_n} g_n = -\log((I - W\exp(-\beta))_{ji}^{-1})/\sqrt{2\beta_n} g_n.$$

- LTHT is closely related to rooted page rank [3] and potential distance [4];
- LTHT puts greater weight on shorter paths;
- LTHT avoids instability of shortest paths by averaging paths near the geodesic.

PDE Characterization: LTHT is solution to Feynman-Kac boundary value problem.

Theorem 4: Feynman-Kac for the Laplace transform

The LTHT $u(x) = \mathbb{E}[\exp(-\beta T_E^x)]$ is the solution to the boundary value problem with boundary condition $u|_{\partial E} = 1$:

$$\frac{1}{2}\mathrm{Tr}[\sigma^T H(u)\sigma] + \mu(x) \cdot \nabla u - \beta u = 0.$$

LTHT is consistent

Summary: For scale parameter $\beta_n = \Theta(\widehat{\beta} g_n^2)$, a modification of LTHT gives consistent metric recovery. We will consider hitting times to the s-neighborhood of a vertex x_i in G_n, the graph equivalent of the ball $B(x_i, s)$.

Modify LTHT using the following notations:

- **s-neighborhood** $NB_n^s(x)$ = estimated set of vertices within latent distance s of x;
- $\widehat{T}_{B(x_j,s)}^{x_i}$ = hitting time of the transformed walk on G_n from x_i to $NB_n^s(x_j)$.

Theorem 5: Consistency of LTHT

Let $x_i, x_j \in G_n$ be connected by a geodesic not intersecting ∂D. For any $\delta > 0$, there exists a choice of $\widehat{\beta}$ and $s > 0$ so that if $\beta_n = \widehat{\beta} g_n^2$, for large n we have whp

$$\left| -\log(\mathbb{E}[\exp(-\beta_n \widehat{T}_{NB_n^s(x_j),n}^{x_i})])/\sqrt{2\widehat{\beta}} - |x_i - x_j| \right| < \delta.$$

Proof intuition:

1. For Brownian motion W_t started at x_i, log-LTHT for $\overline{T}_{B(x_j,s)}^{x_i}$, the hitting time of W_t to $B(x_j, s)$, recovers the latent metric.

   Lemma 1

   For any $\alpha < 0$, if $\widehat{\beta} = s^\alpha$, as $s \to 0$ we have

   $$-\log(\mathbb{E}[\exp(-\widehat{\beta}\overline{T}_{B(x_j,s)}^{x_i})])/\sqrt{2\widehat{\beta}} \to |x_i - x_j|.$$

2. By Theorem 1, discrete log-LTHT to $B(x_j, s)$ converges to continuous log-LTHT;
3. Hitting time to s-neighborhood converges to hitting time to ball $B(x_j, s)$ of radius s.

Empirical consistency:

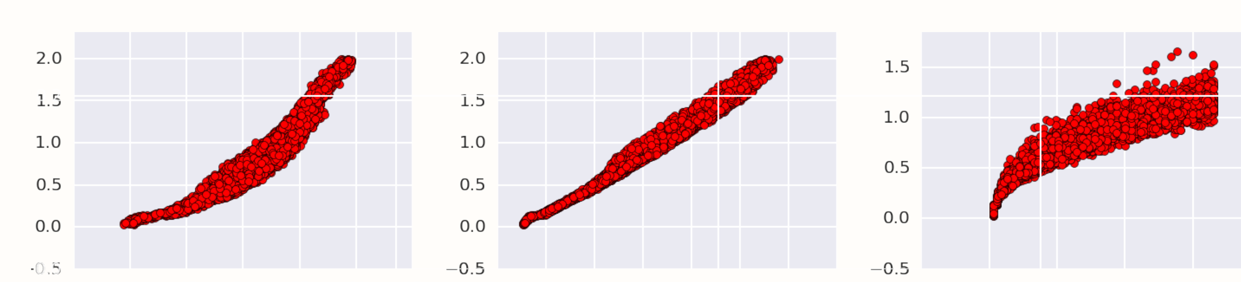

Figure 4 : Estimated distance vs. true latent distance for different values of β on re-weighted walks (simulated dataset) on a k-NN graph with 5000 vertices and $k = 100$.

LTHT preserves clusters

Summary: LTHT learns a cluster preserving metric in 1-D without reweighting.

Consider the density-dependent metric

$$m(x) = \frac{2}{\overline{\varepsilon}(x)^2} + \frac{1}{\beta}\frac{\partial \log(p(x))}{\partial x^2} + \frac{1}{\beta}\left(\frac{\partial \log(p(x))}{\partial x}\right)^2$$

which separates points in different clusters.

Theorem 6: LTHT preserves clusters without debiasing

Suppose $d = 1$ and $h(x) = 1_{x \in [0,1]}$. The hitting time $T_{NB_n^{\hat{\varepsilon}(x_j)g_n}(x_j),n}^{x_i}$ of a simple random walk from x_i to the out-neighborhood of x_j converges to

$$-\log(\mathbb{E}[-\beta T_{NB_n^{\hat{\varepsilon}(x_j)g_n}(x_j),n}^{x_i}])/\sqrt{8\beta} \to \int_{x_i}^{x_j} \sqrt{m(x)} dx + o\left(\log(1 + e^{-\sqrt{2\beta}})/\sqrt{2\beta}\right).$$

LTHT is robust

Summary: LTHT distinguishes between close and far vertices even when a large number of non-geometric noise edges are added to the graph.

Noisy spatial graph: a mixture of spatial graph and Erdös-Renyi noise:

$$\text{Edge from } x_i \text{ to } x_j \text{ with probability } h(|x_i - x_j|\varepsilon_n(x_i)^{-1})(1 - q_j(n)) + q_j(n).$$

Two step log-LTHT: LTHT ignoring immediate hits:

$$M_{ij}^{ts} := -\log(\mathbb{E}[\exp(-\beta T_{x_j,n}^{x_i}) \mid T_{x_j,n}^{x_i} > 1]).$$

Theorem 7: Two-step LTHT reduces to RA index

If $\beta = \omega(\log(g_n^d n))$ and x_i and x_j have at least one common neighbor, then

$$M_{ij}^{ts} - 2\beta \to -\log(R_{ij}) + \log(|NB_n(x_i)|).$$

where the **directed RA index** R_{ij} (well-known local vertex similarity measure) is

$$R_{ij} := \sum_{x_k \in NB_n(x_i) \cap NB_n^{in}(x_j)} |NB_n(x_k)|^{-1}$$

for out and in-neighborhood sets $NB_n(x_i)$ and $NB_n^{in}(x_i)$.

Robustness: RA index distinguishes close and far neighbors even under noise.

Theorem 8: Two-step LTHT is robust under extreme model deviations

If $q_i = q = o(g_n^{d/2})$ for all i, there are c_1, c_2 and h_n so that for any i, j we have whp

- $|x_i - x_j| < \min\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ if $R_{ij} h_n < c_1$;
- $|x_i - x_j| > 2\max\{\varepsilon_n(x_i), \varepsilon_n(x_j)\}$ if $R_{ij} h_n > c_2$.

Evaluation: Link prediction task

Summary: LTHT-based methods outperform benchmarks on two link prediction tasks.

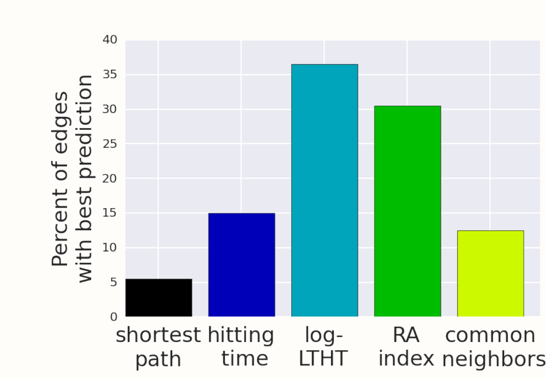

Figure 5 : LTHT recovers deleted edges most consistently on a citation network

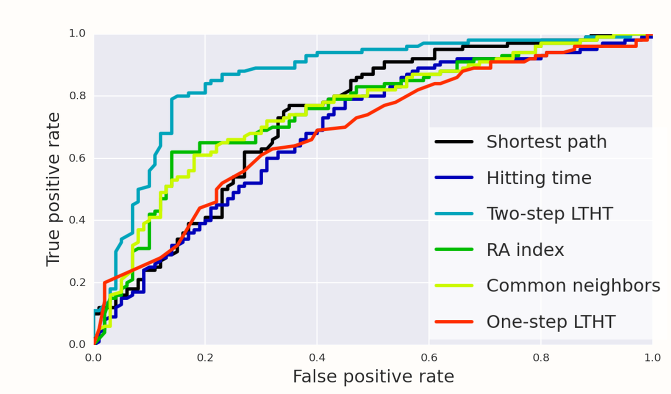

Figure 6 : Two-step LTHT outperforms other techniques at word similarity estimation.

Link prediction: (Figure 5, KDD 2003 challenge data, 11,042 vertices and 222,027 edges): Given a network with edges deleted, how often does each method rank the true, deleted edge above other missing edges. Log-LTHT and RA index both perform well.

Link weight prediction: (Figure 6, Associative thesaurus, 7754 vertices and 246,609 edges): Predict whether two words were associated often in a survey. Task is a classification problem for whether a link is strong (>10 mentions) or weak. Two-step LTHT performs better than other methods.

References

[1] T. Hashimoto, Y. Sun, and T. Jaakkola. Metric recovery from directed unweighted graphs, AISTATS 2015.

[2] U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs, JMLR 2014.

[3] D. Liben-Nell, J. Kleinberg. The link-prediction problem for social networks, CKIM 03.

[4] I. Kivimaki, M. Shimbo, M. Saerens. Developments in the theory of randomized shortest paths, Physica A.

NIPS 2015