Graph-based Semi-Supervised and Active Learning for Edge Flows

Junteng Jia, Michael T. Schaub, Santiago Segarra and Austin R. Benson

jj585@cornell.edu, mschaub@mit.edu, segarra@rice.edu, arb@cs.cornell.edu

https://github.com/000Justin000/ssl_edge

Motivation & Problem Statement

Consider the problem of monitoring traffic flows in a region. Setting up sensors on all roads would provide accurate measurements, but is costly. Given traffic flow measurements on a subset of the roads, can we estimate the remaining flows?

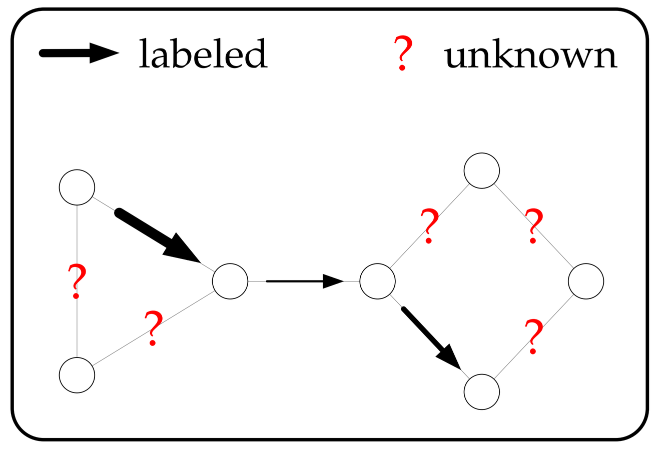
Problem statement

Given:

- a graph topology $G = (\mathcal{V}, \mathcal{E})$
- flows on a subset of the edges $\mathcal{E}^{\mathrm{L}}$

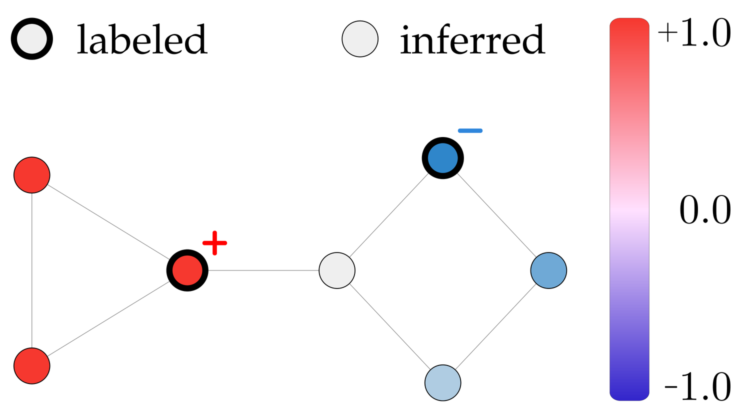
Infer:

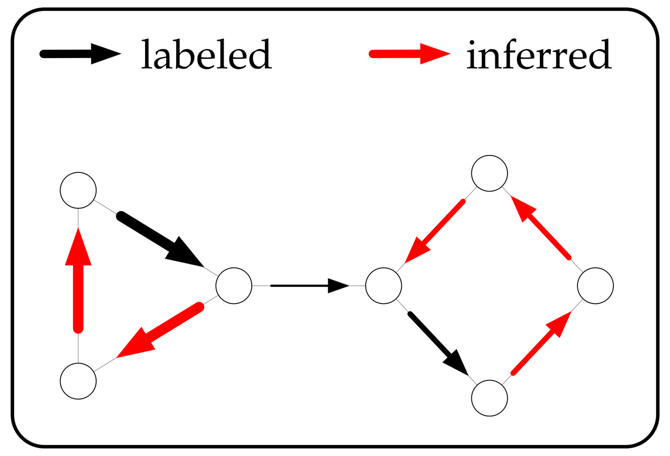
- unknown flows on $\mathcal{E}^{\mathrm{U}} \equiv \mathcal{E} / \mathcal{E}^{\mathrm{L}}$.



Key insight: a suitable learning assumption for edge flows.
Flow conservation – flows that enter/exit a node must balance.

Edge- *vs.* vertex-based semi-supervised learning





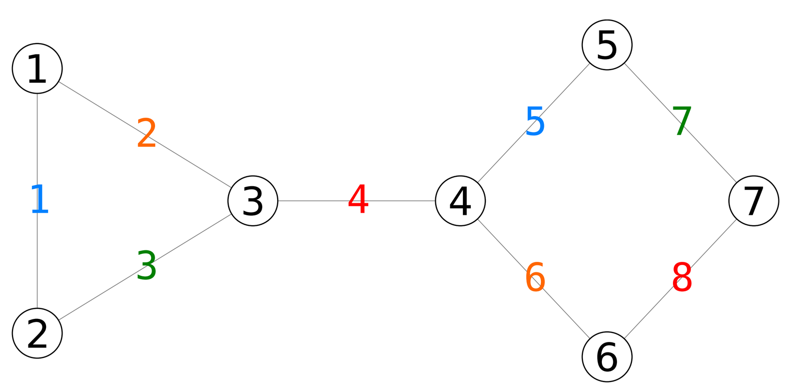
Vertex-based learning

- given some vertex labels
- impose *smoothness* assumption
- interpolate unknown vertices

Edge-based flow learning

- given some edge flows
- impose *flow conservation*
- infer unknown edge flows

Formulation & Inference Algorithm



$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

incidence matrix $\mathbf{B}$

- undirected graph $G = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = n$ and $|\mathcal{E}| = m$
- vertex label vector $\mathbf{y} \in \mathbb{R}^n$
- define (net) edge flow as alternating function $f \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$

$$f(i,j) = \begin{cases} -f(j,i), & \forall\, (i,j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases}$$

edge flow vector $\mathbf{f} \in \mathbb{R}^m$ with $\mathrm{f}_r = f(i,j)$ if $\mathcal{E}_r \equiv (i,j), i < j$

- vertex-edge incidence matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ (right panel)

Computations: edge *vs.* vertex-based learning

Vertex-based learning

- $\|\mathbf{B}^\intercal \mathbf{y}\|^2 = \sum_{(i,j)\in\mathcal{E}} (y_i - y_j)^2$ measures "unsmoothness"
- minimize sum-of-squares difference

$$\mathbf{y}^* = \arg\min_{\mathbf{y}} \|\mathbf{B}^\intercal \mathbf{y}\|^2 \qquad \text{s.t.} \quad y_i = \hat{y}_i, \quad \forall \mathcal{V}_i \in \mathcal{V}^{\mathrm{L}}.$$

Edge flow learning

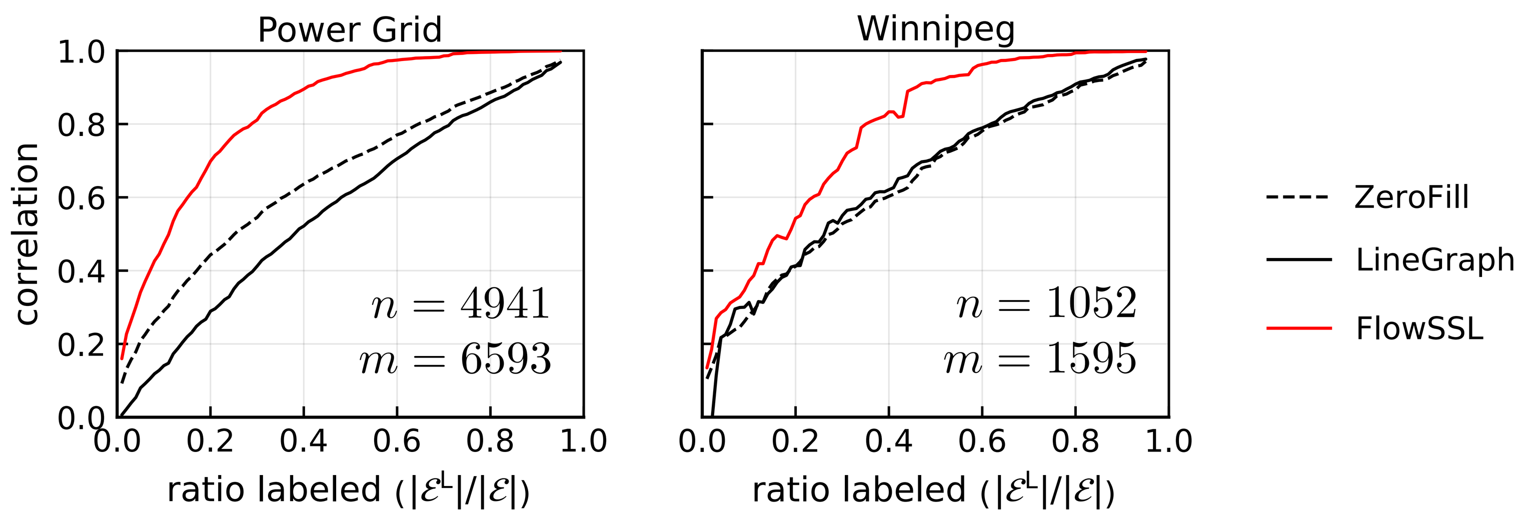
- $(\mathbf{Bf})_i$ measures the flow "divergence" on the i^{th} vertex
- minimize sum-of-square divergence, with regularization

$$\mathbf{f}^* = \arg\min_{\mathbf{f}} \|\mathbf{Bf}\|^2 + \lambda^2 \cdot \|\mathbf{f}\|^2 \qquad \text{s.t.} \quad \mathrm{f}_r = \hat{\mathrm{f}}_r, \quad \forall \mathcal{E}_r \in \mathcal{E}^{\mathrm{L}}.$$

- least-square solution (null space method $\mathbf{f} = \mathbf{f}^0 + \mathbf{\Phi}\mathbf{f}^{\mathrm{U}}$)

$$\mathbf{f}^{\mathrm{U}*} = \arg\min_{\mathbf{f}^{\mathrm{U}}} \left\| \begin{bmatrix} \mathbf{B\Phi} \\ \lambda \cdot \mathbf{I} \end{bmatrix} \mathbf{f}^{\mathrm{U}} - \begin{bmatrix} -\mathbf{Bf}^0 \\ 0 \end{bmatrix} \right\|^2 .$$

Empirical results & Reconstruction error bound



- Transforming flow learning problem with LineGraph to be solved as vertex label learning problem performs no better than consistently predicting zero.
- FlowSSL, our proposed semi-supervised edge-flow learning algorithm, outperforms the baselines by a large margin.

Theorem: Assume the ground truth flow $\hat{\mathbf{f}} = \mathbf{f} + \delta$, where $\mathbf{f}$ is a divergence free flow; and we have flow measurements on a subset $\mathcal{C}$ edges with cardinality at least $m - n + 1$. Denote the null-space of the incidence matrix as $\mathbf{V} = \mathtt{Null}(\mathbf{B})$. Then as the regularization parameter $\lambda \to 0$ in our method, the reconstruction error is bounded by $[\sigma_{\min}^{-1}(\mathbf{V}_{\mathcal{C},:}) + 1] \cdot \|\delta\|$, where $\sigma_{\min}(\cdot)$ is the smallest singular value of a matrix.

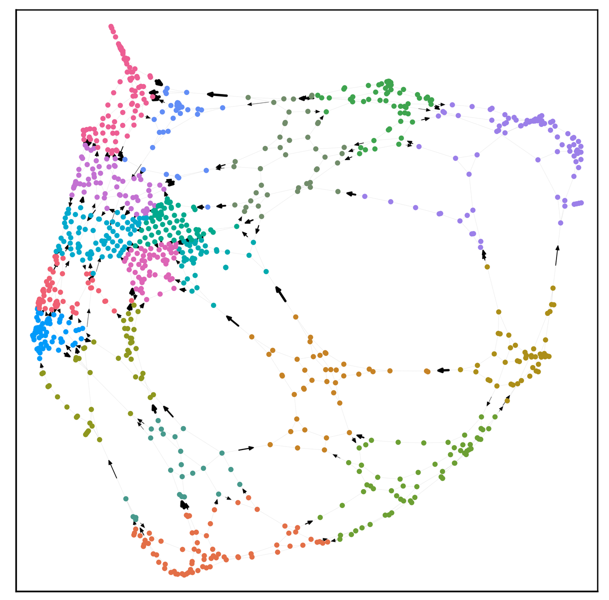
Active Learning Problem & Strategies

Goal: Select a set of edges $|\mathcal{E}^{\mathrm{L}}| = m^{\mathrm{L}}$ to minimize reconstruction error (optimal sensor deployment with a limited budget).

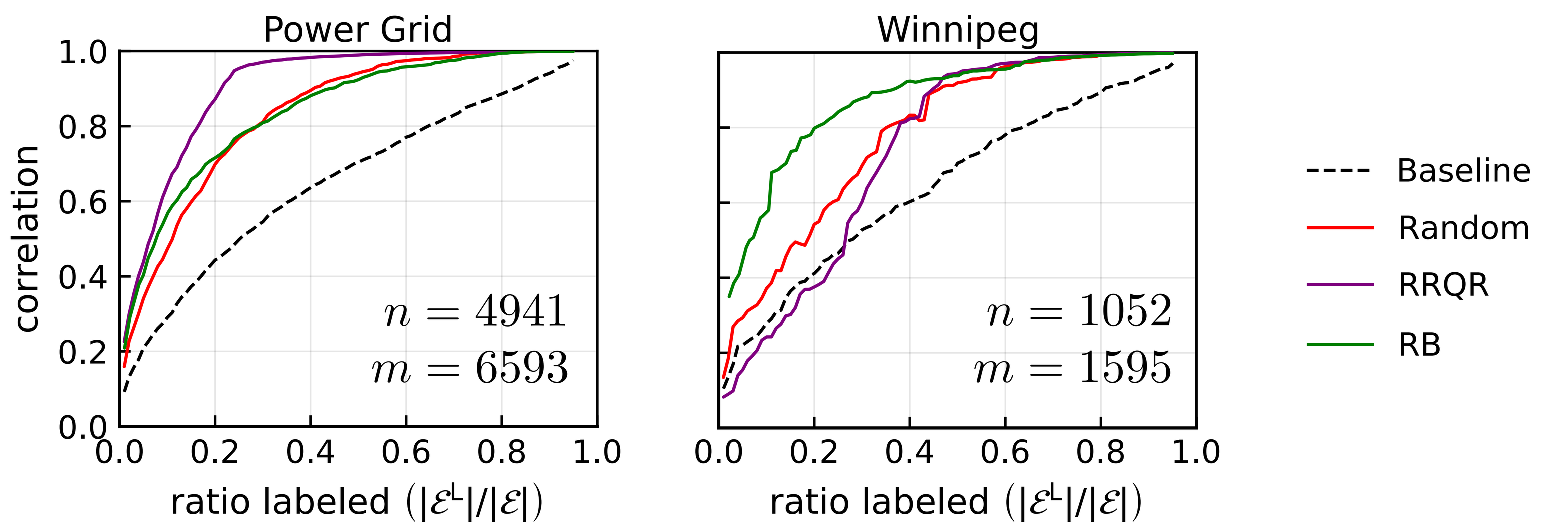
1. RRQR – minimize error bound
   - use rank revealing QR (RRQR) to select well conditioned rows

   $$\mathbf{V}_{\mathcal{C},:}^\intercal\, \Pi = Q\, [R_1 \; R_2]\,.$$

   - $\mathcal{E}^{\mathrm{L}}$ from leading columns of Π
2. RB – select bottleneck edges
   - capture global flow trends
   - recursively bisect (RB) & select edges that bridge clusters



black arrow: edges selected by RB



Findings: RRQR provides additional gains for approximately divergence-free flows (left), RB works well for flows with global trends (right).

This research was supported by NSF award DMS-1830274, ARO Award W911NF-19-1-0057, and European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 702410.