From random walks to distances on unweighted graphs

Authored by:

Tommi Jaakkola
Yi Sun
Tatsunori Hashimoto

Abstract

Large unweighted directed graphs are commonly used to capture relations between entities. A fundamental problem in the analysis of such networks is to properly define the similarity or dissimilarity between any two vertices. Despite the significance of this problem, statistical characterization of the proposed metrics has been limited.We introduce and develop a class of techniques for analyzing random walks on graphs using stochastic calculus. Using these techniques we generalize results on the degeneracy of hitting times and analyze a metric based on the Laplace transformed hitting time (LTHT). The metric serves as a natural, provably well-behaved alternative to the expected hitting time. We establish a general correspondence between hitting times of the Brownian motion and analogous hitting times on the graph. We show that the LTHT is consistent with respect to the underlying metric of a geometric graph, preserves clustering tendency, and remains robust against random addition of non-geometric edges. Tests on simulated and real-world data show that the LTHT matches theoretical predictions and outperforms alternatives.

1 Paper Body

Many network metrics have been introduced to measure the similarity between any two vertices. Such metrics can be used for a variety of purposes, including uncovering missing edges or pruning spurious ones. Since the metrics tacitly assume that vertices lie in a latent (metric) space, one could expect that they also recover the underlying metric in some well-defined limit. Surprisingly, there are nearly no known results on this type of consistency. Indeed, it was recently shown [19] that the expected hitting time degenerates and does not measure any notion of distance. We analyze an improved hitting-time metric ? Laplace transformed hitting time (LTHT) ? and rigorously evaluate its consistency, cluster-preservation, and robustness under a general network model which encapsulates the latent space assumption. This network model, specified in Section 2, posits that vertices lie in a latent metric space, and edges are drawn between nearby vertices in that space. To analyze the LTHT, we develop two key technical tools. We establish a correspondence between functionals of hitting time for random walks on graphs, on the one hand, and limiting It? processes (Corollary 4.4) on the other. Moreover, we construct a weighted random walk on the graph whose limit is a Brownian motion (Corollary 4.1). We apply these tools to obtain three main results. First, our Theorem 3.5 recapitulates and generalizes the result of [19] pertaining to degeneration of expected hitting time in the limit. Our proof is direct and demonstrates the broader applicability of the techniques to general random walk based algorithms. Second, we analyze the Laplace transformed hitting time as a one-parameter family of improved distance estimators based on random walks on the graph. We prove that there exists a scaling limit for the parameter ? such that the LTHT can become the shortest path distance (Theorem S5.2) or a consistent metric estimator averaging over many paths (Theorem 4.5). Finally, we prove that the LTHT captures the advantages 1

of random-walk based metrics by respecting the cluster structure (Theorem 4.6) and robustly recovering similarity queries when the majority of edges carry no geometric information (Theorem 4.9). We now discuss the relation of our work to prior work on similarity estimation. Quasi-walk metrics: There is a growing literature on graph metrics that attempts to correct the degeneracy of expected hitting time [19] by interpolating between expected hitting time and shortest path distance. The work closest to ours is the analysis of the phase transition of the p-resistance metric in [1] which proves that p-resistances are nondegenerate for some p; however, their work did not address consistency or bias of p-resistances. Other approaches to quasi-walk metrics such as logarithmic-forest [3], distributed routing distances [16], truncated hitting times [12], and randomized shortest paths [8, 21] exist but their statistical properties are unknown. Our paper is the first to prove consistency properties of a quasi-walk metric. Nonparametric statistics: In the nonparametric statistics literature, the behavior of k-nearest neighbor and ?-ball graphs has been the focus of extensive study. For undirected graphs, Laplacian-based techniques have yielded consistency for clusters [18] and shortest paths [2] as well as the degeneracy of expected hitting time [19]. Algorithms for exactly embedding k-nearest neighbor graphs are similar and generate metric estimates, but require knowledge of the graph construction method, and their consistency properties are unknown [13]. Stochastic differential equation techniques similar to ours were applied to prove Laplacian convergence results in [17], while the process-level convergence was exploited in [6]. Our work advances the techniques of [6] by extracting more robust estimators from process-level information. Network analysis: The task of predicting missing links in a graph, known as link prediction, is one of the most popular uses of similarity estimation. The survey [9] compares several common link prediction methods on synthetic benchmarks. The consistency of some local similarity metrics such as the number of shared neighbors was analyzed under a single generative model for graphs in [11]. Our results extend this analysis to a global, walk-based metric under weaker model assumptions.

2 2.1

Continuum limits of random walks on networks Definition of a spatial graph

We take a generative approach to defining similarity between vertices. We suppose that each vertex i of a graph is associated with a latent coordinate xi ? Rd and that the probability of finding an edge between two vertices depends solely on their latent coordinates. In this model, given only the unweighted edge connectivity of a graph, we define natural distances between vertices as the distances between the latent coordinates xi . Formally, let X = {x1 , x2 , . . .} ? Rd be an infinite sequence of points drawn i.i.d. from a differentiable density with bounded log gradient p(x) with compact support D. A spatial graph is defined by the following: Definition 2.1 (Spatial graph). Let ?n : Xn ? R¿0 be a local scale function and h : R?0 ? [0, 1] a piecewise continuous function with h(x) = 0 for x ¿ 1, h(1) ¿ 0, and h left-continuous at 1. The spatial graph Gn corresponding to ?n and h is the random graph with vertex set Xn and a directed edge from xi to xj with probability pij = h(—xi ? xj —?n (xi)?1). This graph was proposed in [6] as the generalization of k-nearest neighbors to isotropic kernels. To make inference tractable, we focus on the large-graph, small-neighborhood limit as n ? ? and ?n (x) ? 0. In particular, we will suppose that there exist scaling constants gn and a deterministic continuous function ? : D ? R¿0 so that gn ? 0,

1

1

gn n d+2 log(n)? d+2 ? ?,

?n (x)gn?1 ? ?(x) for x ? Xn ,

where the final convergence is uniform in x and a.s. in the draw of X . The scaling constant gn represents a bound on the asymptotic sparsity of the graph. We give a few concrete examples to make the quantities h, gn , and ?n clear. 1. The directed k-nearest neighbor graph is defined by setting h(x) = 1x?[0,1] , the indicator function of the unit interval, ?n (x) the distance to the k th nearest neighbor, and gn = (k/n)1/d the rate at which ?n (x) approaches zero. 2

2. A Gaussian kernel graph is approximated by setting h(x) = exp(?x2 /? 2)1x?[0,1] . The truncation of the Gaussian tails at ? is an analytic convenience rather than a fundamental limitation, and the bandwidth can be varied by rescaling ?n (x). 2.2

Continuum limit of the random walk

Our techniques rely on analysis of the limiting behavior of the simple random walk Xtn on a spatial graph Gn , viewed as a discrete-time Markov process with domain D. The increment at step t of Xtn is a jump to a random point in Xn which lies within the ball of radius ?n (Xtn) around Xtn . We observe three effects: (A) the random walk jumps more frequently towards regions of high density; (B) the random walk moves more quickly whenever ?n (Xtn) is large; (C) for ?n small and a large step count t, the random variable Xtn ? X0n is the sum of many small independent (but not necessarily identically distributed) increments. In the n ? ? limit, we may identify Xtn with a continuous-time stochastic process satisfying (A), (B), and (C) via the following result, which is a slight strengthening of [6, Theorem 3.4] obtained by applying [15, Theorem

11.2.3] in place of the original result of Stroock-Varadhan. Theorem 2.2. The simple random walk Xtn converges uniformly in Skorokhod space D([0, ?), D) after a time scaling b t = tgn2 to the It? process Ybt valued in the space of continuous functions C([0, ?), D) with reflecting boundary conditions on D defined by ? dYbt = ? log(p(Ybt))?(Ybt)2 /3db t + ?(Ybt)/ 3dWbt . (1) Effects (A), (B), and (C) may be seen in the stochastic differential equation (1) as follows. The 2 direction of the drift is controlled by ? log(p(Ybt)), the rate of drift is controlled by b t) , and the ? ?(Y noise is driven by a Brownian motion Wbt with location-dependent scaling ?(Ybt)/ 3.1 We view Theorem 2.2 as a method to understand the simple random walk Xtn through the continuous walk Ybt . Attributes of stochastic processes such as stationary distribution or hitting time may be defined for both Ybt and Xtn , and in many cases Theorem 2.2 implies that an appropriately-rescaled version of the discrete attribute will converge to the continuous one. Because attributes of the continuous process Ybt can reveal information about proximity between points, this provides a general framework for inference in spatial graphs. We use hitting times of the continuous process to a domain E ? D to prove properties of the hitting time of a simple random walk on a graph via the limit arguments of Theorem 2.2.

3

Degeneracy of expected hitting times in networks

The hitting time, commute time, and resistance distance are popular measures of distance based upon the random walk which are believed to be robust and capture the cluster structure of the network. However, it was shown in a surprising result in [19] that on undirected geometric graphs the scaled expected hitting time from xi to xj converges to inverse of the degree of xj . In Theorem 3.5, we give an intuitive explanation and generalization of this result by showing that if the random walk on a graph converges to any limiting It? process in dimension d ? 2, the scaled expected hitting time to any point converges to the inverse of the stationary distribution. This answers the open problem in [19] on the degeneracy of hitting times for directed graphs and graphs with general degree distributions such as directed k-nearest neighbor graphs, lattices, and power-law graphs with convergent random walks. Our proof can be understood as first extending the transience or neighborhood recurrence of Brownian motion for d ? 2 to more general It? processes and then connecting hitting times on graphs to their It? process equivalents. 3.1

Typical hitting times are large

We will prove the following lemma that hitting a given vertex quickly is unlikely. Let Txxji,n be the hitting time to xj of Xtn started at xi and TExi be the continuous equivalent for Ybt to hit E ? D . 1 Both the variance ?(?n (x)2) and expected value ?(? log(p(x))?n (x)2) of a single step in the simple random walk are ?(gn2). The time scaling b t = tgn2 in Theorem 2.2 was chosen so that as n ? ? there are gn?2 discrete steps taken per unit time, meaning the total drift and variance per unit time tend to a non-trivial limit.

3

Lemma 3.1 (Typical hitting times are large). For any d ? 2, c ¿ 0, and ? ¿ 0, for large enough n we have P(Txxji,n ¿ cgn?2) ¿ 1 ? ?. To prove Lemma 3.1, we require the following tail bound following from the Feynman-Kac theorem. Theorem 3.2 ([10, Exercise 9.12] Feynman-Kac for the Laplace transform). The Laplace transform of the hitting time (LTHT) u(x) = E[exp(??TEx)] is the solution to the boundary value problem with boundary condition u—?E = 1: 1 Tr[? T H(u)?] + ?(x) ? ?u ? ?u = 0. 2 This will allow us to bound the hitting time to the ball B(xj , s) of radius s centered at xj . x

Lemma 3.3. For x, y ? D, d ? 2, and any ? ¿ 0, there exists s ¿ 0 such that E[e?TB(y,s)] ¡ ?. Proof. We compare the Laplace transformed hitting time of the general It? process to that of Brownian motion via Feynman-Kac and handle the latter case directly. Details are in Section S2.1. We now use Lemma 3.3 to prove Lemma 3.1. xi a.s. for Proof of Lemma 3.1. Our proof proceeds in two steps. First, we have Txxji,n ? TB(x j ,s),n any s ¿ 0 because xj ? B(xj , s), so by Theorem 2.2, we have xi

?2
lim E[e?Txj ,n gn] ? lim E[e
n??
x
g ?2 j ,s),n n
i ?TB(x
n??
x
] = E[e
i ?TB(x
j ,s)
].
(2)
xi ?TB(x j ,s)
] ¡ 12 ?e?c for some s ¿ 0. For large enough n, this Applying Lemma 3.3, we have E[e xi ?2 ?c combined with (2) implies P(Txj ,n ? cgn)e ¡ ?e?c and hence P(Txxji,n ? cgn?2) ¡ ?. 3.2

Expected hitting times degenerate to the stationary distribution

To translate results from It? processes to directed graphs, we require a regularity condition. Let qt (xj , xi) denote the probability that Xtn = xj conditioned on X0n = xi . We make the following technical conjecture which we assume holds for all spatial graphs. (?) For t = ?(gn?2), the rescaled marginal nqt (x, xi) is a.s. eventually uniformly equicontinuous.2 Let ?X n (x) denote the stationary distribution of Xtn . The following was shown in [6, Theorem 2.1] under conditions implied by our condition (?) (Corollary S2.6). R Theorem 3.4. Assuming (?), for a?1 = p(x)2 ?(x)?2 dx, we have the a.s. limit p(x) ? b(x) := lim n?X n (x) = a . n?? ?(x)2 We may now express the limit of expected hitting time in terms of this result. Theorem 3.5. For d ? 2 and any i, j, we have E[Txxji,n] a.s. 1 . ? n ? b(xj) Proof. We give a sketch. By Lemma 3.1, the random walk started at xi does not hit xj within cgn?2 steps with high probability. By Theorem S2.5, the simple random walk Xtn mixes at exponential rate, implying in Lemma S2.8 that the probability of first hitting at step t ¿ cgn?2 is approximately the stationary distribution at xj . Expected hitting time is then shown to approximate the expectation of a geometric random variable. See Section S2 for a full proof. Theorem 3.5 is illustrated in Figures 1A and 1B, which show with only 3000 points, expected hitting times on a k-nearest neighbor graph degenerates to the stationary distribution. 3 2 Assumption (?) is related to smoothing properties of the graph Laplacian and is known to hold for undirected graphs [4]. No directed analogue is known, and [6] conjectured a weaker property for all spatial graphs. See Section S1 for further details. 3 Surprisingly, [19] proved that 1-D hitting times diverge despite convergence of the continuous equivalent. This occurs because the discrete walk can jump past the target point. In Section S2.4, we consider 1-D hitting

4

Figure 1: 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 3.5). C, D, and E show that log-LTHT interpolate between hitting time and shortest path.

4

The Laplace transformed hitting time (LTHT)

In Theorem 3.5 we showed that expected hitting time is degenerate because a simple random walk mixes before hitting its target. To correct this we penalize longer paths. More precisely, consider for x b x b 2 the Laplace transforms E[e??T E] and E[e??n TE,n] of T x and T x ?b ¿ 0 and ?n = ?g n E,n . E These Laplace transformed hitting times (LTHT?s) have three advantages. First, while the expected hitting time of a Brownian motion to a domain is dominated by long paths, the LTHT is dominated by direct paths. Second, the LTHT for the It? process can be derived in closed form via the FeynmanKac theorem, allowing us to make use of techniques from continuous stochastic processes to control the continuum LTHT. Lastly, the LTHT can be computed both by sampling and in closed form as a matrix inversion (Section S3). Now define the scaled log-LTHT as p xi ? log(E[e??n Txj ,n])/ 2?n gn . Taking different scalings for ?n with n interpolates between expected hitting time (?n ? 0 on a fixed graph) and shortest path distance (?n ? ?) (Figures 1C, D, and E). In Theorem 4.5, we show b 2) yields a consistent distance measure retaining the unique that the intermediate scaling ?n = ?(?g n properties of hitting times. Most of our results on the LTHT are novel for any quasi-walk metric. While considering the Laplace transform of the hitting time is novel to our work, this metric has been used in the literature in an ad-hoc manner in various forms as a similarity metric for collaboration networks [20], hidden subgraph detection [14], and robust shortest path distance [21]. However, these papers only considered the elementary properties of the limits ?n ? 0 and ?n ? ?. Our consistency proof demonstrates the advantage of the stochastic process approach. 4.1

Consistency

It was shown previously that for n fixed and ?n ? ?, ? log(E[??n Txxji,n])/?n gn converges to shortest path distance from xi to xj . We investigate more precise behavior in terms of the scaling of ?n . There are two regimes: if ?n = ?(log(gnd n)), then the shortest path dominates and the LTHT b 2), the graph log-LTHT converges to shortest path distance (See Theorem S5.2). If ?n

= ?(?g n b converges to its continuous equivalent, which for large ? averages over random walks concentrated b 2), we proceed in three steps: (1) we around the geodesic. To show consistency for ?n = ?(?g n reweight the random walk on the graph so the limiting process is Brownian motion; (2) we show that log-LTHT for Brownian motion recovers latent distance; (3) we show that log-LTHT for the reweighted walk converges to its continuous limit; (4) we conclude that log-LTHT of the reweighted walk recovers latent distance. (1) Reweighting the random walk to converge to Brownian motion: We define weights using the estimators pb and ?b for p(x) and ?(x) from [6]. times to small out neighbors which corrects this problem and derive closed form solutions (Theorem S2.12). This hitting time is non-degenerate but highly biased due to boundary terms (Corollary S2.14).

5

Theorem 4.1. Let pb and ?b be consistent estimators of the density and local scale and A be the b n defined below converges to a Brownian motion. adjacency matrix. Then the random walk X t (b(xj)?1 PAi,j p b(xi)?2 i 6= j n n A p b(xk)?1 ? b b i,k k P(Xt+1 = xj — Xt = xi) = 1 ? ?b(xi)?2 i=j Proof. Reweighting by pb and ?b is designed to cancel the drift and diffusion terms in Theorem 2.2 by ensuring that as n grows large, jumps have means approaching 0 and variances which are asymptotically equal (but decaying with n). See Theorem S4.1. 4 (2) Log-LTHT for a Brownian motion: Let Wt be a Brownian motion with W0 = xi , and let xi T B(xj ,s) be the hitting time of Wt to B(xj , s). We show that log-LTHT converges to distance. Lemma 4.2. For any ? ¡ 0, if ?b = s? , as s ? 0 we have q b xi ? log(E[exp(??T)])/ 2?b ? —xi ? xj —. B(xj ,s) Proof. We consider hitting time of Brownian motion started at distance —xi ? xj — from the origin to distance s of the origin, which is controlled by a Bessel process. See Subsection S6.1 for details. b 2): To compare continuous and discrete log-LTHT?s, we (3) Convergence of LTHT for ?n = ?(?g n will first define the s-neighborhood of a vertex xi on Gn as the graph equivalent of the ball B(xi , s). Definition 4.3 (s-neighborhood). Let ?b(x) be the consistent estimate of the local scale from [6] so that ?b(x) ? ?(x) uniformly a.s. as n ? ?. The ?b-weight of a path xi1 ? ? ? ? ? xil is the sum Pl?1 b(xim) of vertex weights ?b(xi). For s ¿ 0 and x ? Gn , the s-neighborhood of x is m=1 ? NBsn (x) := {y — there is a path x ? y of ?b-weight ? gn?1 s}. xi For xi , xj ? Gn , let TbB(x be the hitting time of the transformed walk on Gn from xi to NBsn (xj). j ,s) We now verify that hitting times to the s-neighborhood on graphs and the s-radius ball coincide. d

xi

xi Corollary 4.4. For s ¿ 0, we have gn2 TbNB s (x),n ? T B(xj ,s) . j n

Proof. We verify that the ball and the neighborhood have nearly identical sets of points and apply Theorem 2.2. See Subsection S6.2 for details. (4) Proving consistency of log-LTHT: Properly accounting for boundary effects, we obtain a consistency result for the log-LTHT for small neighborhood hitting times. Theorem 4.5. Let xi , xj ? Gn be connected by a geodesic not intersecting ?D. For any ? ¿ 0, b 2 , for large n we have with high probability there exists a choice of ?b and s ¿ 0 so that if ?n = ?g n q

b ? —xi ? xj — ¡ ?. ? log(E[exp(??n Tbxi s 2 ?)])/ NBn (xj),n

Proof of Theorem 4.5. The proof has three steps. First, we convert to the continuous setting via Corollary 4.4. Second, we show the contribution of the boundary is negligible. The conclusion follows from the explicit computation of Lemma S6.1. Full details are in Section S6. The stochastic process limit based proof of Theorem 4.5 implies that the log-LTHT is consistent and robust to small perturbations to the graph which preserve the same limit (Supp. Section S8). 4 This is a special case of a more general theorem for transforming limits of graph random walks (Theorem S4.1). Figure S1 shows that this modification is highly effective in practice.

6

4.2

Bias

Random walk based metrics are often motivated as recovering a cluster preserving metric. We now show that the log-LTHT of the un-weighted simple random walk preserves the underlying cluster structure. In the 1-D case, we provide a complete characterization. be Theorem 4.6. Suppose the spatial graph has d = 1 and h(x) = 1x?[0,1] . Let T xi ?(x b j)gn NBn

(xj),n

the hitting time of a simple random walk from xi to the out-neighborhood of xj . It converges to Z xj p

? p p ? 2?])/ ? log(E[??T xi ?(x 8? ? m(x)dx + o log(1 + e)/ 2? , b j)gn NBn

where m(x) =

2 ?(x)2

+

(xj),n

1 ? log(p(x)) ? ?x2

xi

+

1 ?

? log(p(x)) ?x

2

defines a density-sensitive metric.

Proof. Apply the WKBJ approximation for Schrodinger equations to the Feynman-Kac PDE from Theorem 3.2. See Corollary S7.2 and Corollary S2.13 for a full proof. The leading order terms of the density-sensitive metric appropriately penalize crossing regions of large changes to the log density; this is not the case for the expected hitting time (Theorem S2.12). 4.3

Robustness

While shortest path distance is a consistent measure of the underlying metric, it breaks down catastrophically with the addition of a single non-geometric edge and does not meaningfully rank vertices that share an edge. In contrast, we show that LTHT breaks ties between vertices via the resource allocation (RA) index, a robust local similarity metric under Erd?os-R?nyi-type noise. 5 Definition 4.7.

The noisy spatial graph Gn over Xn with noise terms q1 (n), . . ., qn (n) is constructed by drawing an edge from xi to xj with probability pij = h(—xi ? xj —?n (xi)?1)(1 ? qj (n)) + qj (n). Define the directed RA index in terms of the out-neighborhood set NBn (xi) and the in-neighborhood P ts set NBin —NBn (xk)—?1 and two step log-LTHT by Mij := n (xi) as Rij := xk ?NBn (xi)?NBin n (xj) xi xi 6 ? log(E[exp(??Txj ,n) — Txj ,n ¿ 1]). We show two step log-LTHT and RA index give equivalent methods for testing if vertices are within distance ?n (x). Theorem 4.8. If ? = ?(log(gnd n)) and xi and xj have at least one common neighbor, then ts Mij ? 2? ? ? log(Rij) + log(—NBn (xi)—).

Proof. Let Pij (t) be the probability of going from xi to xj in t steps, and Hij (t) the probability of not hitting before time t. Factoring the two-step hitting time yields ?

X Pij (t) ts Hij (t)e??(t?2) . Mij = 2? ? log(Pij (2)) ? log 1 + P (2) t=3 ij

Let kmax be the maximal out-degree in Gn . The contribution of paths of length greater than 2 2 vanishes because Hij (t) ? 1 and Pij (t)/Pij (2) ? kmax , which is dominated by e?? for ? = Rij n ?(log(g n)). Noting that Pij (2) = —NBn (xi)— concludes. For full details see Theorem S9.1. d/2

For edge identification within distance ?n (x), the RA index is robust even at noise level q = o(gn). Modifying the graph by changing fewer than gn2 /n edges does not affect the continuum limit of the random graph, and therefore preserve the LTHT with parameter ? = ?(gn2). While this weak bound allows on average o(1) noise edges per vertex, it does show that the LTHT is substantially more robust than shortest paths without modification. See Section S8 for proofs. 6 The conditioning Txxji,n ¿ 1 is natural in link-prediction tasks where only pairs of disconnected vertices are queried. Empirically, we observe it is critical to performance (Figure 3). 5

7

Figure 2: The LTHT recovered deleted edges most consistently on a citation network

Figure 3: The two-step LTHT (defined above Theorem 4.8) outperforms others at word similarity estimation including the basic log-LTHT.

d/2

Theorem 4.9. If qi = q = o(gn) for all i, for any ? ¿ 0 there are c1 , c2 and hn so that for any i, j, with probability at least 1 ? ? we have ? —xi ? xj — ¡ min{?n (xi), ?n (xj)} if Rij hn ¡ c1 ; ? —xi ? xj — ¿ 2 max{?n (xi), ?n (xj)} if Rij hn ¿ c2 . Proof. The minimal RA index follows from standard concentration arguments (see S9.2).

5

Link prediction tasks

We compare the LTHT against other baseline measures of vertex similarity: shortest path distance, expected hitting time, number of common neighbors, and the RA index. A comprehensive evaluation of these quasi-walk metrics was performed in [8] who showed that a metric equivalent to the LTHT performed best. We consider two separate link prediction tasks on the largest connected component of vertices of degree at least five, fixing ? = 0.2.7 The degree constraint is to ensure that local methods using number of common neighbors such as the resource allocation index do not have an excessive number of ties. Code to generate figures in this paper are contained in the supplement. Citation network: The KDD 2003 challenge dataset [5] includes a directed, unweighted network of e-print arXiv citations whose dense connected component has 11,042 vertices and 222,027 edges. We use the same benchmark method as [9] where we delete a single edge and compare the similarity of the deleted edge against the set of control pair of vertices i, j which do not share an edge. We count the fraction of pairs on which each method rank the deleted edge higher than all other methods. We find that LTHT is consistently best at this task (Figure 2). 8 Associative Thesaurus network: The Edinburgh associative thesaurus [7] is a network with a dense connected component of 7754 vertices and 246,609 edges in which subjects were shown a set of ten words and for each word was asked to respond with the first word to occur to them. Each vertex represents a word and each edge is a weighted, directed edge where the weight from xi to xj is the number of subjects who responded with word xj given word xi . We measure performance by whether strong associations with more than ten responses can be distinguished from weak ones with only one response. We find that the LTHT performs best and that preventing one-step jumps is critical to performance as predicted by Theorem 4.8 (Figure 3).

6

Conclusion

Our work has developed an asymptotic equivalence between hitting times for random walks on graphs and those for diffusion processes. Using this, we have provided a short extension of the proof for the divergence of expected hitting times, and derived a new consistent graph metric that is theoretically principled, computationally tractable, and empirically successful at well-established link prediction benchmarks. These results open the way for the development of other principled quasi-walk metrics that can provably recover underlying latent similarities for spatial graphs. 7 8

Results are qualitatively identical when varying ? from 0.1 to 1; see supplement for details. The two-step LTHT is not shown since it is equivalent to the LTHT in missing link prediction.

8

2 References

[1] M. Alamgir and U. von Luxburg. Phase transition in the family of p-resistances. In Advances in Neural Information Processing Systems, pages 379?387, 2011. [2] M. Alamgir and U. von Luxburg. Shortest path distance in random k-nearest neighbor graphs. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), pages 1031?1038, 2012. [3] P. Chebotarev. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. Discrete Applied Mathematics, 159(5):295?302, 2011. [4] D. A. Croydon and B. M. Hambly. Local limit theorems for se-

quences of simple random walks on graphs. Potential Analysis, 29(4):351?389, 2008. [5] J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 KDD Cup. ACM SIGKDD Explorations Newsletter, 5(2):149?151, 2003. [6] T. B. Hashimoto, Y. Sun, and T. S. Jaakkola. Metric recovery from directed unweighted graphs. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, pages 342?350, 2015. [7] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of English and its computer analysis. The Computer and Literary Studies, pages 153?165, 1973. [8] I. Kivim?ki, M. Shimbo, and M. Saerens. Developments in the theory of randomized shortest paths with a comparison of graph node distances. Physica A: Statistical Mechanics and its Applications, 393:600?616, 2014. [9] L. L? and T. Zhou. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6):1150?1170, 2011. [10] B. ?ksendal. Stochastic differential equations: An introduction with applications. Universitext. SpringerVerlag, Berlin, sixth edition, 2003. [11] P. Sarkar, D. Chakrabarti, and A. W. Moore. Theoretical justification of popular link prediction heuristics. In IJCAI Proceedings-International Joint Conference on Artificial Intelligence, volume 22, page 2722, 2011. [12] P. Sarkar and A. W. Moore. A tractable approach to finding closest truncated-commute-time neighbors in large graphs. In In Proc. UAI, 2007. [13] B. Shaw and T. Jebara. Structure preserving embedding. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 937?944. ACM, 2009. [14] S. T. Smith, E. K. Kao, K. D. Senne, G. Bernstein, and S. Philips. Bayesian discovery of threat networks. IEEE Transactions on Signal Processing, 62:5324?5338, 2014. [15] D. W. Stroock and S. S. Varadhan. Multidimensional diffussion processes, volume 233. Springer Science & Business Media, 1979. [16] A. Tahbaz-Salehi and A. Jadbabaie. A one-parameter family of distributed consensus algorithms with boundary: From shortest paths to mean hitting times. In Decision and Control, 2006 45th IEEE Conference on, pages 4664?4669. IEEE, 2006. [17] D. Ting, L. Huang, and M. I. Jordan. An analysis of the convergence of graph Laplacians. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 1079?1086, 2010. [18] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. The Annals of Statistics, pages 555?586, 2008. [19] U. von Luxburg, A. Radl, and M. Hein. Hitting and commute times in large random neighborhood graphs. Journal of Machine Learning Research, 15:1751?1798, 2014. [20] M. Yazdani. Similarity Learning Over Large Collaborative Networks. PhD thesis, ?cole Polytechnique F?d?rale de Lausanne, 2013. [21] L. Yen, M. Saerens, A. Mantrach, and M. Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785?793. ACM, 2008.

9