A METRIC ON DIRECTED GRAPH NODES BASED ON HITTING PROBABILITIES

Zachary M. Boyd, Nicolas Fraiman, Jeremy L. Marzuola, Peter J. Mucha, Braxton Osting, Jonathan Weare

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

Summary

We introduce a distance function on directed graphs using the *hitting probability* of an ordered pair of nodes, which is the probability that a random walker starting at the first node will reach the second before returning to the first. Our metric uncovers direction-based structure, such as looping and dynamical trapping, that is invisible to other directed graph metrics and symmetrizations. Example applications, such structure discovery, weak community detection in dense graphs, multiscale analysis, and partitioning of Markov chains, will be discussed. Full details are available in the arXiv preprint 2006.14482, which is currently under review.

Metric

The three undirected graph metrics (shortest path, commute time/effective resistance, and diffusion distance [6]) have been widely applied to tasks such as manifold learning, link prediction, and route planning [6, 8, 1]. Most machine learning methods expect the data to lie in a metric space as well.

For directed graphs, metric notions become much less obvious. For example, if there is an edge from i to j and not in the other direction, does this mean that i is nearer to j than j is to i? Of course, such an idea is contrary to the spirit of a metric, so one is led to search for formulations that account for the directionality in some way without losing the metric symmetry. Few previous papers have suggested metrics using generalized effective resistance [14, 11, 3, 4] or Markov chain curvature [13], plus many machine learning proposals [7]. A related idea is graph symmetrization, which seeks to convert a directed graph into an undirected graph, to which many other tools can then be applied. The simplest such symmetrization simply forgets edge directionality, but others have been proposed [9].

In this work, we derive a directed graph metric based on hitting probabilities. Given a directed, strongly connected graph G, with nodes i and j, the hitting probability from i to j, $Q_{i,j}$, is the probability that a random walker starting from node i will reach node j before returning to i. Using ideas from Markov chain theory, we prove that $\phi_i Q_{i,j} = \phi_j Q_{j,i}$, where ϕ is the invariant distribution. This implies that

$$A^{(\mathrm{hp},\beta)}_{i,j} = \frac{\phi_i^\beta}{\phi_j^{1-\beta}} Q_{i,j}$$

for $\beta \in [\frac{1}{2}, 1]$ gives the adjacency matrix of an undirected, symmetric graph. We furthermore show that when $\beta \neq \frac{1}{2}$,

$$d^\beta_{i,j} = -\log A^{(\mathrm{hp},\beta)}_{i,j}$$

is a metric.

When $\beta = \frac{1}{2}$, it is possible for certain sets of "bottleneck" nodes to be at distance zero from each other, making $d^{\frac{1}{2}}$ only a pseudometric. Such bottlenecks occur when with $Q_{i,j} = 1$, i.e. any walk beginning at i must reach j before returning (and vice versa). We develop an interesting structure theory of directed graphs that places the bottleneck nodes into equivalence classes using $d^{\frac{1}{2}}$ and the other nodes into *segments* that lie between bottleneck nodes. Intuitively, graphs with bottleneck nodes have a global cyclical structure for each equivalence class, and segment membership is a way to track progress through global cycles. Using the structure theory, we give a graph quotienting process based on [10] which collapses equivalent nodes and preserves intra-segment distances. We also prove tight bounds on the distortion of inter-segment distances.

The only other paper of which we are aware that uses Q for network science is [2], which finds it effective for link prediction. In particular, [2] does not consider symmetrization or metric properties.

Numerics and examples

Our metric can be computed using recent tools from Markov process approximation theory [12]. The technique is summarized as follows: Instead of computing $Q_{i,j}$ directly, one considers $Q_{i,j,k}$, the probability that a random walk starting at i reaches j before encountering k. These quantities are related by a one-hop recurrence, which leads to a linear algebra formula requiring the computation of N $N \times N$ matrix inverses. It can be shown

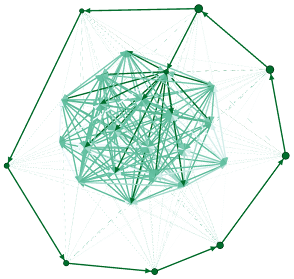

Figure 1: A graph where d^β gives useful information not recovered by other approaches. This is a directed Erdős-Rényi (ER) graph, together with a directed cycle. There are uniformly random edges from the ER graph to the cycle, and one node in the cycle has out edges to the ER graph. Thus, a random walker on this graph would see the cycle as a trap that takes on average a long time to escape relative to similarly sized subsets of the ER nodes. A spectral approach using d^β identifies the cycle, whereas spectral approaches based on naive symmetrization or [5] do not.

that each of these N matrices is related to the others by a rank 2 perturbation, so that the Woodbury matrix identity allows the work to be reduced to a single matrix inversion plus some adjustments. Our implementation of this approach in MATLAB is at `github.com/zboyd2/hitting_probabilities_metric`.

In examples, we compute d^β for a graph with 38 million edges in 31 seconds on a desktop computer. The asymptotic complexity is $O(N^2)$ space (where N is the number of nodes) and $O(N^3)$ time, plus the time for computing ϕ.

Our metric can be computed analytically on some simple graphs, which we present. We also illustrate the results of our method on a synthetic graphs with planted structures, including loops and correlated directions (see Fig. 1 for one example). Simple spectral computations based on $A^{(\text{hp},\beta)}$ uncover the planted structure, which is invisible to other graph symmetrizations. We also show that our metric is qualitatively different from other directed graph metrics.

Applications

The metric structure is a natural starting point for a wide variety of applications. Consider four examples:

1. Using a three-community directed stochastic block model, applying k-means to a PCA embedding of $d^{\frac{1}{2}}$ enhanced weak detectability relative to k-means plus PCA applied to A itself.
2. For a New York Taxi transit network, our metric helped with multiscale structure detection, showing the role of Staton Island.
3. Comparison with Euclidean distance on geometric graphs suggests a possible consistency result for $d^{\frac{1}{2}}$, whereas d^1 may measure something very different.
4. The hitting probabilities metric is insensitive to walk length, making it useful for long-time analysis relative to commute time/effective resistance.

References

[1] I. Abraham, A. Fiat, A. Goldberg, and R. Werneck. Highway dimension, shortest paths, and provably efficient algorithms. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2010.

[2] B. Balls-Barker and B. Webb. Link prediction in networks using effective transitions. *arXiv 1909.01076*, 2019.

[3] P. Chebotarev and E. Deza. Hitting time quasi-metric and its forest representation. *Optim. Lett.*, 14:291–307, 2020.

[4] M. C. H. Choi. On resistance distance of Markov chain and its sum rules. *Lin. Alg. Appl.*, 571:14–25, 2019.

[5] F. Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

[6] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Nat. Acad. Sci. U.S.A.*, 102(21):7426–7431, 2005.

[7] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 40(3):42–51, 2017.

[8] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, New York, 2003. ACM Press.

[9] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Phys. Reports*, 533(4):95–142, 2013.

[10] B. Mitavskiy, J. Rowe, A. Wright, et al. Quotients of Markov chains and asymptotic properties of the stationary distribution of the Markov chain associated to an evolutionary algorithm. *Genet. Program Evolvable Mach.*, 9:109—123, 2008.

[11] M. R. Rozinas. Metric on state space of Markov chain. arXiv:1004.4264, 2010.

[12] E. Thiede, B. V. Koten, and J. Weare. Sharp entrywise perturbation bounds for Markov Chains. *SIAM J. Matrix Anal. Appl.*, 36(3):917–941, 2015.

[13] F. Völlering. Constant curvature metrics for Markov chains. arXiv:1712.02762, 2018.

[14] G. F. Young, L. Scardovi, and N. E. Leonard. A new notion of effective resistance for directed graphs—part I: Definition and properties. *IEEE Trans. Automatic Control*, 61(7):1727–1736, 7 2016.