Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Path Boxplots: A Method for Characterizing Uncertainty in Path Ensembles on a Graph

Mukund Raj, Mahsa Mirzargar, Robert Ricci, Robert M. Kirby & Ross T. Whitaker

To cite this article: Mukund Raj, Mahsa Mirzargar, Robert Ricci, Robert M. Kirby & Ross T. Whitaker (2017) Path Boxplots: A Method for Characterizing Uncertainty in Path Ensembles on a Graph, Journal of Computational and Graphical Statistics, 26:2, 243-252, DOI: 10.1080/10618600.2016.1209115

To link to this article: http://dx.doi.org/10.1080/10618600.2016.1209115

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 2, 243–252
http://dx.doi.org/10.1080/10618600.2016.1209115

Taylor & Francis
Taylor & Francis Group

Path Boxplots: A Method for Characterizing Uncertainty in Path Ensembles on a Graph

Mukund Raj[a], Mahsa Mirzargar[a], Robert Ricci[b], Robert M. Kirby[a], and Ross T. Whitaker[a]

[a]Scientific Computing and Imaging Institute and School of Computing, University of Utah, Salt Lake City, Utah; [b]School of Computing, University of Utah, Salt Lake City, Utah

ABSTRACT

Graphs are powerful and versatile data structures that can be used to represent a wide range of different types of information. In this article, we introduce a method to analyze and then visualize an important class of data described over a graph—namely, ensembles of paths. Analysis of such path ensembles is useful in a variety of applications, in diverse fields such as transportation, computer networks, and molecular dynamics. The proposed method generalizes the concept of *band depth* to an ensemble of paths on a graph, which provides a center-outward ordering on the paths. This ordering is, in turn, used to construct a generalization of the conventional boxplot or whisker plot, called a *path boxplot*, which applies to paths on a graph. The utility of path boxplot is demonstrated for several examples of path ensembles including paths defined over computer networks and roads. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2015
Revised June 2016

KEYWORDS
Data depth; Descriptive statistics; Ensemble; Graph; Paths

1. Introduction

Making sense of sets of information defined over graphs can often be a challenging task. This is because graphs are typically used to represent abstract data that may not be easily representable in a flat, or Euclidean, space. Here, we define a graph $G(V, E, W)$ as a set of vertices (or nodes) V, a set of edges $E \subseteq V \times V$ and a set of edge weights, $W : E \mapsto \mathbb{R}^+$, assigned to each edge. In this article, we describe a method to gain insight into a particular type of data represented on graphs—namely, collections or *ensembles* of paths on graphs, henceforth referred to as *path ensembles*. We define a path (a special type of subgraph) as a sequence of vertices $p = (v_i : 1 \le i \le m)$, where $v_i \in V$ and each consecutive pair of vertices in the sequence have an associated edge, $(v_i, v_{i+1}) \in E \ \forall\ i = \{1, \dots, m-1\}$. We define a path ensemble as a collection of paths on a particular graph.

Paths on a graph are natural structures used to describe and analyze data in a range of applications. For instance, in transportation urban planners study ensembles of paths of commuters (e.g., from recorded GPS data) to identify important travel corridors to plan new routes (Evans et al. 2013). Analysis is performed on a graph whose vertices are usually transition points (road intersections, airports). These vertices have a geographical location and an abstract, logical meaning. The edges in the graph represent direct transportation connections between vertices (segments of roads, routes of airplanes), and they often encode, as weights, information about transit time or cost. A path on this graph is an abstraction of a commuter's path.

In computer networks, system administrators try to detect anomalies or attacks by keeping track of the paths taken by the network traffic over a period of time (Butler et al. 2010). Analysis is performed on a graph whose vertices are Internet subdomains known as *autonomous systems* (ASes) and edges represent a direct data link between ASes, which can encode, as weights, transfer capacity. A path on this *AS graph* represents path of a packet on the Internet.

In molecular dynamics where scientists are interested in studying the protein folding process, various possible configurations (also known as *states*) of a specific protein structure are known while the sequence of discrete intermediate states in the process of protein folding is not. Analysis is performed on a *configuration graph* whose vertices represent the possible protein configurations and weights on edges denote the respective transition probabilities between the associated pair of configurations. In this case, a path is a sequence of potential discrete intermediate states and may be identified by carrying out simulations that incorporate stochastic transitions. These simulations result in an ensemble of possible paths for a folding process on the graph associated with a molecule (Apaydin et al. 2003). In path analysis (Wright 1934), graphs are used to model dependencies (encoded as edges) among a set of variables (encoded as vertices). Direct and indirect dependencies between variables can be represented as edges and paths, respectively, in a model (graph).

Recently, researchers began considering the problem of systematically analyzing and visualizing path ensembles. One of the first challenges is how to summarize or aggregate the information in path ensembles. One approach of aggregation relies on specialized heuristics that often incorporate statistics of low-dimensional descriptors of paths. In road networks, the average travel time between two nodes becomes a salient feature (Hua and Pei 2010). In the analysis of computer networks, one might

CONTACT Mukund Raj ✉ mraj@cs.utah.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

quantify the amount of traffic passing through a node in a computer network (Butler et al. 2010). In molecular dynamics, the product of transition probabilities along folding paths is considered (Apaydin et al. 2003).

Another aggregation approach proposed by researchers is to compare paths directly, rather than using low-dimensional descriptors. Aggregate operations on path ensembles often rely on a definition of the distance between two paths such as Hausdorff (Von Landesberger et al. 2011) or Fréchet (Eiter and Mannila 1994) metrics, which are, in turn, based on distances between individual vertices. From these distances, one can generalize the classical notions of statistical summaries such as median and mean (Evans et al. 2012; Agarwal et al. 2014), as well as clustering (Kharrat et al. 2008). Such aggregate characteristics for a path ensemble can help in understanding the structure of the ensemble.

To fully understand structure of path ensemble by evaluating relationships between paths, applications need to consider not only distances between vertices in a path, but also patterns or differences in the (global) structure or *shapes* of paths. For instance, some paths may deviate from a central or most representative path, but may deviate in either typical or atypical ways. State-of-the-art aggregation techniques for path ensembles typically ignore the relationships that may exist between patterns of vertices in a path.

A growing body of research in analysis methods based on the notion of *data depth* robustly account for nonlocal relationships (correlation) among variables in multidimensional data, in essence capturing their global structure faithfully. Data depth is a method from descriptive statistics that provides a way to quantify *centrality* of multivariate points in an ensemble and derive a center outward ordering, with few assumptions about the underlying distribution. Data depth has been shown to generalize to multidimensional data, and data depth formulations, which account for relationships among variables, have been developed for specialized data types such as functions (López-Pintado and Romo 2009; Sun and Genton 2012), isocontours (Whitaker et al. 2013), and curves (Mirzargar et al. 2014; López-Pintado et al. 2014). Motivated by formulations of data depth for ensembles of multidimensional data, we propose a generalization of data depth for path ensembles on graphs, which we call *path band depth*. At a high level, our generalization comprises of the following two parts, which it shares with earlier formulations for functions and curves: (i) definition of *band* formed by a set of ensemble members and (ii) definition of path band depth. We also propose a visualization strategy for path ensembles, which we call *path boxplots*, based on the order statistics induced by the depth assigned to the paths.

This article is organized as follows. In Section 2, we briefly discuss distance metrics that are currently used to analyze path ensembles. This is followed by the notion of data depth and *band depth*, a type of data depth, and its existing formulations to specialized data types such as functions and curves. In Section 3, we develop our generalization of band depth for paths. In Section 4, we develop our proposed path boxplot visualization strategy. In Section 5, we compare our generalization to distance metric-based alternatives using synthetic data and present two real applications: transportation and computer networks.

2. Background and Related Work

We begin with a brief discussion of current methods for analysis of path ensembles. To select a representative path, Evans et al. (2012, 2013) proposed a generalization of Hausdorff distance for sets of vertices on graphs, which they call network Hausdorff distance (NHD). The classical Hausdorff distance is a measure of dissimilarity between sets and is defined as the maximum of distances from a set of points to their respective nearest neighbor in another set. For paths, we let p_a and p_b denote the sets of vertices for two paths within a weighted graph, then the network Hausdorff distance is defined as (Evans et al. 2013)

$$d_{\mathrm{H}}(p_a, p_b) = \max_{v_a \in p_a} \min_{v_b \in p_b} d_g(v_a, v_b), \tag{1}$$

where $d_g(v_a, v_b)$ is the geodesic (or shortest path) distance between vertices v_a and v_b. The path minimizing the sum of distances from all other paths in an ensemble is the most representative path, a natural generalization of the median.

Alternatively, Eiter and Mannila (1994) used the *discrete Fréchet distance* (DFD) between paths, as an approximation of the classical Fréchet distance. It relies on the set monotonic orderings of the vertices (correspondences or parameterization between paths). The length associated with a correspondence between two paths is defined as the maximum geodesic distance between corresponding vertices, and the DFD distance is defined as the minimum length over all possible correspondences. As with functions, point-based metrics of geometric distances, such as NHD and DFD, generally do not account for the overall, global structure of objects (paths in this case). Therefore, although such metric account for worst-case, vertex distances, they do not capture what is generally referred to as *shape differences* in the geometric setting.

This article proposes a method for exploratory analysis or *visualization* of path ensembles on graph, with consideration of their global structure. The proposed approach is motivated by the univariate boxplot (see Figure 1(a)) introduced by Tukey (Tukey 1977) as an exploratory data analysis tool, based on data depth to summarize the descriptive statistical summaries of an ensemble, based on rank statistics, such as: median, first and third quartile, nonoutlying minimum and maximum values, and identified outliers.

A widely adopted strategy for evaluating the depth of a data sample with respect to a data ensemble is *band depth*. Band depth is a formulation of data depth that relies on the probability that a data point lies *between* a random selection of other points from the distribution. For multivariate data, the *simplicial depth* of a n-dimensional point is the probability of a data point lying in the simplex formed by $n + 1$ (distinct) randomly chosen points from the distribution (Liu 1990). Lopez-Pintado and Romo proposed a concept of band depth for functions (López-Pintado and Romo 2009), in a way that goes beyond point-wise analysis of functions and provides an analysis that accounts for nonlocal correlations that span the function domain. Sun and Genton (2012) used this data ordering to construct *functional boxplots*, a generalization of the conventional whisker plot for visualization of ensembles of functions (see Figure 1(b)). Several authors have proposed extensions of functional band depth to

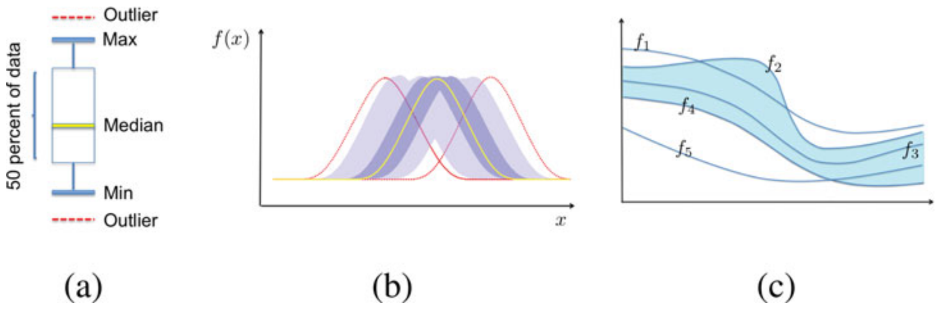

Figure 1. (a) A classic boxplot for univariate data. (b) A functional boxplot for an ensemble of functions. The median function is drawn in yellow, outlier functions in red. The 50% and the 100% data envelope are shown in dark and light purple, respectively. (c) An ensemble of five functions and a sample band formed by three member functions (f_2, f_3, and f_4) from the ensemble.

curves in n dimensions and associated boxplots (López-Pintado et al. 2014; Mirzargar et al. 2014).

The proposed method generalizes the method of function/curve band depth for paths, and therefore we give a brief overview of methodology for band depth on functions/curves (López-Pintado and Romo 2009; López-Pintado et al. 2014; Mirzargar et al. 2014). First, we consider an ensemble of n functions:

$$\mathcal{E} = \{f_1(t), f_2(t), \cdots, f_n(t)\} \subset \mathbb{F}, \quad f_i \in \mathbb{F}, \tag{2}$$

where $\mathbb{F} = \{f | f : \mathbb{R} \mapsto \mathbb{R}\}$ denotes the space of continuous functions on a compact interval. A function g falls within the band $B[\cdot]$ formed by a set of j functions if it lies within their min/max envelope (see Figure 1(c)). That is,

$$g \subset \mathrm{B}\big[\{f_{i_1}, \ldots, f_{i_j}\}\big] \quad \text{iff} \quad \min\left(f_{i_1}(t), \ldots, f_{i_j}(t)\right) \le g(t) \\ \le \max\left(f_{i_1}(t), \ldots, f_{i_j}(t)\right) \quad \forall t. \tag{3}$$

Note that the *band* associated with a random set of functions is the min/max envelope, and the inclusion in the band forms a binary *test* that provides evidence of centrality—not to be confused with other statistical summaries, such as confidence interval or variance on functions.

The band depth of each ensemble member, g, is defined as the probability of its inclusion within the band formed by a random selection of j other functions from the ensemble:

$$\mathrm{BD}_j\left(g\right) = \mathrm{Prob}\left(g \subset B\big[\{f_{i_1}, \ldots, f_{i_j}\}\big]\right). \tag{4}$$

For computation, the probability in Equation (4) is expressed as the expectation of the characteristic function on $g \subset B[\{f_{i_1}, \ldots, f_{i_j}\}]$, and approximated by a sample mean using all choices of j samples from the ensemble (or a random subset, if the ensemble is large):

$$\begin{aligned}&\mathrm{Prob}\left(g \subset B\left[\{f_{i_1}, \ldots, f_{i_j}\}\right]\right) \\ &\quad = E\big[\chi\left(g \subset B\left[\{f_{i_1}, \ldots, f_{i_j}\}\right]\right)\big] \\ &\quad \approx \frac{1}{\binom{n}{j}} \sum_{\{f_{i_1}, \ldots f_{i_j}\} \subset \mathcal{E}} \chi\left(g \subset B\left[\{f_{i_1}, \ldots, f_{i_j})\}\right]\right),\end{aligned} \tag{5}$$

where $\chi(\cdot)$ denotes the characteristic function.

Several practical issues are worth noting. The choice of the number of samples j used to form the band is not specified by the formulation, and may depend on the nature of the data (e.g., variability, number of samples). For larger ensembles, the total number of j-sized subsets may be too large, in which case random subsets may be chosen. Alternatively, the number of j-sized subsets of $\mathcal{E}$ may not be large enough to produce reliable probability estimates and properly order the samples. To address this issue, López-Pintado and Romo (2009) proposed *modified functional band depth*, which replaces the characteristic function χ in Equation (5) with the measure over the domain of $f \in \mathbb{F}$ for which the point-wise inclusion within the band holds. This relaxation can undermine the shape discrimination properties of the depth formulation. Alternatively, Whitaker et al. (2013) proposed an ε-modified band depth (for sets and contours) that relaxes χ to allow a certain amount (e.g., percentage) of the domain to fall outside of the band.

3. Band Depth for Paths on Graphs

In this article, we propose a formulation of band depth for vertices of a graph, and extend that formulation to band depth for paths on graphs. The strategy for building a band for paths mirrors the development of the band depth for curves (i.e., functions $c : \mathbb{R} \mapsto \mathbb{R}^n$) (López-Pintado et al. 2014; Mirzargar et al. 2014), which is to establish a definition of a band for points in the range of the function in $\mathbb{R}^n$ and then apply that band definition for all points in the domain.

In $\mathbb{R}^n$, the band formed by a set of j points has been formulated as the convex hull of $\mathcal{X} = \{x_1, \ldots, x_j\}$ where $x_i \in \mathbb{R}^n$ $\forall i \in \{1, \ldots, j\}$ (Liu 1990). The convex hull of $\mathcal{X}$, $H[\mathcal{X}]$ is the smallest convex region that contains $\mathcal{X}$. $H[\mathcal{X}]$ is a simplex for $j = n + 1$ (and points in general position), and $H[\mathcal{X}]$ has measure zero for $j \le n$. For $n = 1$, the convex hull is the subset of the real numbers bounded by the minimum and maximum of the points in $\mathcal{X}$. Lopez-Pintado et al., as well as Mirzargar et al., generalized the function-band-depth formulation to curves, $\mathcal{C}_j(t) = \{c_1(t), \ldots, c_j(t)\}$ where $c_i : \mathbb{R} \mapsto \mathbb{R}^n$, using the parameterized set of convex hulls for points in $\mathbb{R}^n$. That is $B[\mathcal{C}_j](t) = H[\{c_1(t), \ldots, c_j(t)\}]$. Here, we use a similar generalization strategy for paths on graphs, namely, a parameterized convex hull on the vertices.

We define the *length* of a path p as the sum of weights along its edges, denoted $\|p\|$, while its *cardinality* $|p|$ is the number of constituting vertices. A *geodesic* between two vertices (u, v) is the path between them with the shortest length, and we denote this geodesic distance as $d_g(u, v)$. Geodesic (shortest) paths are not necessarily unique in a graph. In this article, to clarify the discussion, we will generally assume there exists some consistent way to decide among multiple geodesics (in our implementation we use the first geodesic found by Dijkstra's algorithm), while the theory and formulation can be extended to the possibility of multiple geodesics.

We begin with a definition of the band formed by vertices on a graph. Let us define subsets of vertices of size j as follows: $S_j = \{\mathcal{V} \subset \mathcal{P}(V) : |\mathcal{V}| = j\}$ where $\mathcal{P}(V)$ is the power set of V. A vertex v is said to lie in the *band* formed by $\mathcal{V}_j \in S_j$ if and only if it lies in the *convex hull* (Pelayo 2014) of $\mathcal{V}_j$ on G. There are several formulations of convex hulls of a subset of vertices $\mathcal{V}_j$ on G; here we propose to use the *geodesic-convex hull* on G, because of its natural relationship to the simplex and convex hull band depth in $\mathbb{R}^n$. The geodesic-convex set of vertices on a graph is a set of vertices that is closed under geodesic paths (all geodesic paths between all vertices in the set are contained in the set). The convex hull of a set $\mathcal{V}_j$, referred to as a j-simplex, is the smallest geodesic-convex set that contains $\mathcal{V}_j$ (and hence can be thought of as the *geodesic closure* of $\mathcal{V}_j$). We denote the convex hull of $\mathcal{V}_j$ by $H[\mathcal{V}_j]$.

To define band depth, we consider selecting j vertices independently from a probability distribution over the vertex set V given by $\text{Prob}_V(v)$ where $v \in V$. From these vertices we form $\mathcal{V}_j \in \mathcal{S}_j$. We can now ask if a vertex v falls inside the convex hull formed by our random selection of vertices, where the probability of this event is the product of the aforementioned vertex probabilities (by the independence assumption). Once in place we can define the *graph-simplex band depth* of a vertex with respect to the j-simplex to be $v\text{BD}(v) = \text{Prob}(v \in H[\mathcal{V}_j])$, where $\mathcal{V}_j$ is a set of j independent samples taken from the probability distribution we have defined for vertices.

If the graph is finite, the depth of a vertex can be computed in closed form. The band depth of v can be expressed as the expected value of the characteristic function χ for v falling within (or belonging to) a random j-simplex. That is,

$$\begin{aligned} v\text{BD}(v) &= E_{\mathcal{V}_j \in \mathcal{S}_j}\left[\chi\left(v \in H[\mathcal{V}_j]\right)\right] \\ &= \sum_{\mathcal{V}_j \in \mathcal{S}_j} \chi\left(v \in H[\mathcal{V}_j]\right) \prod_{v_m \in \mathcal{V}_k} \text{Prob}_V(v_m). \end{aligned} \tag{6}$$

This form also reveals that the proposed *graph-simplex band depth* is a more general formulation of *graph centrality* from graph theory (Freeman 1977). That is, the *centrality* of a vertex in a graph has been quantified as the number of geodesic paths that pass through that vertex (Freeman 1977), which corresponds to $j = 2$ and $\text{Prob}_V(v) = 1/|V|$ in Equation (6). Thus, graph-simplex band depth characterizes both the structure of the graph itself (and the centrality of points), as well as the probability distribution on the vertices.

The extension from vertices to paths proceeds as in the case of curves, with some additional technicalities. For this, we formulate a path on a graph as a mapping $p : \mathcal{I} \mapsto V$ over an index set $\mathcal{I} = [1, 2, \ldots, m]$ onto the vertex set V, and we use the notation $p(l)$ to denote the vertex of path p that is mapped from index $l \in \mathcal{I}$. The band formed by j paths sharing a common index set is the parameterized set of j-simplex bands formed by their corresponding vertices. Thus, we can index a set of j paths, $\mathcal{P}_j$, such that $\mathcal{P}_j(l) \in \mathcal{S}_j$ for all $l \in \mathcal{I}$.

The formulation for testing a path p against the band formed by a set of paths $\mathcal{P}_j$ that are parameterized over $\mathcal{I}$ is

$$p \in B[\mathcal{P}_j] \quad \text{iff} \quad p(l) \in H[\{p_1(l)), \ldots, p_j(l)\}] \quad \forall l \in \mathcal{I}. \tag{7}$$

The *band depth* of a path p is $\text{Prob}(p \in B[\mathcal{P}_j])$ where $\mathcal{P}_j$ is a set of j, independently drawn paths from the distribution $\text{Prob}(P = p)$. Similar to other notions of band depth, the path band depth can be computed as the expectation of the characteristic function of p being in the band of a randomly chosen set from the distribution of paths:

$$p\text{BD}(p) = E\left[\chi\left(p \in B(\mathcal{P}_j)\right)\right], \tag{8}$$

where $\mathcal{P}_j$ again represents a set of j, independently drawn paths from the distribution $\text{Prob}_{\mathcal{P}}(p)$ over all possible paths $\mathcal{P}$.

The expectation over the bands is approximated as a sample mean, from a random collection of j-sized subsets of an ensemble. In some cases, small sample sizes may interfere with the ability to estimate this expectation with sufficient accuracy to resolve differences in samples with low band depth. Thus, modified versions can either use a measure over an index set rather than a binary characteristic function (López-Pintado and Romo 2009) or relax the "for all" condition in Equation (7) to allow a certain number of vertices to fall outside the simplex band, as proposed by Whitaker et al. (2013).

The proposed formulation for band depth on paths requires $\mathcal{P}_j$ and p to share a common index $\mathcal{I}$, which is effectively a discrete parameterization. However, in most applications, paths are specified as sequences of vertices, without a corresponding index set. Thus, one of the contributions of this work is a strategy for forming these common index sets as part of the construction of bands for paths.

A common index set between a collection of paths establishes a *correspondence* between vertices on a path such that for each vertex on each path there is a (nonempty) set of corresponding vertexes on every other path. Because the paths may be of different lengths, the correspondences are not unique. However, we propose that the mapping from the index set to a path should be monotonic with respect to the sequence of the vertices on the path (order of the vertices in paths is respected), and thus, the correspondences are monotonic between every pair of paths.

The correspondence between a collection of paths is computed using an optimal matching strategy, similar to what is used for string matching in computer science and sequence alignment in biological protein analysis (Needleman and Wunsch 1970). The intuition behind this method is to assign correspondences such that the correspondences are *monotonic* and the overall sum of geodesic lengths between *corresponding* vertices along the paths is minimized. We first describe the method for finding correspondences between two paths. Given two paths p_l and p_m, an optimal correspondence is established by a pair of monotonic mappings from a common index set $\mathcal{I}$ to the paths, such that the distances between vertices are minimized. Thus, we are trying to find two mappings that minimize:

$$\left(\sum_{k \in \mathcal{I}} d_g\left(p_l(k), p_m(k)\right)\right) \tag{9}$$

where $p_l(k)$ is the vertex on path p_l that is mapped from the index $k \in \mathcal{I}$ and $d_g(,)$ denotes the geodesic distance between

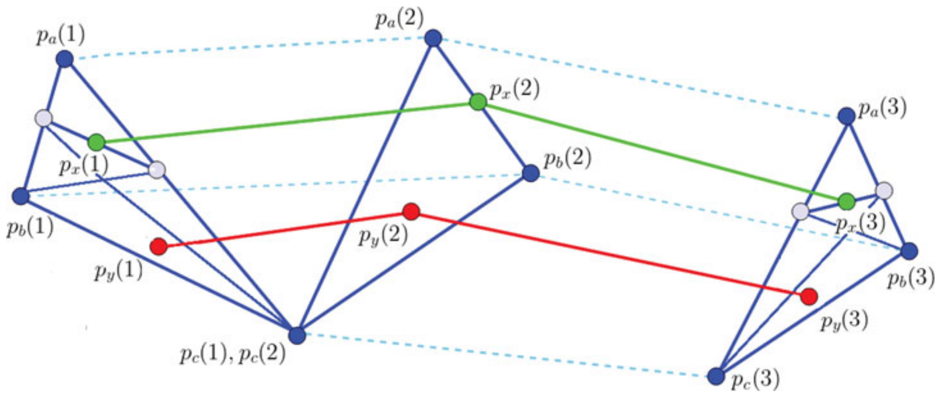

Figure 2. Band formed by three dashed paths on a complete graph whose edge weights are equal to Euclidean distance between vertices (only selected edges are drawn). The green path is completely contained within the band according to definition in Equation (7) while the red path falls completely outside the band. Solid blue edges constitute the geodesics connecting vertices within graph simplices.

two vertices. This formulation generalizes to collections of paths (> 2), by minimizing the sums of all pairs of distances among corresponding vertices in the collection of paths.

To find the correspondences among a set of paths, we use the classical method of dynamic programming (DP) on the matrix/tensor consisting of all possible correspondences—for example, the Needleman-Wunsch algorithm (Carrillo and Lipman 1988). All pairwise distances are organized in a tensor with an order that is the number of paths to be aligned. Thus, the number of distances considered in the optimization is $\prod_{l=1}^{j+1} |p_l|$, which grows exponentially with the number of paths forming the band (generally, the problem is NP-Hard (Just 2001)). There are existing efficient, approximate algorithms for large numbers of paths (Carrillo and Lipman 1988), but that issue is beyond the scope of this paper. For the results presented here we use $j \leq 3$ and rely on the basic (full enumeration of tensor) approach for optimization.

In Figure 2, we see a band formed by three paths— p_a, p_b, and p_c. Here, the elements from common index $\mathcal{I} = [1, 2, 3]$ are mapped to vertices on the graph from each of the paths. Path p_x is completely contained within the band as all of its vertices are part of a j-simplex formed by corresponding vertices that are mapped from the same element in $\mathcal{I}$ to p_a, p_b, and p_c. Similarly, we observe that no vertex from p_y is contained in any j-simplex. Also, two elements from $\mathcal{I}$ are mapped to a single vertex in p_c as it is shorter than the other paths. Once we are able to describe a band formed by a set of paths, we can generate order statistics on an ensemble of paths by calculating the path band depth of each member within the ensemble.

An ordering of the data based on *path band depth* readily yields a set of rank statistics. The median is the path with the highest probability of falling within a random band—(i.e., the *deepest* ensemble member). The 50% band consists of paths whose probabilities are in upper half percentile of all probabilities. The 100% envelope is formed by excluding the outliers. We define outliers (as in Sun and Genton (2012)): $p\mathrm{BD}(p) < p\mathrm{BD}(p_{\mathrm{median}}) - \alpha \times \left(p\mathrm{BD}(p_{\mathrm{median}}) - p\mathrm{BD}(p_{50\%})\right)$ where $p_{50\%}$ is the band depth value that splits the ensemble into equal parts, and $\alpha = 1.5$ is a typical value as found in the literature (Sun and Genton 2012). For the results shown in this article we used values of α in the range 2.4 to 3.7 to flag only the most nonrepresentative paths as outliers. Furthermore, we used the modified formulation of band depth (López-Pintado and Romo 2009), to resolve depth with sufficient accuracy to avoid ties.

By convention, data depth formulations in flat spaces (e.g., simplex depth in $\mathbb{R}^n$) are considered desirable if they demonstrate a set of properties that are consistent with classical methods on certain classes of distributions. For instance, Zuo and Serfling (2000) proposed affine invariance, maximal depth around a point of symmetry, monotonic fall off with distance from a central point, and zero depth for points at infinity. While some of these properties have yet to be developed for general graph structures, in the appendix we prove the asymptotic depth property for points at infinity for vertices and paths.

4. Path Boxplot Visualization

Here we develop a visualization for the proposed analysis in a manner similar to what has been proposed for functions, contours, and curves (Sun and Genton 2012; Genton et al. 2014; Whitaker et al. 2013; Mirzargar et al. 2014). The proposed visualization approach is motivated by the classical whisker plot or boxplot, and relays a display of the median, 50% band, 100% band, and outliers for graph-based path ensembles. Figure 3(a) shows a synthetically generated path ensemble with each path drawn using a random color. Figure 3(c) and 3(d) shows two variations of our proposed visualization described next.

We render the visualizations in a way that it describes rank statistics of the distribution or ensemble. We first establish the placement of vertices and edges either intrinsically or via a layout algorithm (Gibson et al. 2013). Next, we use color and width/thickness on edges and vertices to represent their rank. The paths in the 100% band are drawn thickest in light blue. The paths in the 50% band are drawn using a thinner dark blue stroke on top of the thicker light blue band. This drawing of the thinner dark blue stroke *over* the thicker light blue stroke is done to indicate that the path in the 50% band is contained within the 100% band as well. Continuing this strategy, the *median* path is drawn using a thin yellow stroke drawn over a thicker dark blue stroke, which in turn, is drawn over the thickest light blue stroke. To signify that the outlier paths lie outside even the 100% envelope, they are drawn using only a thin red stroke. Figure 3(c) shows a version of the path boxplot that uses the described encoding for paths. A variation of this approach as seen in Figure 3(d) where

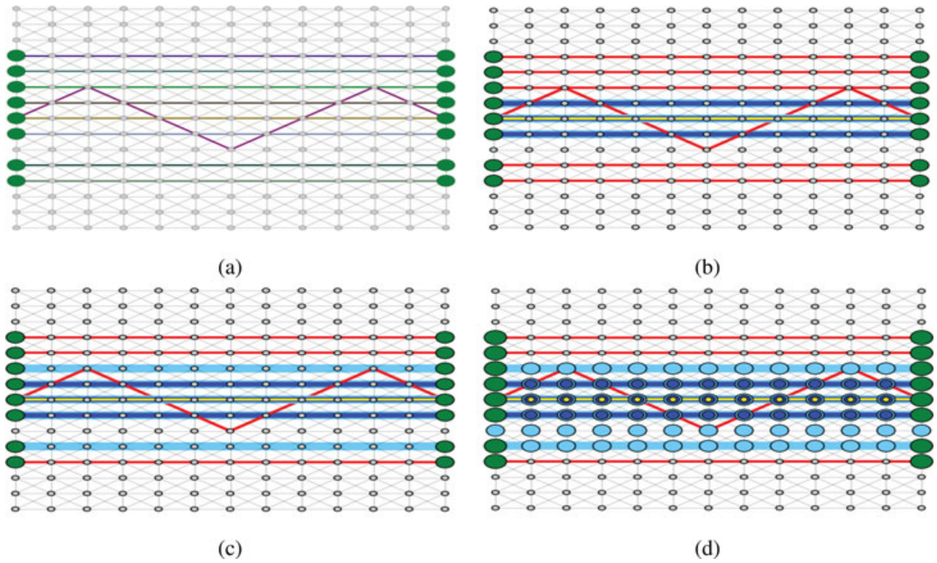

Figure 3. Synthetic example 1. (a) A path ensemble with each path rendered with a random color. (b) Path boxplot using rank statistics based on sum of Fréchet distances. (c) Path boxplot based on path band depth (visualization *without* vertex encoding). (d) Path boxplot based on path band depth (visualization *with* vertex encoding).

the vertices are also encoded, based on their position, in addition to the edges in the graph. Vertices that are not part of the convex hull formed by any set of corresponding vertices between paths are drawn as small gray circles. Vertices that are in the convex hulls formed by paths in the 50% band are drawn using a light blue circle. Analogous to the encoding for the paths, vertices in the convex hulls formed by paths in the 50% band are drawn using a deep blue circle contained within a larger light blue circle while vertices lying on the median path are marked with an additional yellow circle drawn within the deep blue circle which is itself contained within a light blue circle.

The sections that follow demonstrate applications of the proposed method on synthetic examples and datasets from applications in transportation and computer networks. We use the visualization approach *with* vertex encoding (as seen in Figure 3(d)) for all further path boxplot visualizations based on path band depth in this article except when vertices on the graph are not rendered (see Figure 5(b)).

5. Results

We begin by showing results for two synthetically generated path ensembles on graph. For these ensembles, we show path boxplot visualizations generated using rank statistics obtained by path band depth analysis, as well as, the Fréchet distance metric analysis (Eiter and Mannila 1994). For these path ensembles, the results were identical on replacing Fréchet metric by Hausdorff metric, and therefore we show only one of these methods. When using a distance metric, we rank each path using the sum of its distances from all the other paths in the ensemble. Hence, the path that *minimizes* this sum is identified as the median. Note that this is different from path band depth where the median path has *maximum* depth. The underlying graph in both our examples is associated with a regular, *diagonal grid* (constructed from including diagonals in a conventional, structured quadrilateral grid).

For the first of these examples (see Figure 3), we generate an ensemble of 20 paths by sampling with replacement from a set of straight paths (all vertices in path have same ordinate) spanning the horizontal extent of the grid. The ordinate of each path comes from a random variable associated with a normal distribution centered at central ordinate of the grid. We complete the ensemble by adding a simulated outlier in the form of a zigzag path (see Figure 3(a)). In Figure 3(b), we see the path boxplot visualization of Fréchet distance-based depth. Figure 3(c) and 3(d) shows two versions of path boxplot of our path band depth analysis. In this simple example, we see that the result from path band depth analysis is very similar to distance metric-based analysis with both approaches identifying the zigzag and peripheral paths as outliers.

We now present an example where distance metric-based methods fail to detect the general structure (median) and anomalous path (outlier) in an ensemble. Further, we see that path band depth analysis is able to correctly make this determination by capturing the nonlocal correlations in the path ensemble. Here, we produce an ensemble of 20 straight paths spanning the grid's horizontal extent, starting and ending at vertices with same ordinate (see Figure 4). In this case, however, each path is required to undergo flips when traversing the flip regions as seen in Figure 4(a). The vertex within each zone where the flip occurs is chosen uniformly from among the vertices in each zone. We add a simulated outlier to this ensemble in the form of a path with no flips (Figure 4(a)). In this case, we see that the distance-based metrics (Figure 4(b)) identify the simulated outlier as the median (most representative) while the path band depth method (Figure 4(c)) selects one of the randomly sampled paths as the median. The simulated outlier is *closest* to other paths with regard to the distance metrics while identified as an outlier by the path band depth analysis.

5.1. *Transportation Networks*

We used publicly available road data from OpenStreetMaps (OSM) (Haklay and Weber 2008) for a randomly chosen region in Los Angeles, California. Figure 5(a) shows a part of the road

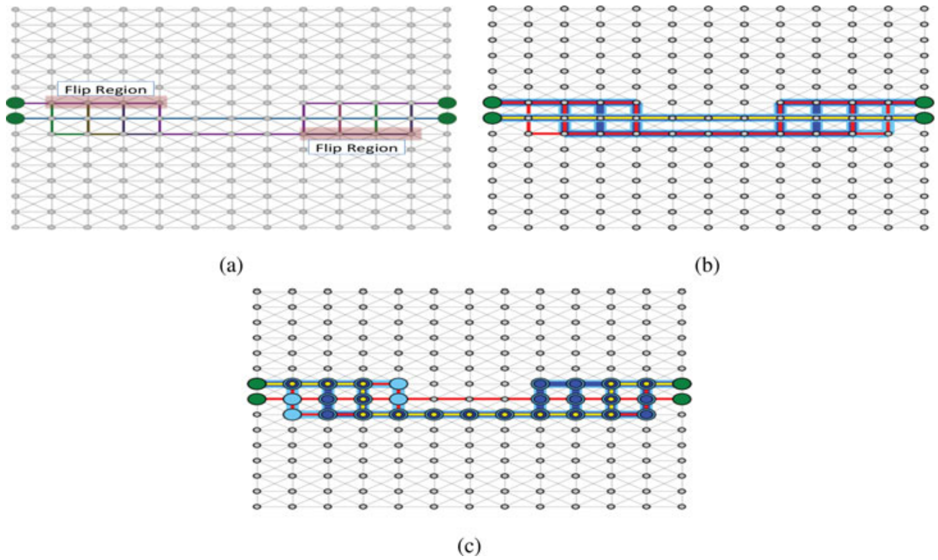

Figure 4. Synthetic example 2. (a) A path ensemble with each path rendered with a random color. (b) Path boxplot using order statistics based on sum of Fréchet distances. (c) Path boxplot based on path band depth.

graph overlaid on a map. We used *expected travel time* between the two adjacent vertices, obtained by querying the open source routing engine Gosmore, as the weight of each edge. Travel time along a short road segments can be modeled using a normal distribution (He et al. 2002). We obtained an ensemble of 20 paths between two random vertices by repeatedly finding the lowest cost path on graph whose edge weights were picked, after each iteration, from a normal distribution centered at the expected travel time for that edge.

For visualizing the paths, we use the geographical coordinates of the vertices on the road graph for layout. A map, also based on OSM data, is provided in the background for context

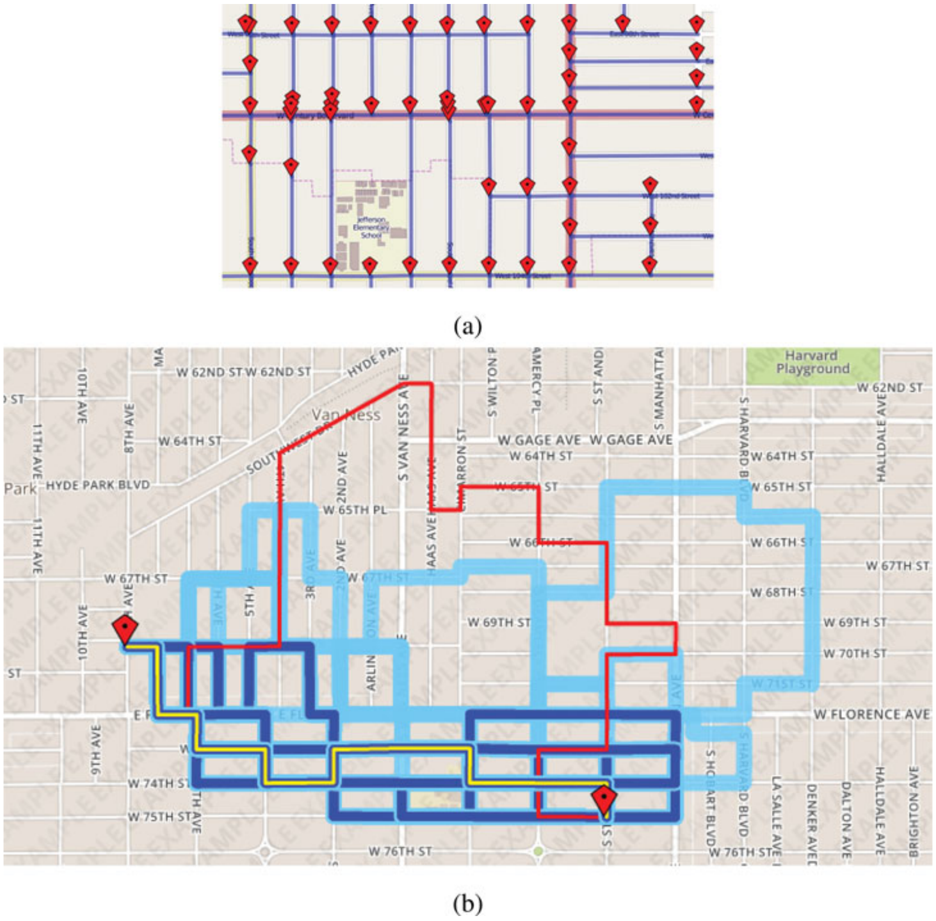

Figure 5. Road network. (a) A section of the road graph overlaid on a map representing actual spatial embedding of vertices and edges. (b) Path boxplot for an ensemble of paths on a road network.

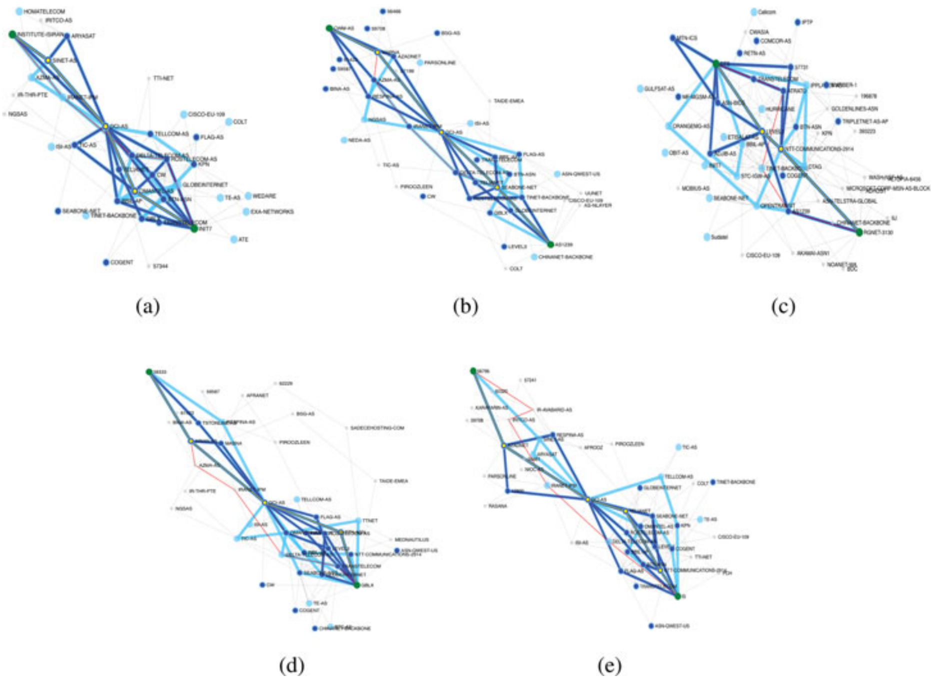

Figure 6. Outlier paths on AS graph: (a) Class 1 outlier: No unique vertices/edges in the outlier path. (b) Class 2 outlier: One unique edge, no unique vertices in the outlier path. (c) Class 3 outlier: Outlier appears to be one hop bypassing a more normal route. (d) Class 4 outlier: Outlier is two hops around a more normal route. (e) Class 5 outlier: Outlier takes several hops around the usual path.

in accordance to the common practice for viewing geographical routes (see Figure 5(a) and 5(b)). To have the overlaid paths align with underlying roads on the map and also be feasible with regard to traffic restrictions, we used the Gosmore routing service to obtain the geographic coordinates of the spatial path drawn between every pair of adjacent vertices along each paths in our ensemble. This is necessary for drawing road segments that are curved or where the direct connection between two vertices is illegal according to local traffic rules.

A path boxplot of a path ensemble on a road graph is shown in Figure 5(b). The most representative path or the median path seen here can be useful when the requirement is to select a particular path from a collection of paths on a road graph. For instance, a median path would be a good choice of a path that affords quick access to a number of alternate paths, which would be useful in situations involving high traffic conditions or blockages. The path boxplot would also find utility for planning bicycle corridors (Evans et al. 2012, 2013).

5.2. Computer Networks (Autonomous Systems)

We used a subset of the AS graph as well as path ensembles of packets traveling between AS's on that graph from a set of path snapshots seen from the Oregon Routeviews server. For clarity, we filtered out vertices in the graph that did not lie on a geodesic between any pair of vertices in the path ensemble. Additionally in the visualization we only include a single geodesic between all pairs of vertices in the ensemble. For graph layout in 2D, we modify the force directed model in Fruchterman and Reingold (1991) by including an extra repulsion between the vertices at the two endpoints, so that they are placed at nearly opposite ends of the layout. Also, the charge/repulsion on each vertex is made proportional to its degree for avoiding congestion near high degree vertices.

We looked at several destinations that had significant variations in their paths throughout the year. Visualizations of a few of these ensembles can be seen in Figure 6. Looking at a selection of these ensembles, there are some special cases identified as outliers. Figure 6(a) shows an outlier where the outlying path is of the same cardinality as the median path and does not contain any unique vertices or edges causing it to be undetectable by common heuristic methods used to analyze network traffic. Other cases include Figure 6(b)–6(e) where the outlier path bypasses other paths through unique edges or vertices. Cases where they depart near one of the endpoints, such as 6(e), may be relatively straightforward for the operators of those edge networks to detect, as their own routers will directly see the change in where traffic enters or exits their networks. Cases such as Figure 6(c), however, exhibit changes in networks that may not be directly visible from the endpoints, and yet affect the overall behavior of traffic to/from these endpoints. These are cases that can be particularly difficult to discover and diagnose; a path boxplot can aid operators in assessing such cases.

6. Conclusion and Future Work

Assigning centrality based ordering for an ensemble of paths is useful in many applications. Although robust band depth-based methods for calculating order statistics have been recently introduced for various kinds of ensembles on a continuous domain, they cannot be employed in cases where the ensemble members are described on a graph. We identify the challenges in extending this approach to paths on a graph and present a solution in the form of a novel notion of depth denoted as path band depth. A visualization scheme based on this new notion of depth called path boxplot is also introduced. This article demonstrates its utility to help understand the overall structure of the ensemble using synthetic data as well as data from two real application areas, path ensembles on autonomous system (AS) graphs and on road graphs.

While being a robust method for generating order statistics for path ensembles, the proposed analysis is computationally intensive due to its combinatorial nature. The topology of the underlying graph as well as the density of its edges also effect the computation time by a constant factor. A practical approach to deal with larger ensembles (with large number of paths) is to trade running time for an approximate solution by randomly selecting a subset from the set of all possible bands as suggested in López-Pintado and Romo (2009). In the case of ensembles with long paths, skipping vertices in the description of the paths may also provide an acceptable compromise between accuracy and performance. Developing a heuristic for skipping vertices in large ensembles of long paths to achieve an optimum trade-off between running time of analysis and the quality of the *solution* would be an interesting avenue for future work. It would also be interesting to explore the application of path boxplot in other areas such as in mobile ad hoc networks, which can be modeled as a graph with dynamic topology (Molnár et al. 2011) and in molecular dynamics, to identify a most representative path as an alternative to computing the mean statistics for the ensemble members.

Appendix

Let graph $G = (V, E, W)$ be an infinite graph (i.e., $|E|$ and $|V|$ are infinite), and W are positive edge weights, as in Section 3. For $\{m, n\} \in V$, we say that $v \in V$ is *geodominated* by $\{m, n\}$ if v lies on some $m - n$ geodesic (Pelayo 2014). If $\mathcal{S} \subset V$, we say *geodesic closure* $I[\mathcal{S}]$ is the union of all vertices geodominated by any pair of vertices in $\mathcal{S}$. We recursively define $I^{k+1}[\mathcal{S}] = I[I^k[\mathcal{S}]]$ with $I^0[\mathcal{S}] = \mathcal{S}$. The geodesic iteration number of $\mathcal{S}$, denoted $\text{gin}(\mathcal{S})$, is the smallest positive integer n such that $I^n[\mathcal{S}] = I^{n+1}[\mathcal{S}]$ (Cáceres et al. 2005). The geodesic closure or convex hull of $\mathcal{S}$ is denoted $H[\mathcal{S}]$. We rely on two other definitions that build on geodesic iterations. We define the diameter of a set of vertices $\mathcal{S} \subset V$, as $D(\mathcal{S}) = \max\{d_g(u, v) : u, v \in \mathcal{S}\}$.

For the following theorems, we consider graphs with bounded geodesic iterations. Thus, we say that a graph has bounded geodesic iteration number for sets of size j (denoted B-gin-j) if and only if there exists k such that $\max_{\mathcal{V}_j \in \mathcal{S}_j}\{\text{gin}(\mathcal{V}_j)\} \leq k$, where $\mathcal{V}_j = \{\mathcal{S} \subset V : |\mathcal{S}| = j\}$ and $\mathcal{S}_j$ denotes the set of all subsets of V of size j.

The asymptotic properties of depth on graphs will depend on the nature of the probability distributions on vertices and paths. We say that $\text{Prob}_V(v)$ is a *transient probability distribution* over the vertices, $v \in V$, if and only if there exists a vertex $v_c \in V$ such that $\text{Prob}(v) \to 0$ as $d_g(v, v_c) \to \infty$. Likewise, for paths, let $\mathcal{P}$ be the set of all possible paths in G. $\text{Prob}_{\mathcal{P}}(p)$ is a transient distribution over paths, if and only if there exists a finite path $p_c \in \mathcal{P}$ such that $\text{Prob}_{\mathcal{P}}(p) \to 0$ as $d_{\text{H}}(p, p_c) \to \infty$ (where d_{H} is the network Hausdorff distance) and $\text{Prob}_{\mathcal{P}}(p) \to 0$ as $|p| \to 0$.

Lemma A.1. $D(I^k[\mathcal{V}_j]) \leq 2^k \times D(\mathcal{V}_j)$ for all $k \geq 0$.

Proof. We prove this by induction on geodesic iterations. The base case $k = 0$ is trivial. Assume $D(I^k[\mathcal{V}_j]) = 2^k \times D(\mathcal{V}_j)$ for some $k \geq 0$. Now consider a vertex $a \in \{I^{k+1}[\mathcal{V}_j] - I^k[\mathcal{V}_j]\}$. It must lie on a geodesic between two points in $I^k[\mathcal{V}_j]$, and thus its distance to the nearest point in $I^k[\mathcal{V}_j]$ is at most $D(I^k[\mathcal{V}_j])/2$, and its distance to any point in $I^k[\mathcal{V}_j]$ is bounded by $3D(I^k[\mathcal{V}_j])/2$. Thus, $d_g(a, b) \leq 2 \times D(I^k[\mathcal{V}_j]) \; \forall \; b \in I^{k+1}[\mathcal{V}_j]$, and hence we have $D(I^{k+1}[\mathcal{V}_j]) \leq 2 \times D(I^k[\mathcal{V}_j])$, which completes the proof. □

Theorem A.1. For an infinite B-gin $-$ j graph G and a transient distribution Prob_V over its vertices, the graph-simplex band depth of a vertex v converges to zero as its distance from v_c tends to infinity.

Proof. The proof is by contraction. For a vertex v let there exist $\epsilon > 0$ such that $\text{Prob}[v \in H(\mathcal{V}_j)] \geq \epsilon$ as $d_g(v, v_c) \to \infty$. Because of the nonzero probability, there exists a set $\mathcal{V}_j \in \mathcal{S}_j$ and $\gamma \in \mathbb{R}^+$ such that $d_g(v', v_c) < \gamma \; \forall \; v' \in \mathcal{V}_j$. This implies that $D(\mathcal{V}_j \cup v_c) < 2\gamma$. This and the B-gin-$j$ property bounds the closure of $\mathcal{V}_j \cup v_c$. Thus, $D(H[\mathcal{V}_j \cup v_c]) < 2^{\mathcal{M}+1}\gamma$ where $\mathcal{M}$ is the gin bound. We can therefore enclose the convex closure of $\mathcal{V}_j$ in a finite-sized ball around v_c. Therefore, $v \in H(\mathcal{V}_j)$ implies $d_g(v, v_c) < 2^{\mathcal{M}+1}\gamma$. This contradicts $d_g(v, v_c) \to \infty$. □

Lemma A.2. For any path $p \in \mathcal{P}$ and index set $\mathcal{I}$, let $l = \text{argmax}_{i \in \mathcal{I}} \, d_g(p(i), p_c(i))$. Then, $d_{\text{H}}(p, p_c) \leq d_g(p(l), p_c(l)) \leq d_{\text{H}}(p, p_c) + A$ where $A \in \mathbb{R}^+$.

Proof. From the definition of Hausdorff distance, we have $d_{\text{H}}(p, p_c) \leq d_g(p(l), p_c(l))$. Now, let m and n be indices such that $d_g(p(m), p_c(n)) = d_{\text{H}}(p, p_c)$. From triangle inequality, we see $d_g(p(l), p_c(l)) \leq d_{\text{H}}(p, p_c) + d_g(p(l), p(m)) + d_g(p_c(l), p_c(n))$. As path lengths are bounded in the transient distribution $\mathcal{P}$, we have that $d_g(p(l), p(m))$ and $d_g(p_c(l), p_c(n))$ are bounded. Hence, $d_g(p_i(l), p_c(l)) \leq d_{\text{H}}(p, p_c) + A$ where $A \in \mathbb{R}^+$. □

Theorem A.2. For a B-gin-j graph, G and a transient distribution $\text{Prob}_{\mathcal{P}}$ of paths in G, the path band depth of an arbitrary path p converges to zero as $d_{\text{H}}(p, p_c) \to \infty$.

Proof. The proof is by contradiction. For a path p let there exist $\epsilon > 0$ such that $\text{Prob}(p \in B(\mathcal{P}_j)) \geq \epsilon$ as $d_{\text{H}}(p, p_c) \to \infty$. Because $\text{Prob}(p \in B(\mathcal{P}_j)) \geq \epsilon$, there exists $K \in \mathbb{R}^+$ and $\mathcal{P}_j$ such that $d_{\text{H}}(p', p_c) \leq K \; \forall \; p' \in \mathcal{P}_j$. For an index set (i.e., correspondence), let l be the index of the vertex $p(l)$ such that $l = \text{argmax}_i \, d_g(p(i), p_c(i))$. Now we consider the vertex set $\mathcal{P}_j(l) = \{p_1(l), \ldots, p_j(l)\}$. From Lemma A.2 and because the Hausdorff distance of this set of paths to p_c is bounded, the distance between corresponding vertices, $d_g(p_i(l), p_c(l))$ where $p_i \in \mathcal{P}_j$, is bounded by $K + A$ where $A \in \mathbb{R}^+$. Let $\|p_c\| = L$, and we have $D(\mathcal{P}_j(l)) \leq 2(K + A) + L$. From Lemma A.1 and the B-gin-j property, we have $D(H(\mathcal{P}_j(l))) \leq 2^{\mathcal{M}}(2(K + A) + L)$ where $\mathcal{M}$ is the gin bound. Therefore, we have $d_g(p(l), p_c(l)) \leq 2^{\mathcal{M}}(2(K + A) + L)$. From Lemma A.2, this means that $d_{\text{H}}(p, p_c) \leq 2^{\mathcal{M}}(2(K + A) + L)$, which contracts the assumption that $d_{\text{H}}(p, p_c)$ is unbounded. □

Supplementary Materials

Python code for path boxplot: All code and datasets are included in the *PathBoxplot.zip* file. The root folder contains scripts to perform analysis and generate visualizations for synthetic, AS network, and road network for all examples in the manuscript. Instructions for running the experiments are included in *readme.txt* located in the root folder.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grants IIS-1212806 and CNS-1302688.

References

Agarwal, P. K., Avraham, R. B., Kaplan, H., and Sharir, M. (2014), "Computing the Discrete Fréchet Distance in Subquadratic Time," *SIAM Journal on Computing*, 43, 156–167. [244]

Apaydin, M. S., Brutlag, D. L., Guestrin, C., Hsu, D., Latombe, J.-C., and Varma, C. (2003), "Stochastic Roadmap Simulation: An Efficient Representation and Algorithm for Analyzing Molecular Motion," *Journal of Computational Biology*, 10, 257–281. [243]

Butler, K., Farley, T., Mcdaniel, P., and Rexford, J. (2010), "A Survey of bgp Security Issues and Solutions," *Proceedings of the IEEE*, 98, 100–122. [243]

Cáceres, J., Márquez, A., Oellermann, O. R., and Puertas, M. L. (2005), "Rebuilding Convex Sets in Graphs," *Discrete Mathematics*, 297, 26–37. [251]

Carrillo, H., and Lipman, D. (1988), “The Multiple Sequence Alignment Problem in Biology,” *SIAM Journal on Applied Mathematics*, 48, 1073–1082. [247]

Eiter, T., and Mannila, H. (1994), “Computing Discrete Fréchet Distance,” Technical Report CD-TR 94/64, Information Systems Department, Technical University of Vienna. [244,248]

Evans, M. R., Oliver, D., Shekhar, S., and Harvey, F. (2012), “Summarizing Trajectories Into k-primary Corridors: A Summary of Results,” in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 454–457. [244,250]

——— (2013), “Fast and Exact Network Trajectory Similarity Computation: A Case-study on Bicycle Corridor Planning,” in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, ACM, pp. 9. [243,244,250]

Freeman, L. C. (1977), “A Set of Measures of Centrality Based on Betweenness,” *Sociometry*, 40, 35–41. [246]

Fruchterman, T. M. J., and Reingold, E. M. (1991), “Graph Drawing by Force-directed Placement,” *Software: Practice and Experience*, 21, 1129–1164. [250]

Genton, M. G., Johnson, C., Potter, K., Stenchikov, G., and Sun, Y. (2014), “Surface Boxplots,” *Stat Journal*, 3, 1–11. [247]

Gibson, H., Faith, J., and Vickers, P. (2013), “A Survey of Two-Dimensional Graph Layout Techniques for Information Visualisation,” *Information Visualization*, 12, 324–357. [247]

Haklay, M., and Weber, P. (2008), “Openstreetmap: User-Generated Street Maps,” *IEEE Pervasive Computing*, 7, 12–18. [248]

He, R. R., Liu, H. X., and Kornhauser, A. L. (2002), “Temporal and Spatial Variability of Travel Time,” Paper UCI-ITS-TS-02, 14, Center for Traffic Simulation Studies. [249]

Hua, M., and Pei, J. (2010), “Probabilistic Path Queries in Road Networks: Traffic Uncertainty Aware Path Selection,” in *Proceedings of the 13th International Conference on Extending Database Technology*, pp. 347–358. [243]

Just, W. (2001), “Computational Complexity of Multiple Sequence Alignment with Sp-score,” *Journal of Computational Biology*, 8, 615–623. [247]

Kharrat, A., Popa, I. S., Zeitouni, K., and Faiz, S. (2008), “Clustering Algorithm for Network Constraint Trajectories,” in *Headway in Spatial Data Handling*, New York: Springer, pp. 631–647. [244]

Liu, R. Y. (1990), “On a Notion of Data Depth Based on Random Simplices,” *The Annals of Statistics*, 18, 405–414. [244,245]

López-Pintado, S., and Romo, J. (2009), “On the Concept of Depth for Functional Data,” *Journal of the American Statistical Association*, 104, 718–734. [244,245,246,247,251]

López-Pintado, S., Sun, Y., Lin, J., and Genton, M. (2014), “Simplicial Band Depth for Multivariate Functional Data,” *Advances in Data Analysis and Classification*, 8, 1–18. [244,245]

Mirzargar, M., Whitaker, R., and Kirby, R. (2014), “Curve Boxplot: Generalization of Boxplot for Ensembles of Curves,” *IEEE Transactions on Visualization and Computer Graphics*, 20, 2654–2663. [244,245,247]

Molnár, M., and Marie, R. (2011), “Stability Oriented Routing in Mobile ad hoc Networks Based on Simple Automatons,” in *Mobile Ad-Hoc Networks: Protocol Design*, ed. X Wang, Rijeka, Croatia: Intech, pp. 363–390. [251]

Needleman, S. B., and Wunsch, C. D. (1970), “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins,” *Journal of Molecular Biology*, 48, 443–453. [246]

Pelayo, I. M. (2014), *Geodesic Convexity in Graphs* (Springer Briefs in Mathematics), New York: Springer. [246,251]

Sun, Y., and Genton, M. G. (2012), “Adjusted Functional Boxplots for Spatio-Temporal Data Visualization and Outlier Detection,” *Environmetrics*, 23, 54–64. [244,247]

Tukey, J. W. (1977), *Exploratory Data Analysis. Behavioral Science: Quantitative Methods*, Boston: Addison-Wesley. [244]

Von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., Van Wijk, J., Fekete, J.-D., and Fellner, Dieter, W. (2011), “Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges,” *Computer Graphics Forum*, 30, 1719–1749. [244]

Whitaker, R. T., Mirzargar, M., and Kirby, R. M. (2013), “Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles,” *IEEE Transactions on Visualization and Computer Graphics*, 19, 2713–2722. [244,245,246,247]

Wright, S. (1934), “The Method of Path Coefficients,” *The Annals of Mathematical Statistics*, 5, 161–215. [243]

Zuo, Y., and Serfling, R. (2000), “General Notions of Statistical Depth Function,” *Annals of Statistics*, 28, 461–482. [247]