Metric recovery from directed unweighted graphs

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

[1]MIT CSAIL

[2]MIT Dept. Mathematics

Summary

- Obtain directed unweighted graph from $x_i \in \mathbb{R}^d$ with edge $i \to j$ with probability $p_{ij} = h(|x_i - x_j|\varepsilon(x_i)^{-1})$.
- We can recover the radius function $\varepsilon(x_i)$ and density $p(x_i)$ given the graph and d.
- We show consistent recovery is possible up to isometric scaling if the vertex degree is $\omega(n^{2/(2+d)}\log(n)^{d/(d+2)})$.

Problem setup

We consider the following construction:

- $p(x)$: given probability density in $\mathbb{R}^d$.
- $\mathcal{X} = \{x_1, x_2, \dots\}$: latent coordinate points drawn independently from $p(x)$.
- $\varepsilon_n(x_i)$: radius function (may depend on $\mathcal{X}$).
- $h(d_{ij})$: connectivity kernel mapping $\mathbb{R}^+ \to [0,1)$ such that $\int_0^1 h(r)r^{d-1}dr > 0$ and $h(r) = 0$ for $r > 1$
- $G_n = (\mathcal{X}_n, E_n)$: unweighted directed graph with vertices $\mathcal{X}_n = \{x_1, \dots, x_n\}$ and edge $i \to j$ with probability $h(|x_i - x_j|\varepsilon^{-1}(x_i))$

Fix a large n. We consider the random directed graph model given by observing G_n. Under the assumptions $(\star)$ specified below, we solve:

Problem: Problem statement

Given G_n and d, form a consistent estimate of $p(x_i)$ and $|x_i - x_j|$ up to proportionality.

We make the following assumptions:

Definition: Assumption $(\star)$

- The density $p(x)$ is differentiable with $\nabla \log(p(x))$ bounded on a connected compact domain $D \subset \mathbb{R}^d$ with smooth boundary ∂D.
- There is a deterministic continuous function $\bar{\varepsilon}(x) > 0$ on D and scaling constants g_n with
$$g_n \to 0 \text{ and } g_n n^{\frac{1}{d+2}} \log(n)^{-\frac{1}{d+2}} \to \infty$$
so that a.s. in the draw of $\mathcal{X}$, $g_n^{-1}\varepsilon_n(x)$ converges uniformly to $\bar{\varepsilon}(x)$.
- The rescaled density functions $n\pi_{X_n}(x)$ are a.s. uniformly equicontinuous.

Some special cases are the following:

Definition: Special cases

- **k-nearest neighbor graphs**
- **ε-ball proximity graph**
- **Gaussian affinity:** $p_{ij} = \exp(-d_{ij}\sigma)$
- **Annulus:** $p_{ij} = 1$ iff $a < d_{ij} < b$

Theoretical results

We consider:

- $X_n(t)$: the simple random walk on G_n.
- $\pi_{X_n}(x)$: the stationary density of $X_n(t)$.

Our main result shows $\pi_{X_n}(x)$ converges to an explicit function of $p(x)$ and $\bar{\varepsilon}(x)$. Combining with an estimate on the out-degree of points in G_n allows us to recover density and scale. Let V_d be the volume of the unit d-ball and $NB_n(x)$ the neighbors of x in G_n.

Theorem: Main result

Given $(\star)$, a.s. in $\mathcal{X}$, we have
$$n\pi_{X_n}(x) \to c\frac{p(x)}{\bar{\varepsilon}(x)^2},$$
for $c^{-1} = \int p(x)^2\bar{\varepsilon}(x)^{-2}dx$.

Corollary: Density estimates

Assuming $(\star)$, we have a.s. in $\mathcal{X}$ that
$$\left(\frac{n^{\frac{d-2}{d}}}{cV_d^{2/d}g_n^2}\right)^{\frac{d}{d+2}} |NB_n(x)|^{\frac{2}{d+2}}\pi_{X_n}(x)^{\frac{d}{d+2}} \to p(x);$$
$$\left(\frac{1}{c^{d/2}V_d n^2 g_n^d}\right)^{\frac{1}{d+2}} |NB_n(x)|^{\frac{1}{d+2}}\pi_{X_n}(x)^{-\frac{1}{d+2}} \to \bar{\varepsilon}(x).$$

The main results follow by proving that the process $X_n(t)$ converges to an Itô process:

Theorem: Continuum limit of the walk

Under $(\star)$, as $n \to \infty$ a.s. in the draw of $\mathcal{X}$ the process $X_n(\lfloor t/h_n \rfloor)$ converges in $D([0,\infty), \bar{D})$ to the isotropic $\bar{D}$-valued Itô process $Y(t)$ with reflecting boundary condition defined by
$$dY(t) = \frac{\nabla p(Y(t))}{3p(Y(t))}\bar{\varepsilon}(Y(t))^2 dt + \frac{\bar{\varepsilon}(Y(t))}{\sqrt{3}}dW(t).$$

Empirical Results

Near perfect reconstruction

We demonstrate near-perfect reconstruction performance on simulated data. Our estimator is nearly indistinguishable from the naive metric ball estimator and substantially outperforms prior work of [1].

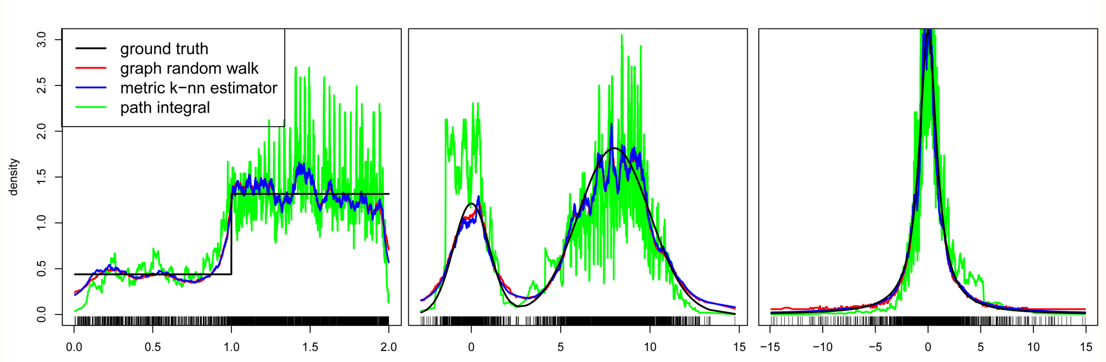

Figure: (red) = our method; (green) = path integral [1]; (blue) = metric k-nearest neighbor; (black) = ground truth

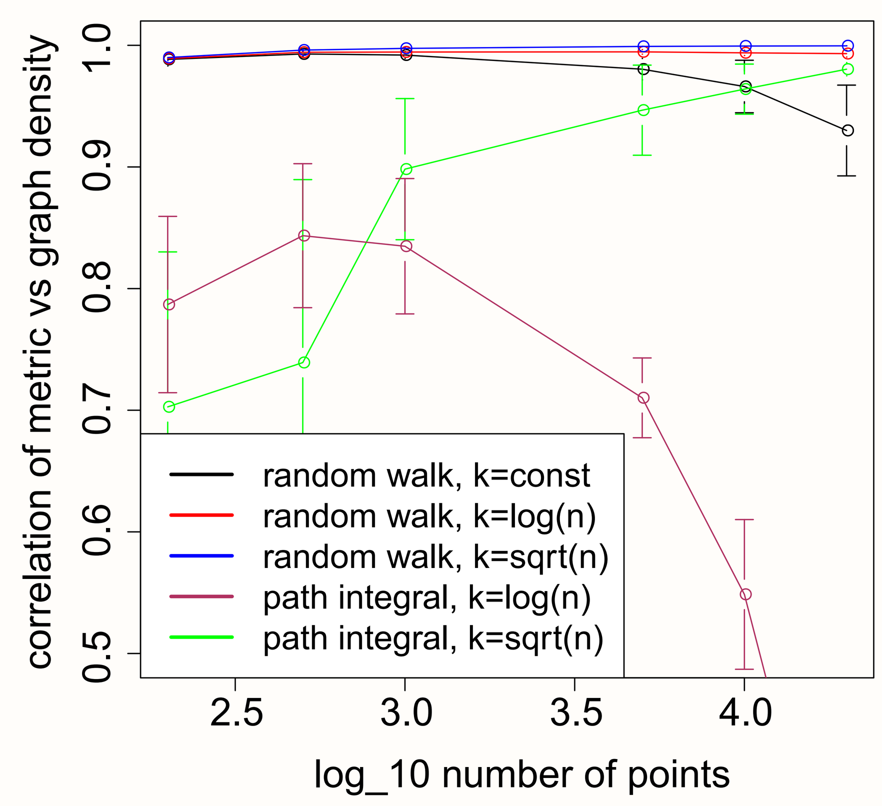

Figure: Accuracy vs sample and neighborhood size. (red, blue, black) = our estimator; (green, maroon) = path integral of [1].

Amazon co-purchasing data

We demonstrate useful embeddings on the Amazon co-purchasing dataset:

- edge from item $i \to j$ if item i is listed as purchased together often with j;
- density estimates correspond to sales rank;
- embedding reproduces product categories.

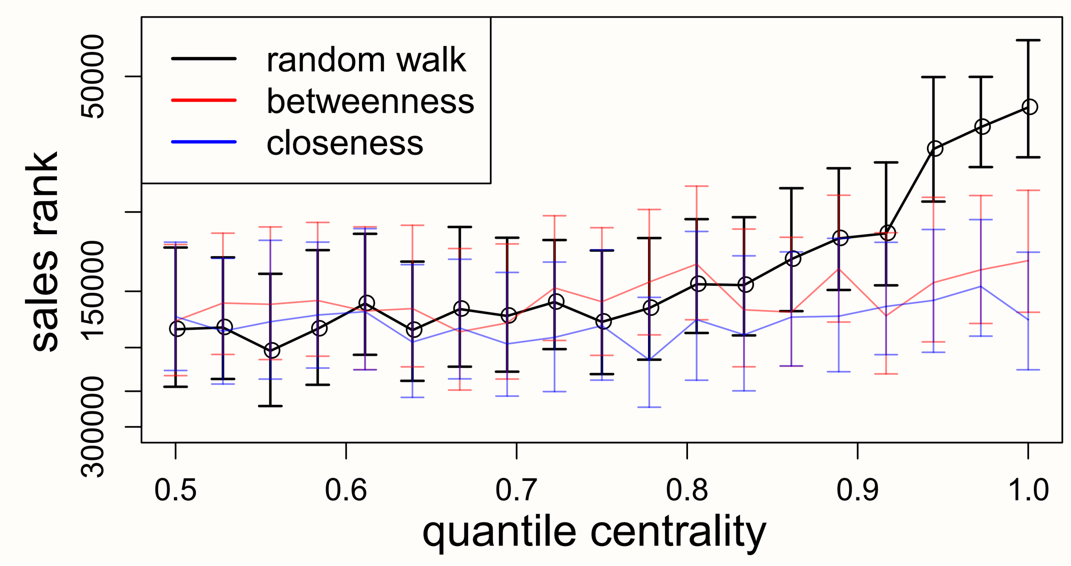

Figure: Density estimates in the graph correlate well with sales rank, unlike other measures of centrality.

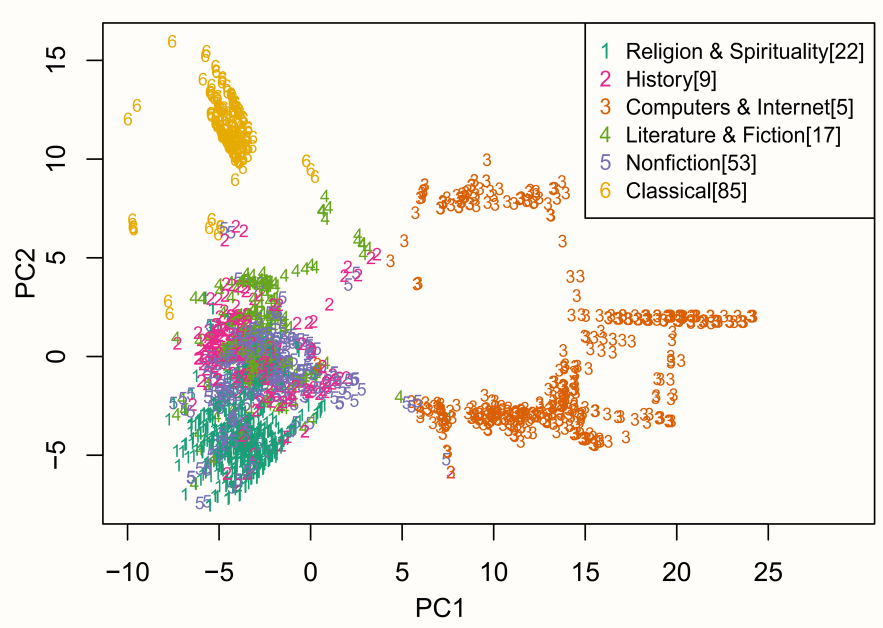

Figure: Embeddings from estimated distances recover separation between different product categories.

References

[1] U. Von Luxburg and M. Alamgir. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. NIPS 2013.

NIPS workshop on Networks, 13 December 2014