Analyzing the Harmonic Structure in Graph-Based Learning

Xiao-Ming Wu[1] Zhenguo Li[2] Shih-Fu Chang[1]

[1]Columbia University [2]Huawei Noah's Ark Lab, Hong Kong

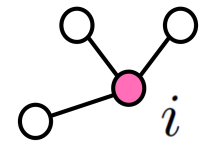
Motivation

- How to measure the fit between a model and a graph?

  Model ⟺ Graph

- Many target functions exhibit a harmonic structure:

  value on a vertex ~ weighted average of its neighbors



$$f : \mathcal{V} \to \mathbb{R} \qquad f(i) \approx \sum_{j \sim i} \frac{w_{ij}}{d_i} f(j)$$

$$\mathcal{G} = (\mathcal{V}, W) \qquad W = [w_{ij}] \in \mathbb{R}^{n \times n} \qquad d_i = \sum_j w_{ij}$$

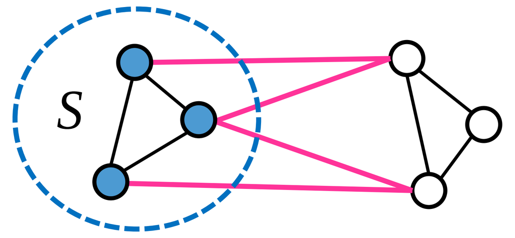
- Objectives of this paper
  - Analyze the harmonic structure;
  - Answer open questions in various graph models;
  - Provide guidelines for various applications.

Harmonic Structure Analysis

Harmonic Loss

$$\mathcal{L}_f(\mathcal{S}) := \sum_{i \in \mathcal{S}} d_i \left(f(i) - \sum_{j \sim i} \frac{w_{ij}}{d_i} f(j) \right)$$

Lemma 2.2. $\mathcal{L}_f(\mathcal{S}) = \sum_{i \in \mathcal{S}, j \in \bar{\mathcal{S}}} w_{ij}(f(i) - f(j))$.



- Quantify discrepancy across the cut.

Definition 2.3 (Superlevel set). $\{i \mid f(i) \geq c\}$

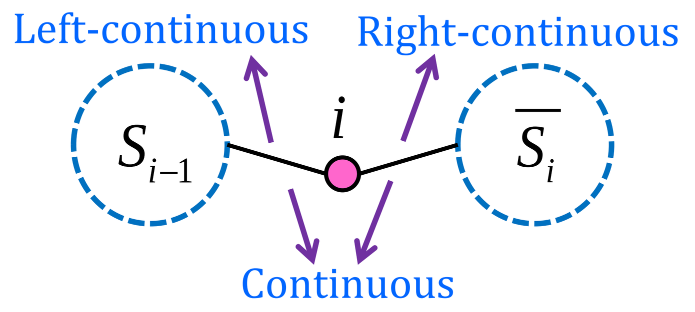
- Assume $f(1) \geq f(2) \geq \cdots \geq f(n-1) \geq f(n)$.

$\mathcal{S}_i := \{1, \ldots, i\}$ - superlevel set with level $f(i)$

Lemma 2.4. $\mathcal{L}_f(\mathcal{S}_i) \geq 0, i = 1, \ldots, n$.

Continuity

$f : \mathcal{V} \to \mathbb{R}$ left-continuous if $i \sim \mathcal{S}_{i-1}$, right-continuous if $i \sim \bar{\mathcal{S}}_i$, continuous if $i \sim \mathcal{S}_{i-1}$ and $i \sim \bar{\mathcal{S}}_i$



- Vertices of similar values are connected.

Proposition 2.6.

$\mathcal{L}_f(i) < 0 \Rightarrow$ left-continuous;

$\mathcal{L}_f(i) > 0 \Rightarrow$ right-continuous;

$\mathcal{L}_f(i) = 0 \Rightarrow$ continuous.

Bounding Function Variation

Theorem 2.7 (Dropping upper bound).

$$f(i) - f(i+1) \leq \frac{\mathcal{L}_f(\mathcal{S}_i)}{w(\mathcal{S}_i, \bar{\mathcal{S}}_i)} = \frac{\mathcal{L}_f(\mathcal{S}_i)}{\Phi(\mathcal{S}_i) \min(d(\mathcal{S}_i), d(\bar{\mathcal{S}}_i))}.$$

- f drops little in a dense area.

Theorem 2.8 (Dropping lower bound).

$$f(u) - f(v) \geq \frac{\mathcal{L}_f(\mathcal{S}_i)}{w(\mathcal{S}_i, \bar{\mathcal{S}}_i)} = \frac{\mathcal{L}_f(\mathcal{S}_i)}{\Phi(\mathcal{S}_i) \min(d(\mathcal{S}_i), d(\bar{\mathcal{S}}_i))},$$

$$u := \arg \max_{j \in \mathcal{S}_i, j \sim \bar{\mathcal{S}}_i} f(j) \quad v := \arg \min_{j \in \bar{\mathcal{S}}_i, j \sim \mathcal{S}_i} f(j)$$

- f drops a lot across a sparse cut.

If the harmonic loss varies slowly, i.e., f is *harmonic almost everywhere*,
→conductance dominates variation of f .

Absorbing Random Walks [1]

- Laplacian regularization seems problematic for classification [2];
- Labeled data - absorbing states; f - absorption probabilities.

$$1 = f(1) > f(2) \geq \cdots \geq f(n-1) > f(n) = 0$$

Harmonic form: $f(i) = \sum_{k \sim i} \frac{w_{ik}}{d_i} f(k), \quad \text{for } i = 2, \ldots, n-1.$

1. $\mathcal{L}_f(\mathcal{S}_i) = \sum_{k \sim 1} w_{1k}(1 - f(k)), i = 1, \ldots, n-1,$ *is a constant.*
2. *f is continuous if f is mutually different on unlabeled data.*

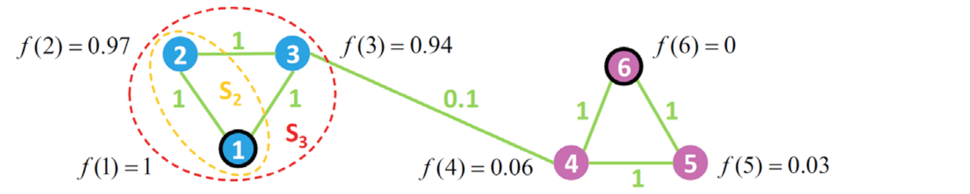
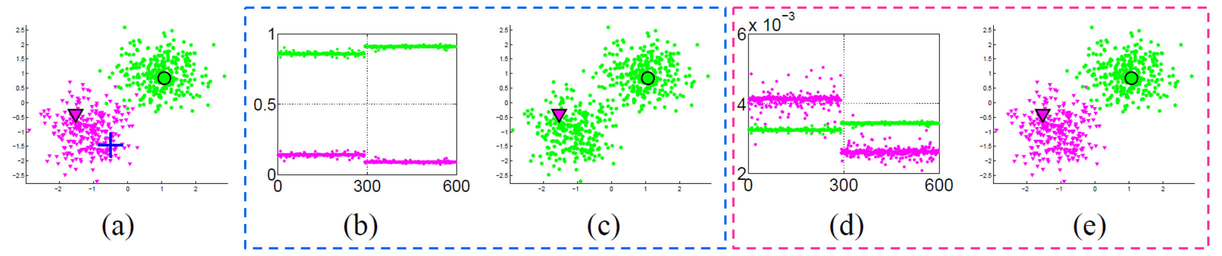


Figure 1: Absorbing random walks on a 6-point graph.

Lemma 3.3. Normalize each function by its mass before comparison.



(a) (b) (c) (d) (e)

ARW-1NN ARW-N-1NN

Partially Absorbing Random Walks [3]

- At each move, a walker gets absorbed at current state with probability $p_{ii} = \frac{\alpha\lambda_i}{\alpha\lambda_i + d_i}$ $\alpha > 0$ $\lambda_i > 0$.
- $A = (\alpha\Lambda + L)^{-1}\alpha\Lambda$ - absorption probability matrix ;
- $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ - regularization matrix ;
- p - first column of $A = [a_{ij}] \in \mathbb{R}^{n \times n}$, i.e., the probabilities starting from every vertex and getting absorbed at the first vertex. Assume $p(1) > p(2) \geq \cdots \geq p(n)$.

Harmonic form:

$$p(1) = \frac{\alpha\lambda_1}{d_1}(1 - p(1)) + \sum_{k \sim 1} \frac{w_{1k}}{d_1} p(k),$$

$$p(i) = -\frac{\alpha\lambda_i}{d_i} p(i) + \sum_{k \sim i} \frac{w_{ik}}{d_i} p(k), \; i = 2, \ldots, n.$$

1. $\mathcal{L}_p(\mathcal{S}_i) = \alpha\lambda_1 \sum_{k \in \bar{\mathcal{S}}_i} a_{1k}, i = 1, \ldots, n-1.$
2. *p is left-continuous.*

- Setting of Λ in [3] is unnecessary;
- A random Λ performs equally well;
- Columns of A are informative, not rows;
- [4] is a special case with $\Lambda = D$.

Pseudo-inverse of Laplacian

- $L^\dagger$ - a valid kernel of commute times (CT);
- CT fails to capture the graph topology [6];
- ℓ - first column of $L^\dagger$. Assume

$$\ell(1) > \ell(2) \geq \cdots \geq \ell(n)$$

Harmonic form:

$$\ell(1) = \frac{1 - \frac{1}{n}}{d_1} + \sum_{k \sim 1} \frac{w_{1k}}{d_1} \ell(k),$$

$$\ell(i) = -\frac{\frac{1}{n}}{d_i} + \sum_{k \sim i} \frac{w_{ik}}{d_i} \ell(k), \; i = 2, \ldots, n.$$

1. $\mathcal{L}_\ell(\mathcal{S}_i) = \frac{|\bar{\mathcal{S}}_i|}{n}, i = 1, \ldots, n-1.$
2. *ℓ is left-continuous.*

- $\mathcal{L}_\ell(\mathcal{S}_i) < 1$ and decreases very slowly in large graphs, since $\mathcal{L}_\ell(\mathcal{S}_i) - \mathcal{L}_\ell(\mathcal{S}_{i+1}) = 1/n$.
- This justifies its superiority in practice [5].

Hitting Times

- The expected number of steps starting from one vertex to hit others is dominated by the local structure around the targets [6];
- $h : \mathcal{V} \to \mathbb{R}$ - hitting times (HT) from every vertex to a particular target;
- Assume vertex n is the target, and

$$h(1) \geq h(2) \geq \cdots \geq h(n-1) > h(n) = 0$$

Harmonic form:

$$h(i) = 1 + \sum_{k \sim i} \frac{w_{ik}}{d_i} h(k), \quad \text{for } i = 1, \ldots, n-1.$$

1. $\mathcal{L}_h(\mathcal{S}_i) = \sum_{1 \leq k \leq i} d_k, i = 1, \ldots, n-1.$
2. *h is right-continuous.*

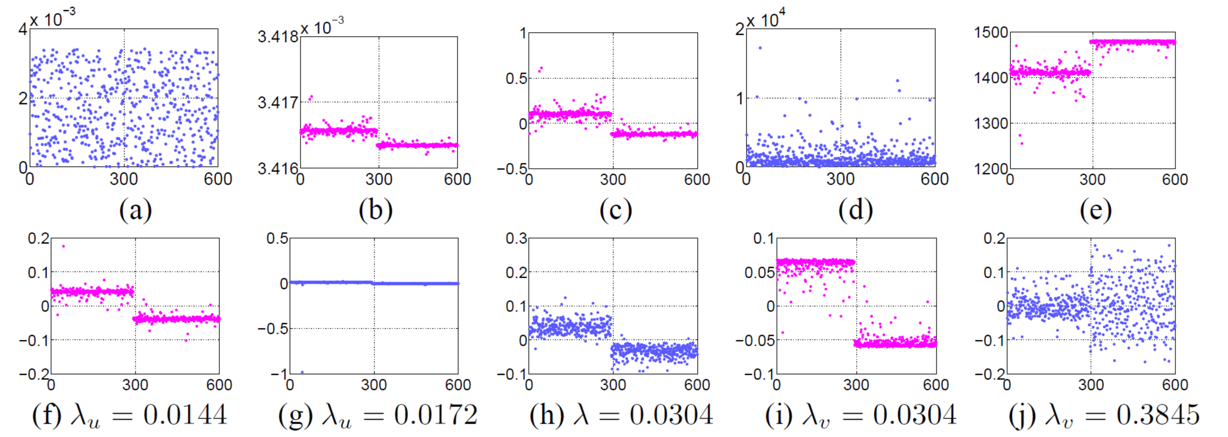
- $\mathcal{L}_h(\mathcal{S}_i)$ is the volume of $\mathcal{S}_i$.
- Variation of h only depends on $\Phi(\mathcal{S}_i)$.
- Our result is complementary to [6].

Eigenvectors of the Laplacian Matrices

$$Lu = \lambda_u u \qquad L_{rw} v = \lambda_v v \qquad L_{rw} := D^{-1}L \qquad L_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

Harmonic form: $u(i) = \sum_{k \sim i} \frac{w_{ik}}{d_i - \lambda_u} u(k), \quad v(i) = \sum_{k \sim i} \frac{w_{ik}}{d_i(1 - \lambda_v)} v(k)$

- The closer λ_u to $\min_i\{d_i\}$, the weaker of harmonic structure of u ;
- If $\lambda_v << 1$, v will enjoy a significant harmonic structure;
- This explains why eigenvectors of L_{rw} are preferred than those of L .



(a) (b) (c) (d) (e)

(f) $\lambda_u = 0.0144$ (g) $\lambda_u = 0.0172$ (h) $\lambda = 0.0304$ (i) $\lambda_v = 0.0304$ (j) $\lambda_v = 0.3845$

- (a) PARW – from one to others; (b) PARW – from others to one;
- (c) A row of $L^\dagger$; (d) HT – from one to others; (e) HT – from others to one;
- (f-g) Eigenvectors of L ($\min_i\{d_i\} = 0.0173$);
- (h) An eigenvector of L_{sym} ; (i-j) Eigenvectors of L_{rw} .

Experimental Results

Table 1: Classification accuracy on 9 datasets.

	USPS	YaleB	satimage	imageseg	ionosphere	iris	protein	spiral	soybean
ARW-N-1NN	**.879**	**.892**	**.777**	**.673**	**.771**	**.918**	**.589**	**.830**	**.916**
ARW-1NN	.445	.733	.650	.595	.699	.902	.440	.754	.889
ARW-CMN	.775	.847	.741	.624	.724	.894	.511	.726	.856
LGC	.821	.884	.725	.638	.731	.903	.477	.729	.816
PARW ($\Lambda = I$)	.880	.906	.781	.665	.752	.928	.572	.835	.905

Table 2: Ranking results (MAP) on USPS.

Digits	0	1	2	3	4	5	6	7	8	9	All
$\Lambda = R$ (column)	**.981**	**.988**	**.875**	**.892**	**.647**	**.780**	**.941**	**.918**	**.746**	**.731**	**.850**
$\Lambda = R$ (row)	.169	.143	.114	.096	.092	.076	.093	.093	.075	.086	.103
$\Lambda = I$	.981	.988	.876	.893	.646	.778	.940	.919	.746	.730	.850

Table 3: Classification accuracy on USPS.

k-NN unweighted graphs	10	20	50	100	200	500
HT($\mathcal{L} \to \mathcal{U}$)	**.8514**	**.8361**	**.7822**	**.7500**	**.7071**	**.6429**
HT($\mathcal{U} \to \mathcal{L}$)	.1518	.1454	.1372	.1209	.1131	.1113
$L^\dagger$	.8512	.8359	.7816	.7493	.7062	.6426

ARW-CMN – [1]

LGC - "Learning with local and global consistency", Zhou et al., NIPS'04.

$\Lambda = R$ - A random positive diagonal matrix.

HT($\mathcal{L} \to \mathcal{U}$) - From all labeled points to hit one unlabeled point.

HT($\mathcal{U} \to \mathcal{L}$)- From one unlabeled point to hit all labeled points.

[1] "Semi-supervised learning using Gaussian fields and harmonic functions." Zhu et al., ICML'03.
[2] "Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data", Nadler et al., NIPS'09.
[3] "Learning with Partially Absorbing Random Walks." Wu et al., NIPS'12.
[4] "Local graph partitioning using pagerank vectors." Anderson et al., FOCS'06.
[5] "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation." Fouss et al., TKDE, 2007.
[6] "Getting lost in space: Large sample analysis of the commute distance." Luxburg et al., NIPS' 2010.