# Graph-based Semi-Supervised & Active Learning for Edge Flows

KDD2019


Junteng Jia
Cornell University
jj585@cornell.edu

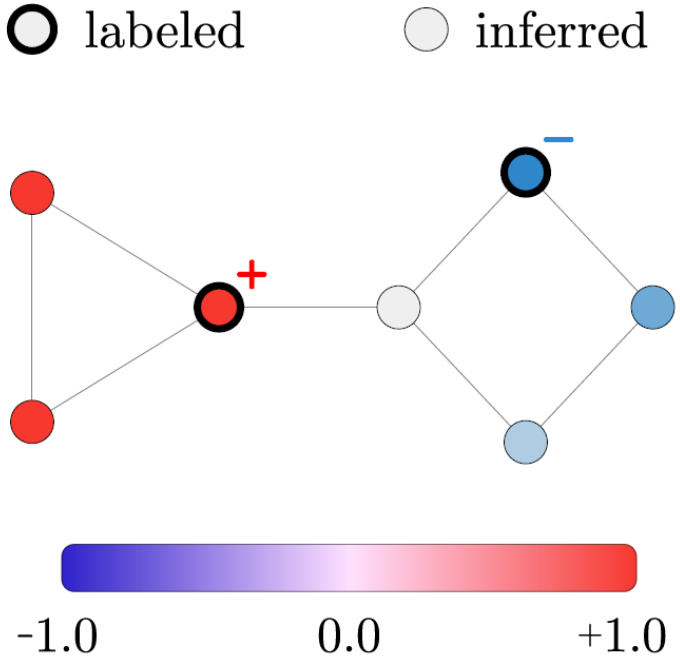
Michael T. Schaub
Massachusetts Institute of Technology
University of Oxford
mschaub@mit.edu

Santiago Segarra
Rice University
segarra@rice.edu

Austin R. Benson
Cornell University
arb@cs.cornell.edu


2019.11.6

# Graph-based SSL for Vertex Labels




Given the labels of a subset of vertices, and our goal is to find a label assignment of the unlabeled vertices such that the labels vary smoothly across neighboring vertices.

$\mathcal{E} = \{\mathcal{E}_1, \ldots, \mathcal{E}_r, \ldots, \mathcal{E}_m\}$ The incidence matrix B is defined as

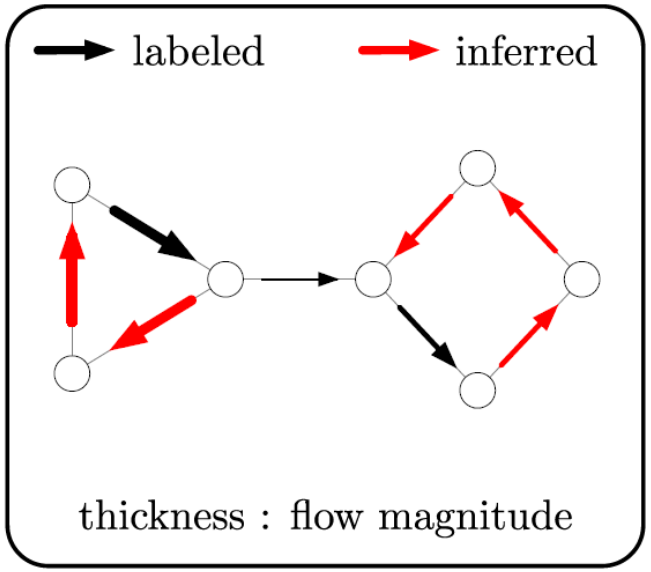
$$B_{kr} = \begin{cases} 1, & \mathcal{E}_r \equiv (i,j),\ k = i,\ i < j \\ -1, & \mathcal{E}_r \equiv (i,j),\ k = j,\ i < j \\ 0, & \text{otherwise.} \end{cases}$$

This notion of smoothness can be defined via a loss function of the form $\|\mathbf{B}^\mathsf{T}\mathbf{y}\|^2 = \sum_{(i,j)\in\mathcal{E}}(y_i - y_j)^2$

$$\|\mathbf{B}^\mathsf{T}\mathbf{y}\|^2 = \mathbf{y}^\mathsf{T}\mathbf{L}\mathbf{y} \quad \Longrightarrow \quad \mathbf{L} = \mathbf{B}\mathbf{B}^\mathsf{T}$$

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

Graph-Based SSL for Edge Flows




problem

The edge flows over a network can be represented with a vector $\mathrm{f}$, where $\mathrm{f}_r > 0$ if the flow orientation on edge $r$ aligns with its reference orientation and $\mathrm{f}_r < 0$ otherwise. In this sense, we are only accounting for the **net flow** along an edge.

assumption

To impose a **flow conservation assumption** for edge flows, we consider the divergence at each vertex, which is the sum of outgoing flows minus the sum of incoming flows at a vertex.

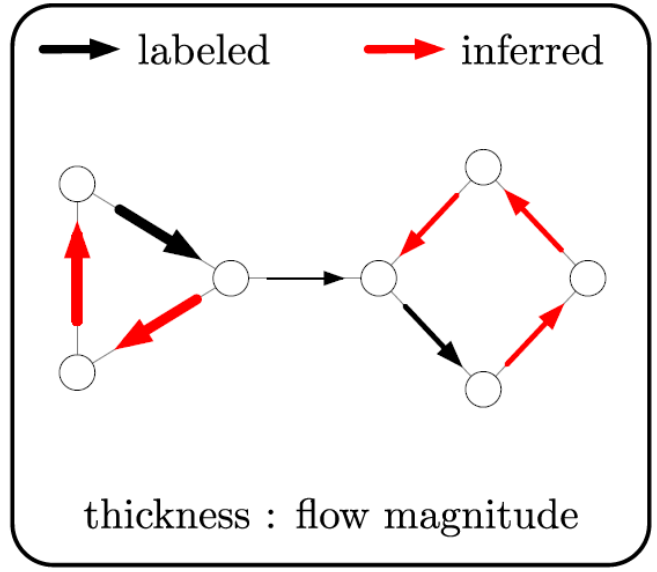
$$(\mathbf{B}\mathrm{f})_i = \sum_{\mathcal{E}_r \in \mathcal{E} \,:\, \mathcal{E}_r \equiv (i,j),\, i<j} \mathrm{f}_r - \sum_{\mathcal{E}_r \in \mathcal{E} \,:\, \mathcal{E}_r \equiv (j,i),\, j<i} \mathrm{f}_r .$$

loss

To create a loss function for edge flows that enforces a notion of flow-conservation, we use the sum-of-squares vertex divergence:

$$\|\mathbf{B}\mathrm{f}\|^2 = \mathrm{f}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}\mathbf{B}\mathrm{f} = \mathrm{f}^{\mathsf{T}}\mathbf{L}_e\mathrm{f}.$$

However, unlike the case for smooth vertex labels, requiring $\mathbf{f}^{\mathsf{T}}\mathbf{L}_e\mathbf{f} = 0$ is actually under-constrained, i.e., even when more than one edge is labeled, many different divergence-free edge-flow assignments may exist that induce zero loss.




$$\mathbf{f}^* = \arg\min_{\mathbf{f}} \|\mathbf{B}\mathbf{f}\|^2 + \lambda^2 \cdot \|\mathbf{f}\|^2$$

$$\text{s.t.} \quad \mathrm{f}_r = \hat{\mathrm{f}}_r, \forall \mathcal{E}_r \in \mathcal{E}^{\mathrm{L}}.$$

## Computation

Let $\mathbf{f}^0$ be a trival feasible point where $\mathbf{f}_r^0 = \hat{\mathbf{f}}_r$ if $r \in \varepsilon^L$ and $\mathbf{f}^r = 0$ otherwise.

Denote the set of indices for unlabeled edges as $\mathcal{E}^{\mathrm{U}} = \{\mathcal{E}_1^{\mathrm{U}}, \mathcal{E}_2^{\mathrm{U}}, \ldots, \mathcal{E}_{m^{\mathrm{U}}}^{\mathrm{U}}\}$

We define the expansion operator $\boldsymbol{\Phi}$ as a linear map from $\mathbb{R}^{m^U}$ to $\mathbb{R}^{\boldsymbol{m}}$ given by $\Phi_{rs} = 1$ if $\varepsilon_r = \varepsilon_s^U$ and 0 otherwise. Let $\mathbf{f}^U \in \mathbb{R}^{m^U}$ be the edge flows on the unlabeled edges.

The original problem can be converted to a linear least-squares problem:

$$\mathbf{f}^* = \arg\min_{\mathbf{f}} \|\mathbf{B}\mathbf{f}\|^2 + \lambda^2 \cdot \|\mathbf{f}\|^2$$

$$\text{s.t.} \quad \mathrm{f}_r = \hat{\mathrm{f}}_r, \, \forall \mathcal{E}_r \in \mathcal{E}^{\mathrm{L}}.$$

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

Any feasible point can be written as $\mathbf{f}^0 + \boldsymbol{\Phi}\mathbf{f}^{\mathrm{U}}$

The least-squares problem can be solved with iterative methods such as LSQR or LSMR.

*LSMR: An Iterative Algorithm for Sparse Least-Squares Problems.(SIAM2017)*

*LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. ACM TOMS (1982).*

## Spectral Graph Theory Interpretations

The spectral decomposition of the graph Laplacian matrix is $\mathbf{L} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}^{\mathsf{T}}$. Because $\mathbf{L} = \mathbf{B}\mathbf{B}^{\mathsf{T}}$ the orthonormal basis $\mathbf{U} \in \mathbb{R}^{n \times n}$ for vertex labels is formed by the left singular vectors of the incidence matrix

$\mathbf{\Sigma} \in \mathbb{R}^{n \times m}$ is the diagonal matrix of ordered singular values with m − n columns of zero-padding on the right

the right singular vectors $\mathbf{V} \in \mathbb{R}^{m \times m}$ is an orthonormal basis for edge flows.

The divergence-minimizing objective can be rewritten in terms of the right singular vectors of $\mathbf{B}$

Let $\mathbf{p} = \mathbf{V}^{\mathsf{T}}\mathbf{f} \in \mathbb{R}^{m}$ represent the spectral coefficients of $\mathbf{f}$ expressed in terms of the basis $\mathbf{V}$.

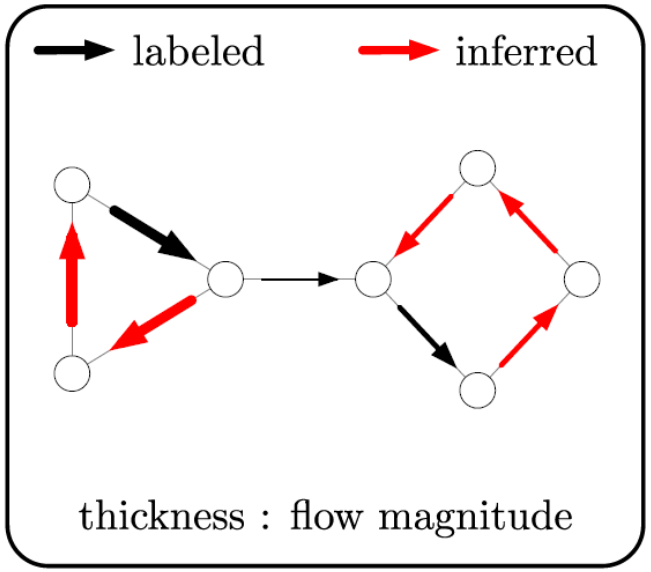
$$\mathbf{f}^{*} = \arg\min_{\mathbf{f}} \|\mathbf{B}\mathbf{f}\|^{2} + \lambda^{2} \cdot \|\mathbf{f}\|^{2} \qquad \text{s.t.} \quad \mathrm{f}_{r} = \hat{\mathrm{f}}_{r},\ \forall \mathcal{E}_{r} \in \mathcal{E}^{\mathrm{L}}.$$

$$\mathbf{f}^* = \mathbf{V} \cdot \arg\min_{\mathbf{p}} (\mathbf{V}\mathbf{p})^\mathsf{T} \mathbf{B}^\mathsf{T} \mathbf{B} (\mathbf{V}\mathbf{p}) + \lambda^2 \cdot (\mathbf{V}\mathbf{p})^\mathsf{T} (\mathbf{V}\mathbf{p})$$

$$= \mathbf{V} \cdot \arg\min_{\mathbf{p}} \mathbf{p}^\mathsf{T} \left( \Sigma^\mathsf{T} \Sigma + \lambda^2 \cdot \mathbf{I} \right) \mathbf{p}$$

$$= \mathbf{V} \cdot \arg\min_{\mathbf{p}} \lambda^2 \cdot \sum_{\alpha} \frac{\sigma_\alpha^2 + \lambda^2}{\lambda^2} p_\alpha^2$$

$$\text{s.t.} \quad (\mathbf{V}\mathbf{p})_r = \hat{\mathrm{f}}_r, \ \forall \mathcal{E}_r \in \mathcal{E}^\mathrm{L},$$




By construction **V** is a complete orthonormal basis for the space of edge flows. This space can be decomposed into two orthogonal subspaces.

The first subspace is the **cut-space** $\mathcal{R} = \mathrm{im}(\mathbf{B}^\mathsf{T})$ spanned by the singular vectors associated with nonzero singular values.

The space $\mathcal{R}$ is also called the space of gradient flows, since any vector may be written as $\mathbf{B}^\mathsf{T}\mathbf{y}$,

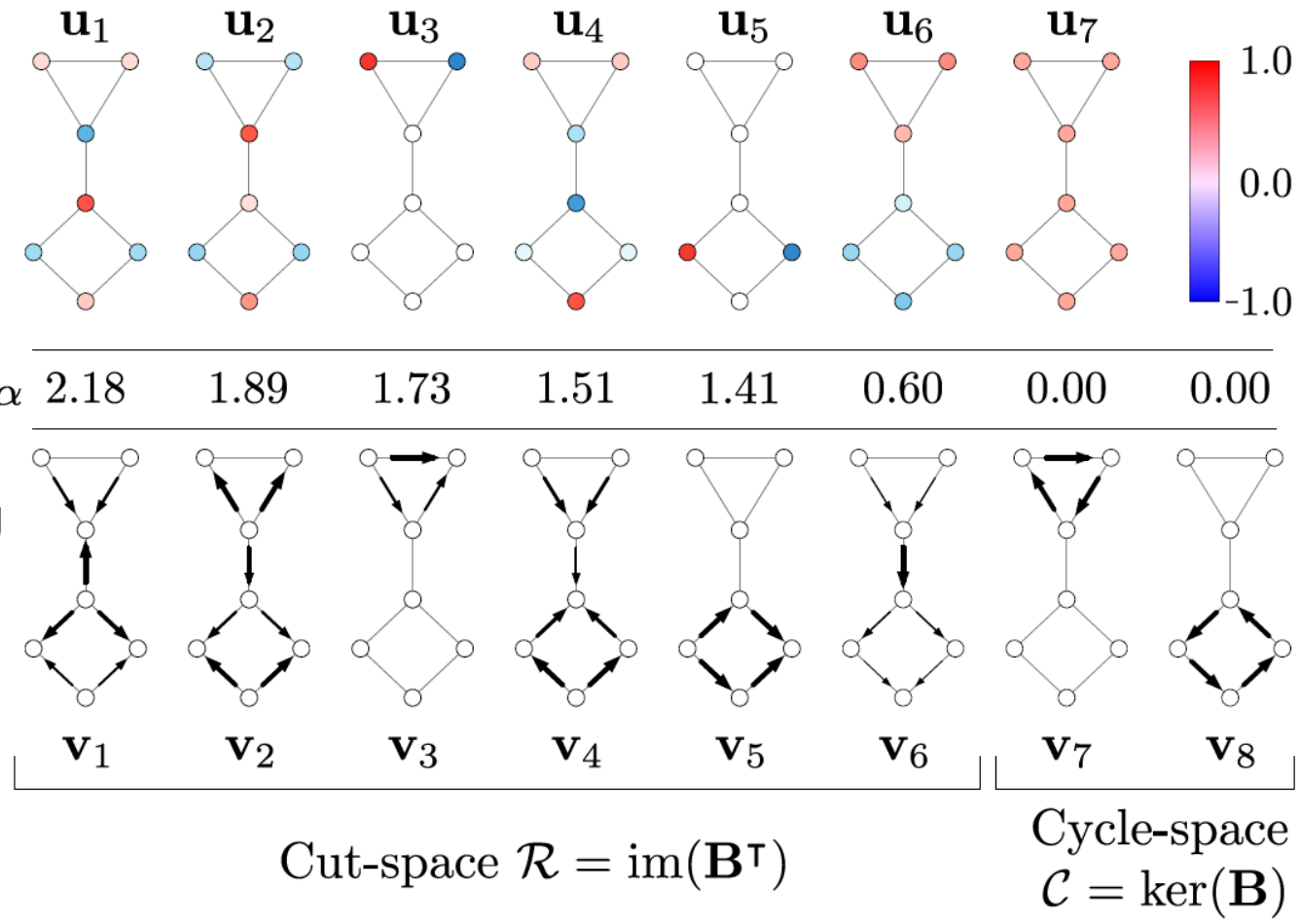
$\mathbf{y}$ is a vector of vertex scalar potentials that induce a gradient flow.

The second subspace is the cycle-space $C = \ker(\mathbf{B})$ spanned by the remaining right singular vectors $\mathbf{V}_C$ associated with zero singular values. Note that any vector $\mathbf{f} \in C$ corresponds to a circulation of flow, and will induce zero cost in the loss function.

Let $\mathbf{u}_\alpha, \sigma_\alpha, \mathbf{v}_\alpha$ a triple of a left singular vector, singular value, and right singular vector.

$\mathbf{u}_\alpha^\mathsf{T}\, \mathbf{L}\, \mathbf{u}_\alpha = \sigma_\alpha^2$ provide a notion of "unsmoothness" of basis vector $\mathbf{u}_\alpha$ representing vertex labels,

$\mathbf{v}_\alpha^\mathsf{T}\, \mathbf{L}_e\, \mathbf{v}_\alpha = \sigma_\alpha^2$ gives the sum-of-squares divergence of basis vector $\mathbf{v}_\alpha$ representing edge flows.



LEMMA 2.1. *Assume the ground truth flows are divergence-free. Then as* $\lambda \longrightarrow 0$, *the solution of Eq.* (8) *can exactly recover the ground truth from some labeled edge set* $\mathcal{E}^{\mathrm{L}}$ *with cardinality* $c = m - n + 1$.

$$\begin{aligned}
\mathbf{f}^* &= \mathbf{V} \cdot \arg\min_{\mathbf{p}} (\mathbf{V}\mathbf{p})^{\intercal}\mathbf{B}^{\intercal}\mathbf{B}\,(\mathbf{V}\mathbf{p}) + \lambda^2 \cdot (\mathbf{V}\mathbf{p})^{\intercal}(\mathbf{V}\mathbf{p}) \\
&= \mathbf{V} \cdot \arg\min_{\mathbf{p}} \mathbf{p}^{\intercal}\left(\Sigma^{\intercal}\Sigma + \lambda^2 \cdot \mathbf{I}\right)\mathbf{p} \qquad \text{s.t.} \quad (\mathbf{V}\mathbf{p})_r = \hat{\mathrm{f}}_r,\ \forall \mathcal{E}_r \in \mathcal{E}^{\mathrm{L}} \\
&= \mathbf{V} \cdot \arg\min_{\mathbf{p}} \lambda^2 \cdot \sum_{\alpha} \frac{\sigma_\alpha^2 + \lambda^2}{\lambda^2} p_\alpha^2
\end{aligned}$$

Recall that $\mathbf{p}_C \in \mathbb{R}^c$ are the spectral coefficients of a basis $\mathbf{V}_C$ in the cycle-space, then the ground truth edge flows can be written as $\hat{\mathbf{f}} = \mathbf{V}_C \mathbf{p}_C$

On the other hand, in the limit $\lambda \longrightarrow 0$ the spectral coefficients of basis vectors with non-zero singular values have infinite weights and are forced to zero.

Therefore, by choosing the set of labeled edges corresponding to $c = m - n + 1$ linearly independent rows from $\mathbf{V}_C$, the ground truth $\hat{\mathbf{f}}$ is the unique optimal solution.

THEOREM 2.2. *Let $\mathbf{V}_C^{\mathrm{L}}$ denote c linearly independent rows of the $\mathbf{V}_C$ that correspond to labeled edges. If the divergence-free edge flows $\mathbf{f}$ are perturbed by $\delta$, then as $\lambda \to 0$, the reconstruction error of the proposed algorithm is bounded by $[\sigma_{\min}^{-1}(\mathbf{V}_C^{\mathrm{L}}) + 1] \cdot \|\delta\|$.*

The ground truth edge flows can be written as $\hat{\mathbf{f}} = \mathbf{f} + \delta = \begin{bmatrix} \mathbf{f}^{\mathrm{L}} \\ \mathbf{f}^{\mathrm{U}} \end{bmatrix} + \begin{bmatrix} \delta^{\mathrm{L}} \\ \delta^{\mathrm{U}} \end{bmatrix}$

Further, the reconstructed edge flows are given by $\mathbf{V}_C(\mathbf{V}_C^{\mathrm{L}})^{-1}(\mathbf{f}^{\mathrm{L}} + \delta^{\mathrm{L}})$

Therefore, we can bound the norm of the reconstruction error as follows:

$$\|\mathbf{V}_C(\mathbf{V}_C^{\mathrm{L}})^{-1}(\mathbf{f}^{\mathrm{L}} + \delta^{\mathrm{L}}) - (\mathbf{f} + \delta)\|$$

# SEMI-SUPERVISED LEARNING RESULTS

Minnesota road network n = 2642,m =3303

US power grid network from **KONECT** n = 4941,m = 6593

The water irrigation network of Balerma city

n = 447,m = 454

An autonomous system network n = 520,m = 1280

Create synthetic flows with spectral coefficients for each basis vector that are inversely proportional to the associated singular values

$$p_\alpha = \frac{b}{\sigma_\alpha + \epsilon} \qquad b = 0.02, \epsilon = 0.1$$

Performance Measurement and Baselines

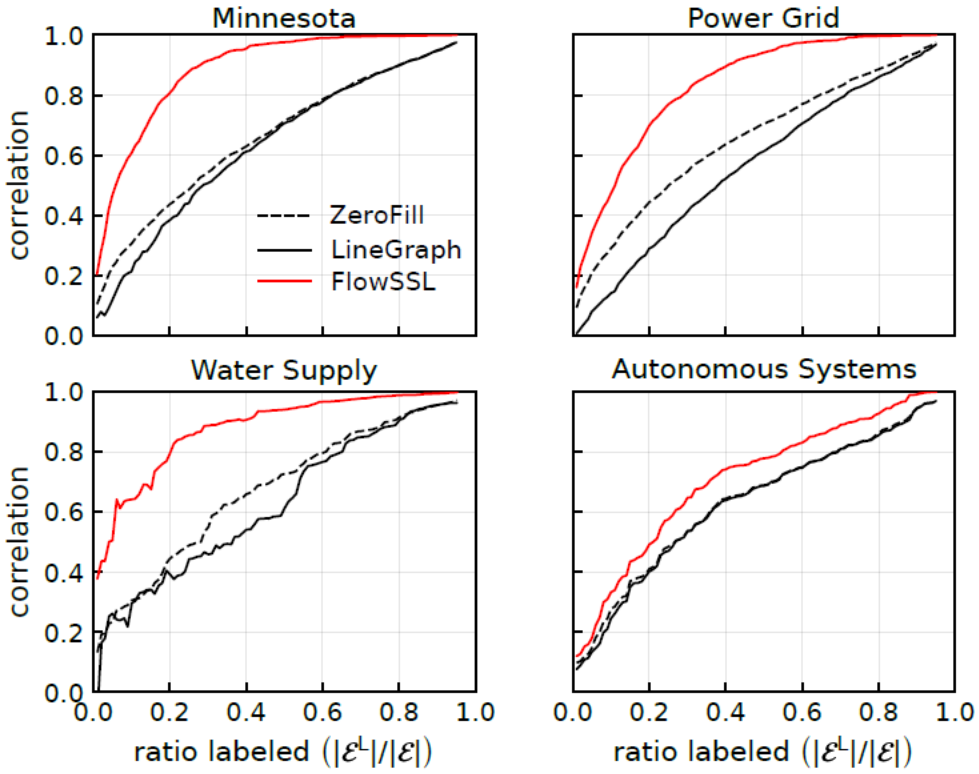
Pearson correlation coefficient $$\rho(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}}$$

First, the **ZeroFill** baseline simply assigns 0 edge flows to all unlabeled edges.

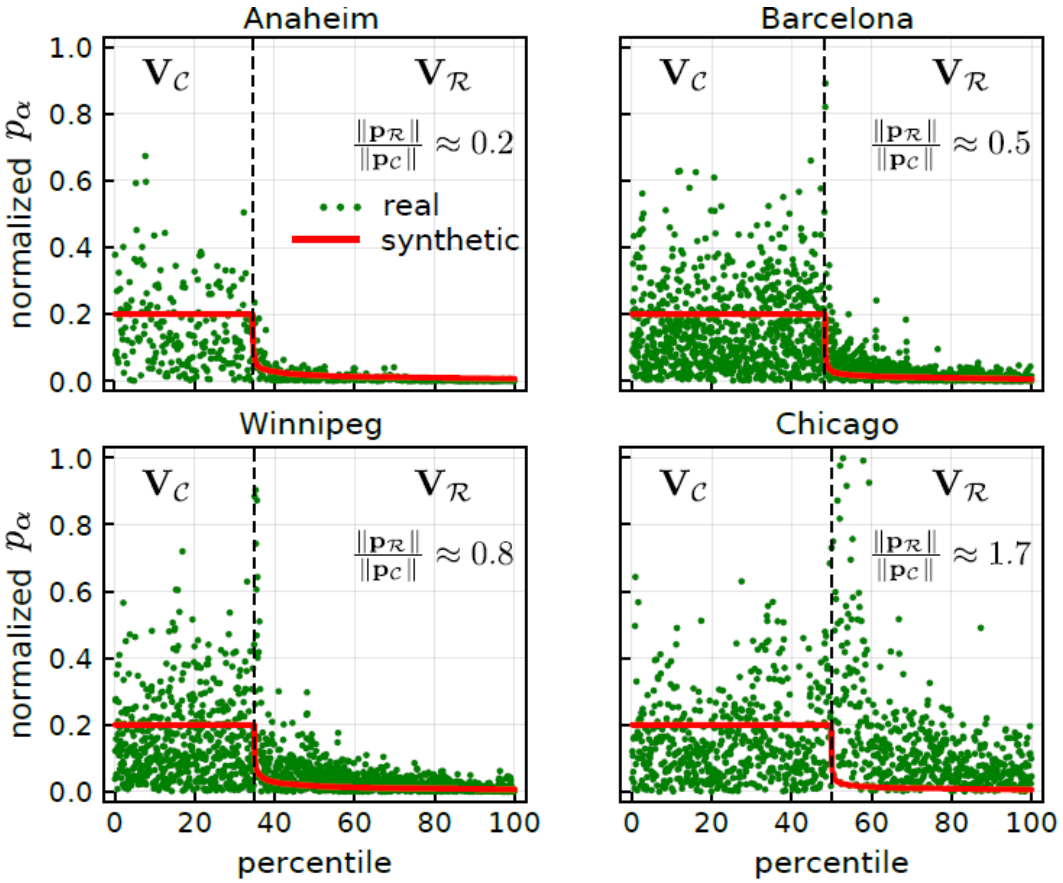
Second, the **LineGraph** baseline uses a line-graph transformation of the network and then applies standard vertex-based SSL on the resulting graph.

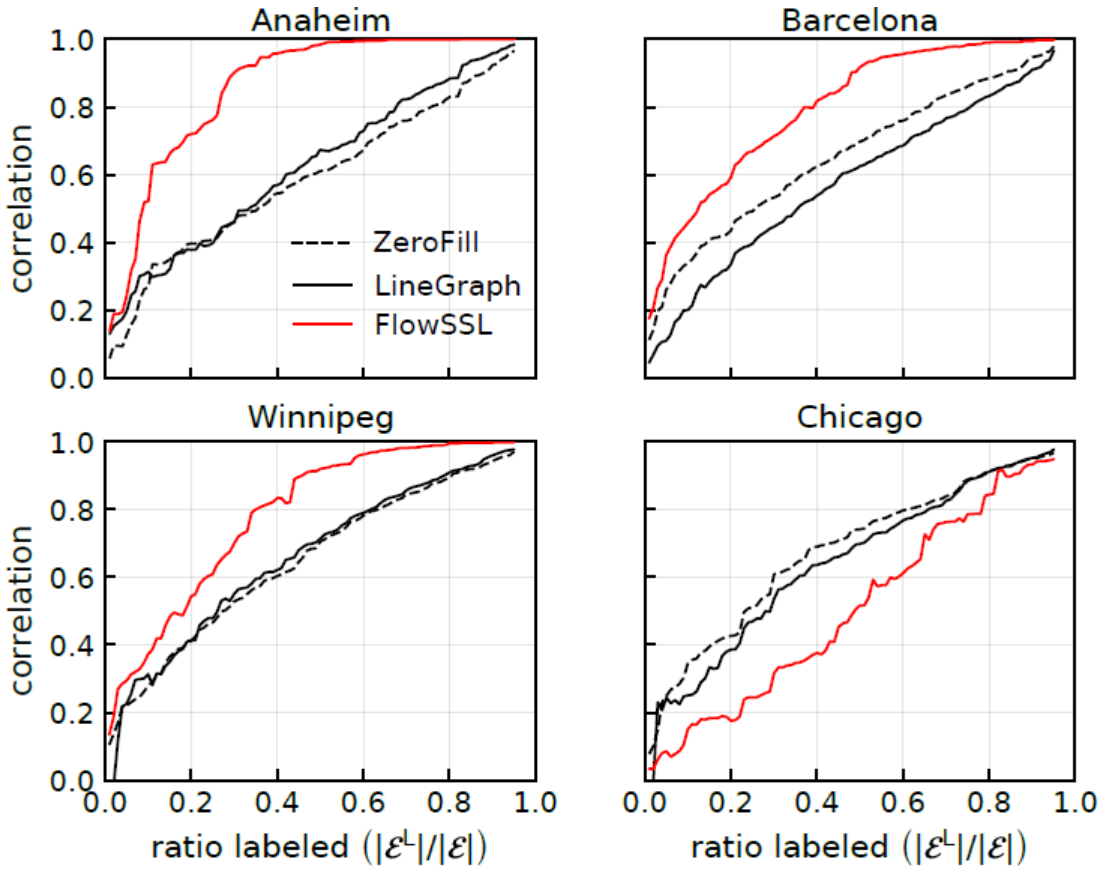
The LineGraph approach performs no better than ZeroFill. This should not be surprising, since the LineGraph approach does not interpret the sign of an edge flow as an orientation but simply as part of a numerical label

In our first set of experiments, the network topology comes from real data, but we use synthetic flows to demonstrate our method.

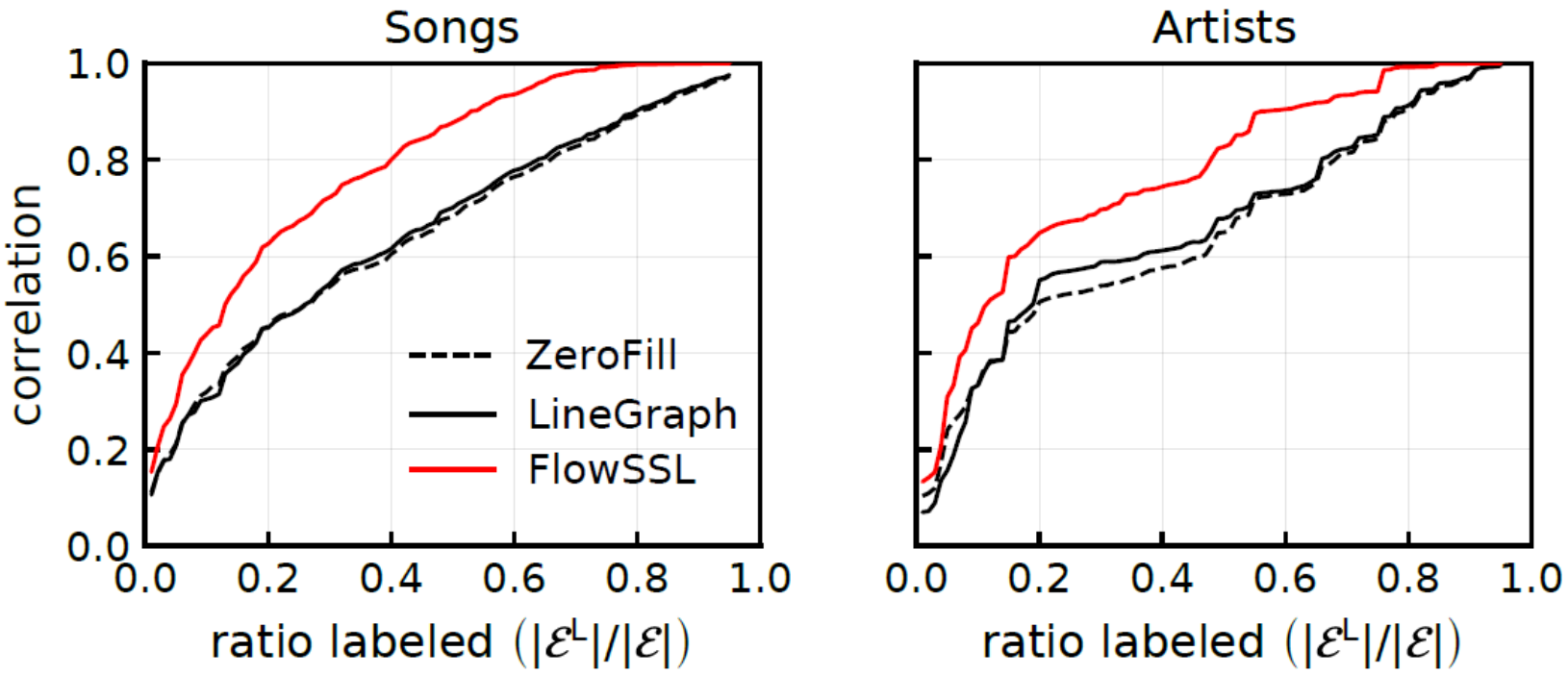


# Learning Real-World Traffic Flows



## Information Flow Networks




Even though the flows are not physical, FlowSSL method still outperforms the baselines.

**Two active learning algorithms**

$$= [\sigma_{\min}^{-1}(\mathbf{V}_C^{\mathrm{L}}) + 1] \cdot \|\delta\|$$

one strategy for selecting $\mathcal{E}^{\mathrm{L}}$ is to choose $m^{\mathrm{L}}$ rows from $\mathcal{V}_0$ that maximize the smallest singular value of the resulting submatrix. This problem is known as optimal column subset selection (maximum submatrix volume) and is NP-hard.

However, a good heuristic is the rank revealing QR decomposition (RRQR) [5], which computes

$$\mathbf{V}_C^{\mathsf{T}}\,\Pi = Q\,[R_1 \quad R_2]$$

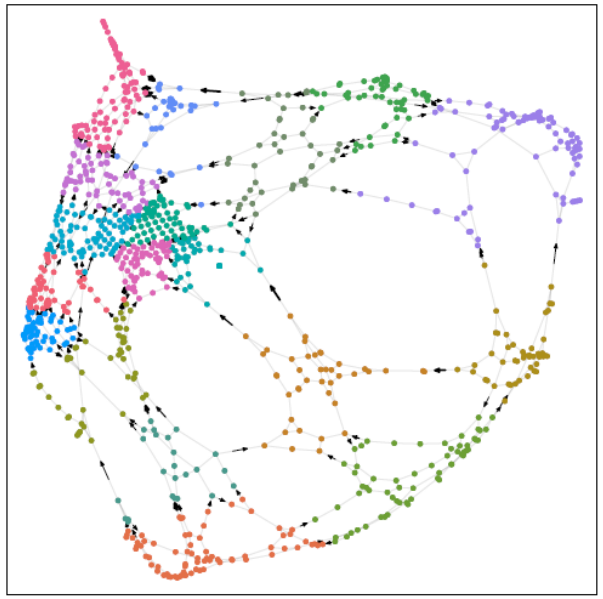
$\Pi$ is a permutation matrix that keeps $R_1$ well-conditioned.

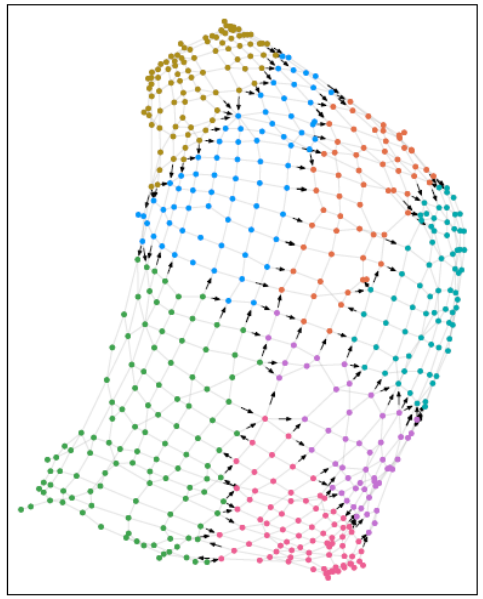
This approach is mathematically similar to graph clustering algorithms that use RRQR to select representative vertices for cluster centers.

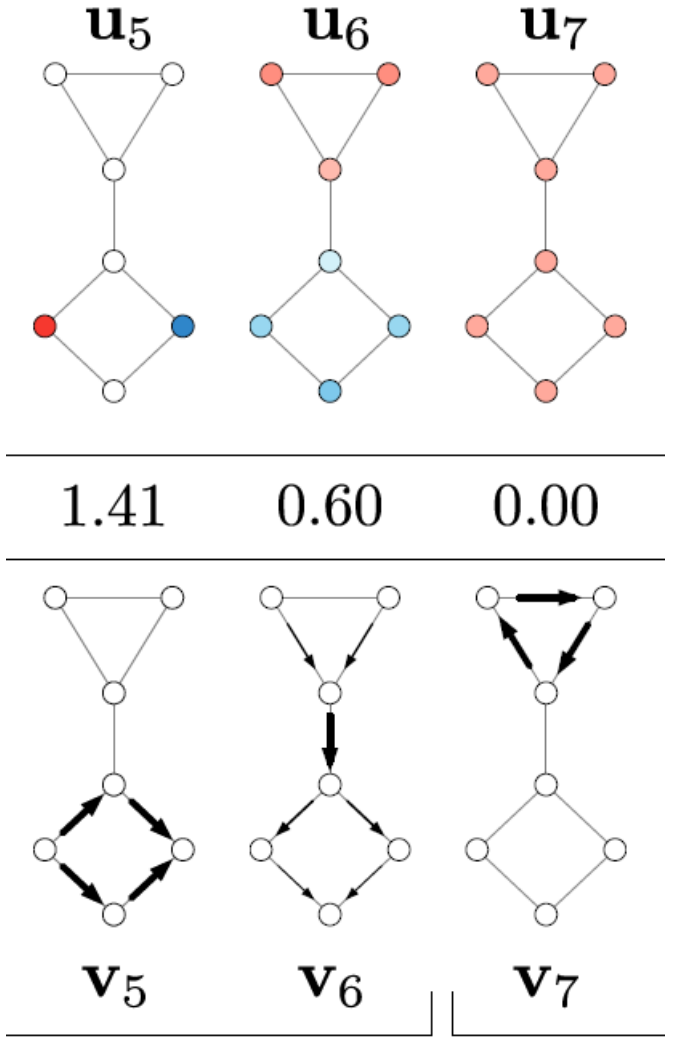
## Recursive Bisection (RB)

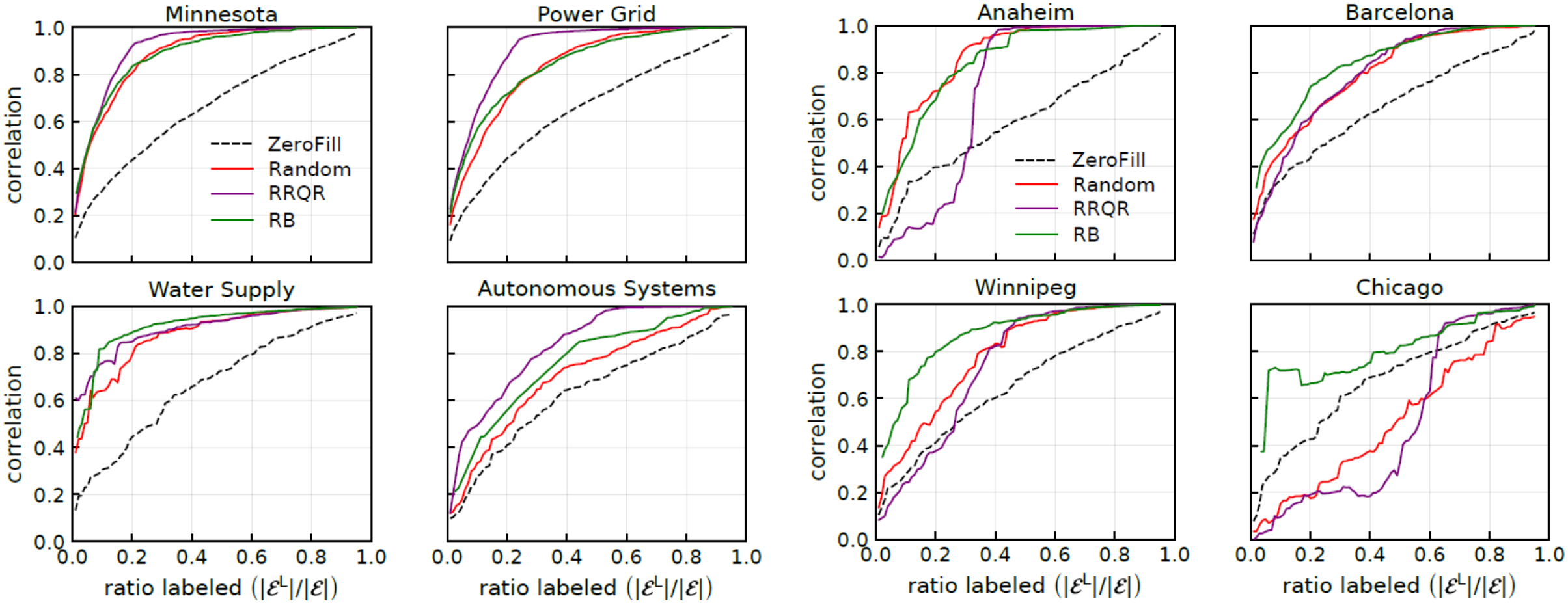
The intuition behind this heuristic is that edge flows on bottleneck-edges, which partition a network, are able to capture global trends in the networks' flow pattern.

We start with an empty labeled set $\mathcal{E}^{\mathrm{L}}$, a target number of labeled edges $m^{\mathrm{L}}$, Next, we recursively partition the largest cluster in the graph with spectral clustering and add every edge that connects the two resulting clusters into $\mathcal{E}^{\mathrm{L}}$








Minnesota
Power Grid
Anaheim
Barcelona
Water Supply
Autonomous Systems
Winnipeg
Chicago
correlation
ratio labeled ($|\mathcal{E}^L|/|\mathcal{E}|$)
ZeroFill
Random
RRQR
RB
0.0
0.2
0.4
0.6
0.8
1.0

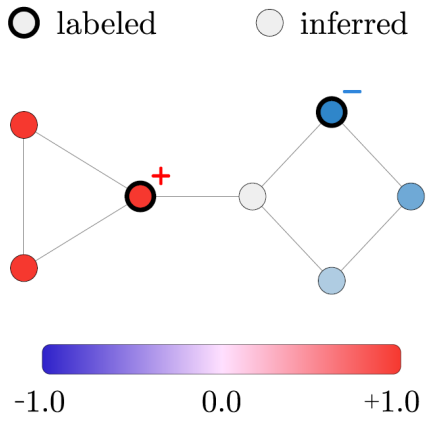
“Similar methods have been shown to be effective in semi-supervised active learning for vertex labels, in these cases, the graph is first clustered, and then one vertex is selected from each cluster.”

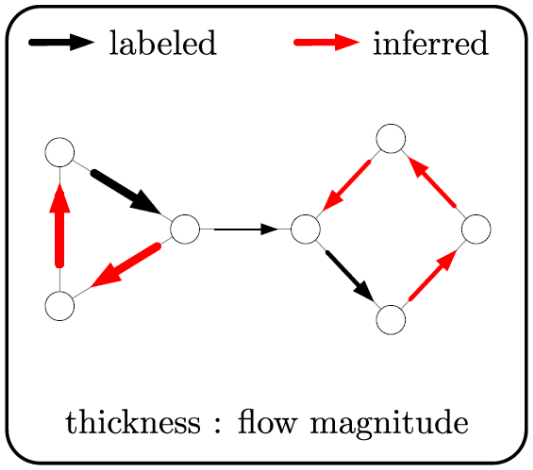
Label Selection on Graphs(NIPS2009)

While any other graph partitioning algorithm could be used and greedy recursive bisection approaches can be sub-optimal.

How Good is Recursive Bisection?(SIAM1998)

$$\mathbf{y}^* = \arg\min_{\mathbf{y}} \|\mathbf{B}^\intercal \mathbf{y}\|^2$$




$$\mathbf{f}^* = \arg\min_{\mathbf{f}} \|\mathbf{B}\mathbf{f}\|^2 + \lambda^2 \cdot \|\mathbf{f}\|^2$$

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

$$\mathbf{V} \cdot \arg\min_{\mathbf{p}} \lambda^2 \cdot \sum_{\alpha} \frac{\sigma_\alpha^2 + \lambda^2}{\lambda^2} p_\alpha^2$$

cut-space, and cycle space

$$\|\mathbf{V}_C(\mathbf{V}_C^{\mathrm{L}})^{-1}(\mathbf{f}^{\mathrm{L}} + \delta^{\mathrm{L}}) - (\mathbf{f} + \delta)\|$$

$$p_\alpha = \frac{b}{\sigma_\alpha + \epsilon}$$

Pearson correlation coefficient

ZeroFill

LineGraph

RRQR

Recursive Bisection