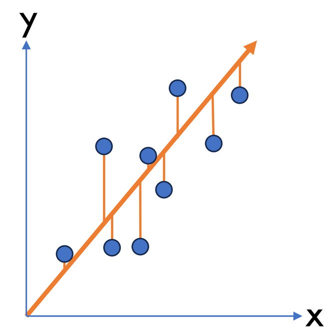
Learning in Metric Spaces: How Can we Best Predict Graphs?

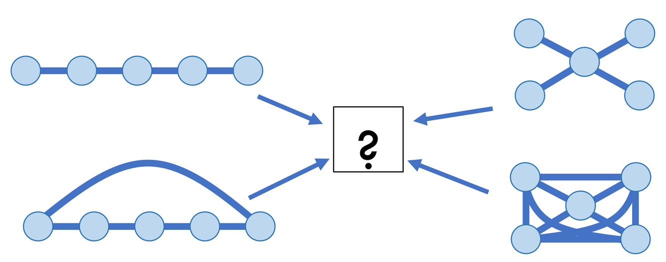
Alex G. Zalles[1], Kai M. Hung[2] Ann E. Finneran[3], Lydia Beaudrot[3], César A. Uribe[4]

[1]Department of Computational Mathematics and Operations Research, [2]Department of Computer Science, [3]Department of Biosciences, [3]Department of Electrical and Computer Engineering

How to Learn in Wasserstein Space?

- Linear Regression, modeling **Euclidean outputs** with real valued predictors
  - Distance = length
- How do we predict outputs that are **not** Euclidean (i.e., graphs)?



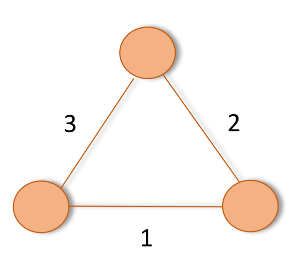


Main theoretical result → **regression** in **metric spaces** is **equivalent** to a **linear combination** in the **Wasserstein** space

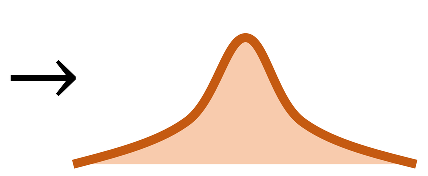
Regression with Wasserstein and Frobenius Metrics

What different distance measures are there?

- **Encode** as **Graph Laplacian,** and then covariance matrix of **Gaussian distribution**



$$\rightarrow \begin{bmatrix} 5 & -2 & -3 \\ -2 & 3 & -1 \\ -3 & -1 & 4 \end{bmatrix} \rightarrow$$



$$d_F(L_1, L_2) = \{\mathrm{tr}[(L_1 - L_2)^T(L_1 - L_2)]\}^{1/2}$$

$$W_2^2(N_1(\mu_1,\Sigma_1), N_2(\mu_2,\Sigma_2)) = \|\mu_1 - \mu_2\|_2^2 + tr\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$$

Global Regression

$$m_G(x) \coloneqq \underset{\omega\in\Omega}{argmin}\, M_G(\omega, x),$$
$$M_G(\bullet, x) = \mathbb{E}[s_G(X,x)d^2(Y,\bullet)]$$

Local Regression

$$m_{L,n}(x) \coloneqq \underset{\omega\in\Omega}{argmin}\, M_{L,n}(\omega, x),$$
$$M_{L,n}(\bullet, x) = \mathbb{E}[s_L(X,x)d^2(Y,\bullet)]$$

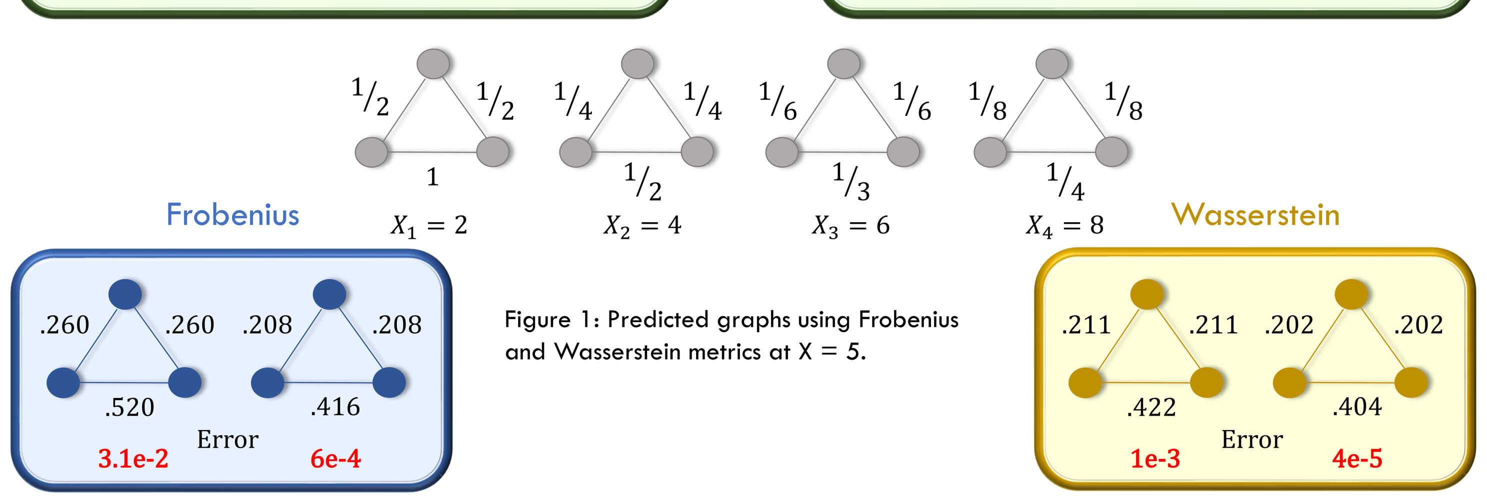


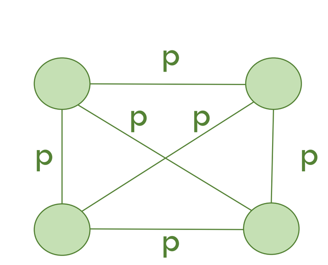
Figure 1: Predicted graphs using Frobenius and Wasserstein metrics at X = 5.

Regression on Non-Deterministic Graphs

Testing our model on less deterministic examples

- **Erdős-Rényi** graphs randomly generated

ER(n = 4, p) →

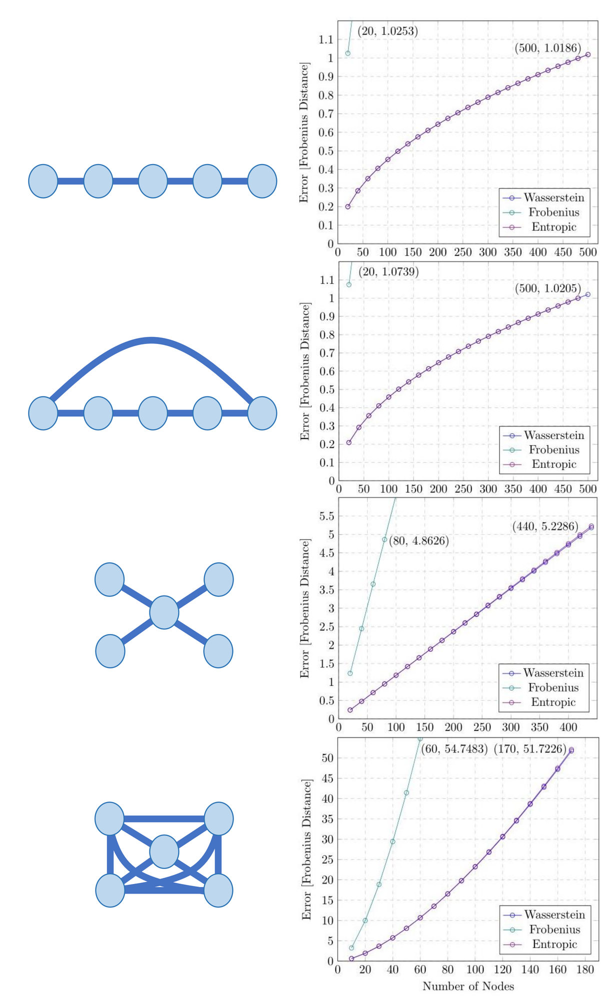


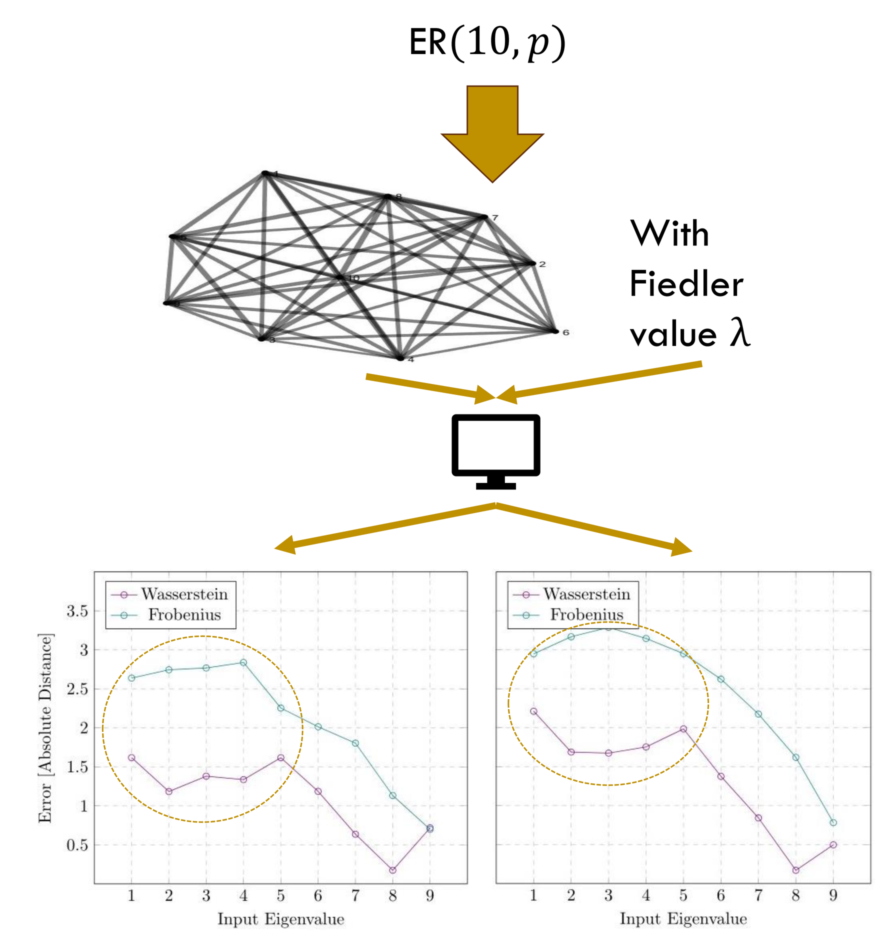
The Wasserstein barycenter $(\overline{m}, \overline{\Sigma})$ of gaussian distributions $(m_1, \Sigma_1), \ldots, (m_n, \Sigma_n)$ with weights $\lambda_1, \ldots, \lambda_n$ satisfies the equations

$$\overline{m} = \sum_{i=1}^{n} \lambda_i m_i \quad \overline{\Sigma} = \sum_{i=1}^{n} \lambda_i \left(\overline{\Sigma}^{1/2}\Sigma_i\overline{\Sigma}^{1/2}\right)^{1/2}$$

Solved using **Fixed-point Recursion**

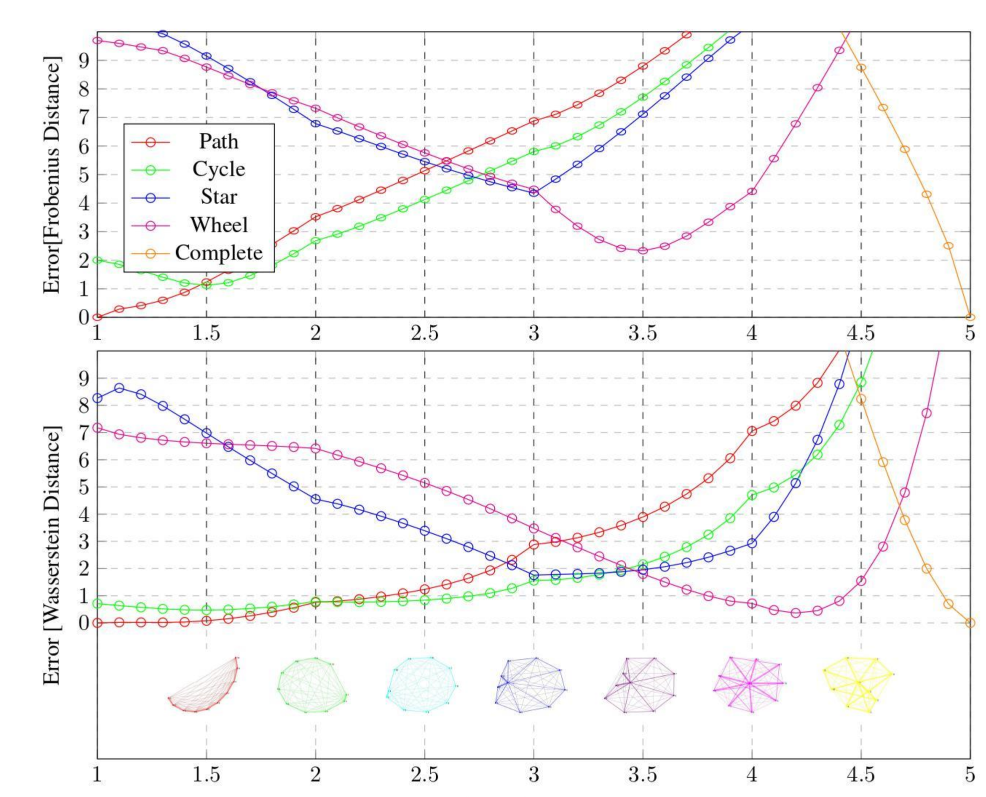
Training Over Graph Connectivity





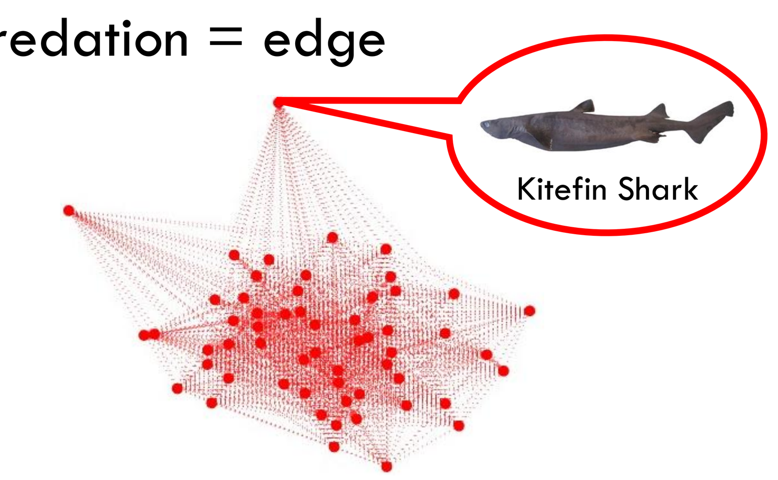
Wasserstein performs better, best on low connectivity

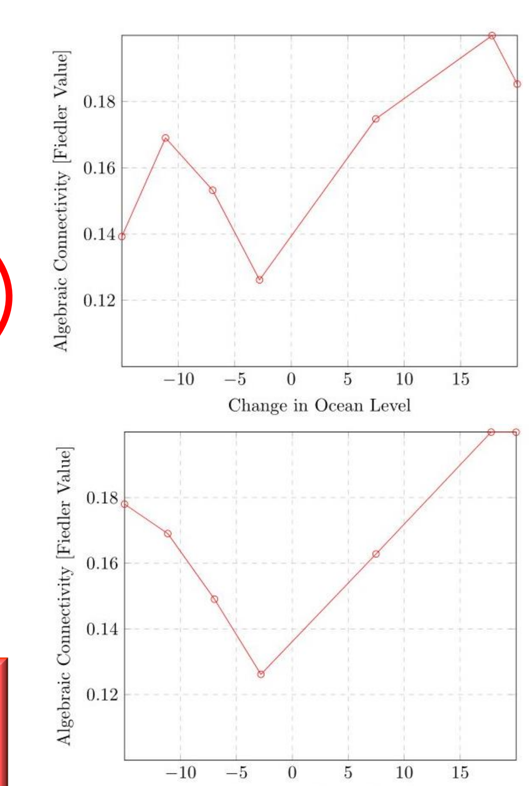
Training Over Graph Topology



Results: Food Webs

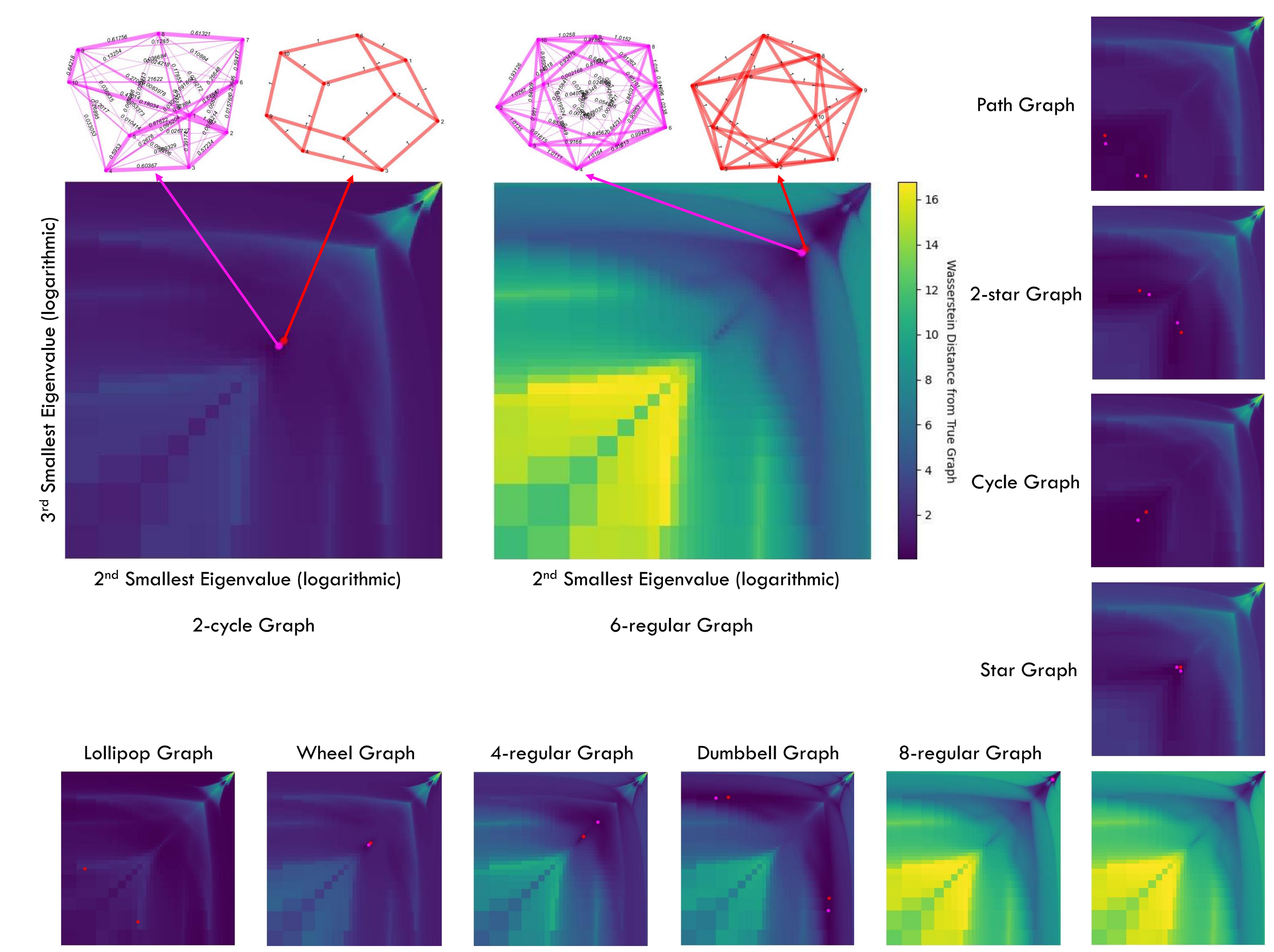
- Translate food webs into graphs
  - Organisms = nodes
  - Predation = edge





Future work will focus on utilizing regression models to understand connectivity and robustness of food webs for conservation

Training Over Spectral Properties



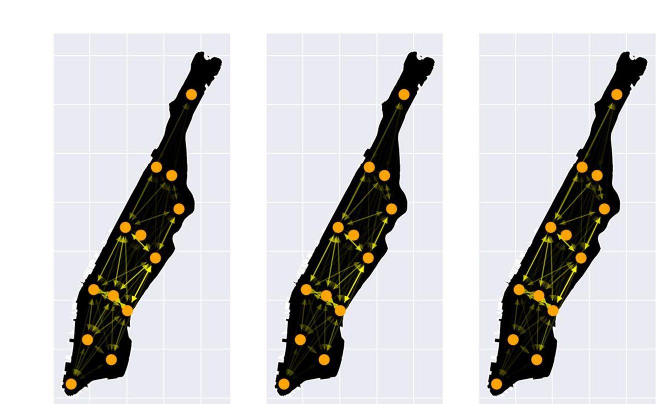
Training on more **graphs** and more **variables** results in **greater accuracy**

Results: COVID-19 and Taxi Trips

- Predicting travel as a response to COVID-19 cases

R^2 coefficient

Frobenius: 0.433
Power Metric: 0.453
Wasserstein: 0.607



Distance used	% MSPE of Frobenius
Power Metric	96.4%
Wasserstein (Prediction) Frobenius (Error)	95.995%
Wasserstein (Prediction) Wasserstein (Error)	86.375%

- Compute Mean Square Prediction error (MSPE) with ten-fold cross-validation, averaging over 100 iterations

Future Works

- **Directed** food webs, Laplacians become asymmetric
- Extending to graphs of different number of nodes with **Gromov-Wasserstein** distance
- Robust approaches to barycenters with **negative weights**
- All encapsulated in Sub-Saharan African food web data set

For more information, you can reach me at agz2@rice.edu and view my LinkedIn profile via the QR code. This work is generously supported by NSF with grant number 2213568.