Learning with Partially Absorbing Random Walks

Xiao-Ming Wu[1] Zhenguo Li[1] Anthony Man-Cho So[2] John Wright[1] Shih-Fu Chang[1]
[1]Columbia University [2]The Chinese University of Hong Kong

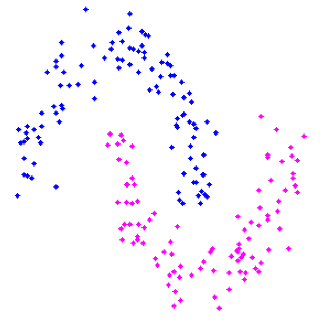
Motivation

- Problems with Random Walks
  - Stationary distribution $\propto (d_1, d_2, \ldots, d_n)^\top$
  - Hitting and commute times

$$\frac{1}{\mathrm{vol}(G)} H_{uv} \approx \frac{1}{d_v} \qquad \frac{1}{\mathrm{vol}(G)} C_{uv} \approx \frac{1}{d_u} + \frac{1}{d_v}$$

Though intended to model non-local structure, a random walk model may end up capturing only local information.

- Cluster Assumption
  - Semantics vary slowly in regions of high density.
    - Learned boundary should be placed in regions of low density.

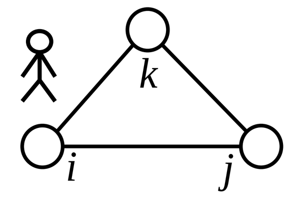


- Our Goals
  - A new rigorous model to implement cluster assumption.
  - A unifying view of existing random walk models that provides new deeper insights.
  - Better retrieval and classification accuracy.

Our Model: Partially Absorbing Random Walks (PARWs)

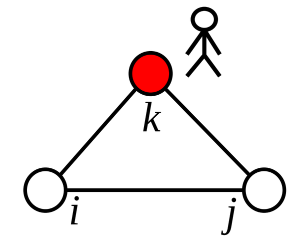
The Partially Absorbing Concept

States: Non-absorbing Fully-absorbing Partially-absorbing



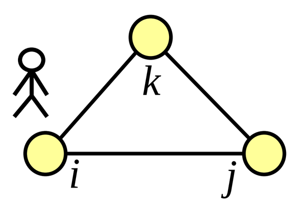
Ordinary Random Walks:

At each step, follow a random edge out.



Absorbing Random Walks:

At each step, either follow a random edge out or get absorbed at current state.



Partially Absorbing Random Walks:

At each step, there is some probability to be absorbed at the current state.

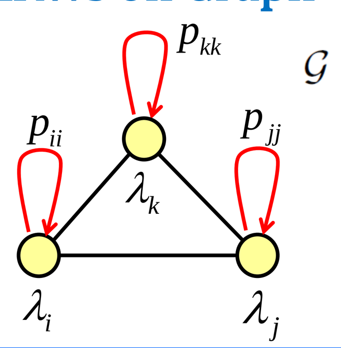
Definition & Transition Probabilities

PARWs are defined by second-order Markov Chain:

$$\mathbb{P}(X_{t+2}=j|X_{t+1}=i, X_t=k) = \begin{cases} 1, & i=j, i=k, \\ 0, & i\neq j, i=k, \\ \mathbb{P}(X_{t+2}=j|X_{t+1}=i)=p_{ij}, & i\neq k, \end{cases}$$

but fully specified by the first–order transition probabilities p_{ij}.

PARWs on Graph



$\mathcal{G}=(\mathcal{V},W)$ $W=[w_{ij}]\in\mathbb{R}^{n\times n}$ $\lambda_1,\ldots,\lambda_n\geq 0$ $d_i=\sum_j w_{ij}$

$D=\mathrm{diag}(d_1,d_2,\ldots,d_n)$ $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$

$$p_{ij} = \begin{cases} \frac{\lambda_i}{\lambda_i+d_i}, & i=j, \\ \frac{w_{ij}}{\lambda_i+d_i}, & i\neq j. \end{cases}$$

p_{ii} — Absorption rates

define

Λ — Regularizer matrix

First–order transition probability matrix:

$$P=(\Lambda+D)^{-1}(\Lambda+W)$$

Absorption Probabilities

What is the probability that a PARW starting from i gets absorbed by j?

Absorption probability matrix:

$$A=[a_{ij}]\in\mathbb{R}^{n\times n}$$

By first step analysis:

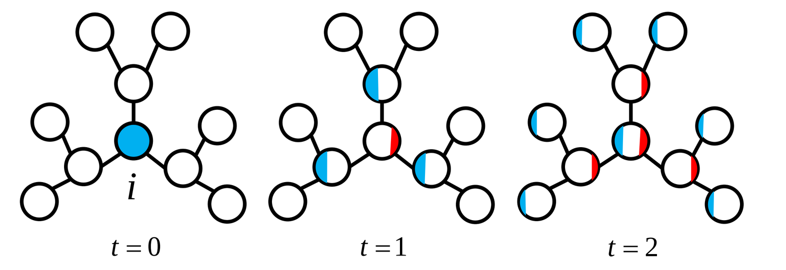
$$\begin{cases} a_{ii} = \frac{\lambda_i}{\lambda_i+d_i}\times 1 + \sum_{j\neq i}\frac{w_{ij}}{\lambda_i+d_i}a_{ji}, \\ a_{ij} = \sum_{k\neq i}\frac{w_{ik}}{\lambda_i+d_i}a_{kj}. \end{cases}$$

$$\boxed{A=(\Lambda+L)^{-1}\Lambda}$$

where $L=D-W$ is the graph Laplacian.

Equivalent Interpretations

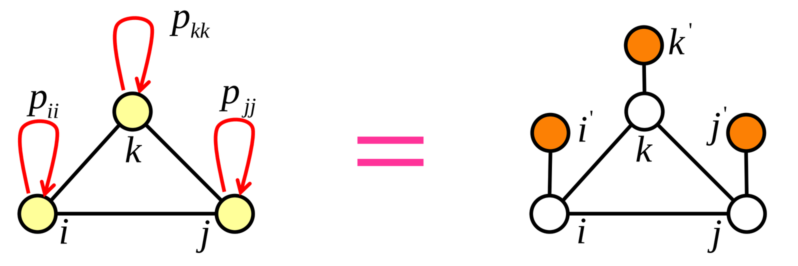
Flow Diffusion



Blue – Running flow Red – Retained flow

a_{ij} = the amount of flow stored in vertex j

Absorbing Random Walks



Partially absorbing random walks = Absorbing random walks + N fully-absorbing states

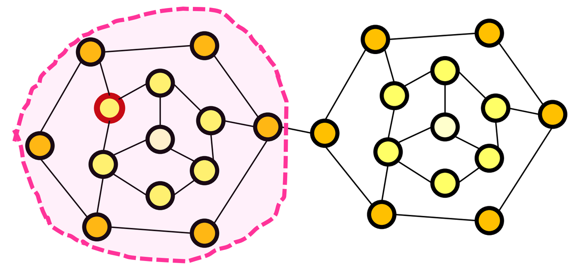
$a_{ij} = a_{ij'}$

Model Selection: Implementing the Cluster Assumption

A Unifying Framework

- PageRank is a PARW with constant absorption rates β (i.e., the restart probability).
- Absorbing random walks are PARWs with 0/1 absorption rates.
- 6 popular label propagation models are PARWs with different absorption rates, from
  - Zhu et al., ICML'03
  - Zhou et al., NIPS'04
  - Chapelle et al., AISTATS'05
  - Kveton et al., AISTATS'10
  - Bengio et al., 2006

Which model is more desirable? And in what sense?

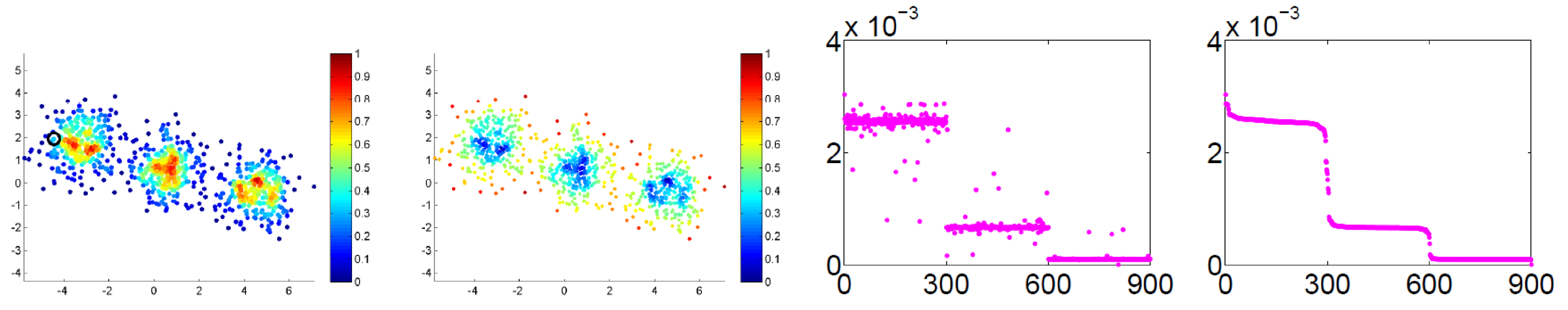


A PARW starting from a set $\mathcal{S}$ of low conductance is mostly absorbed in $\mathcal{S}$.

— $\mathcal{S}$ — The starting vertex

With $\Lambda=\alpha I$ $p_{ii}=\frac{\lambda_i}{\lambda_i+d_i}=\frac{\alpha}{\alpha+d_i}$

High/low density regions have low/high absorption rates.



Graph density Absorption rates Absorption probabilities Sorted Absorption probabilities

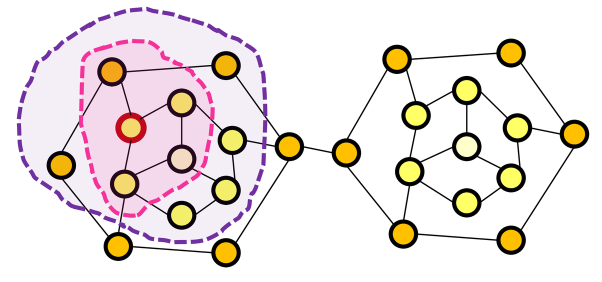
Intuition: building bounding walls around clusters to prevent the random walk from escaping.

The absorption probabilities vary slowly inside the cluster, while dropping sharply outside.

— The starting vertex

— $\mathcal{S}_j$ — $\mathcal{S}_k$

Lighter colors indicate smaller absorption rates.



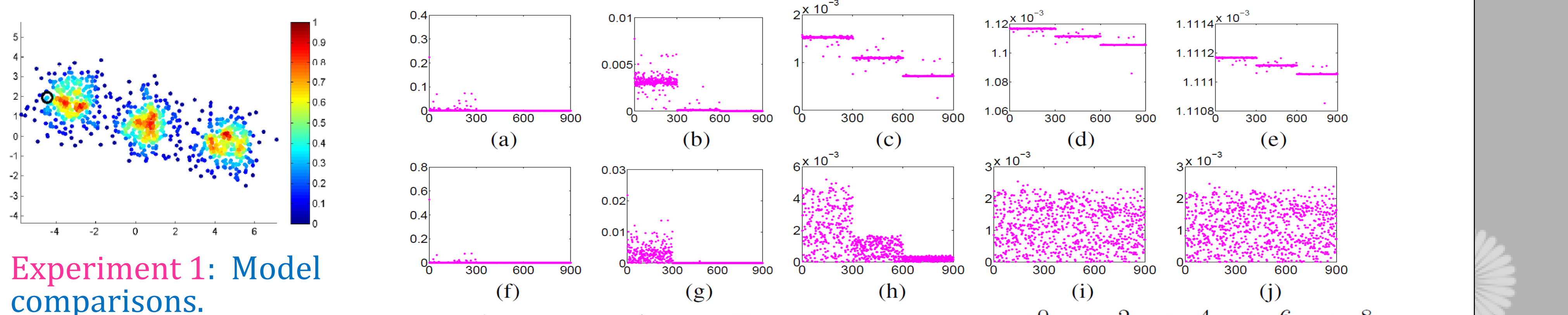
If $\Phi(\mathcal{S}_j)\geq 2\phi$, then there exists a $k>j$ such that

$$d(\mathcal{S}_k)\geq(1+\phi)d(\mathcal{S}_j) \quad \textit{and} \quad \mathbf{a}(k)\geq\mathbf{a}(j)-\frac{\alpha\left(1-\sum_{k=1}^{j}\mathbf{a}(k)\right)}{\phi d(\mathcal{S}_j)}.$$

If $\mathcal{S}_j$ has high conductance, then there will be a set $\mathcal{S}_k$ much larger than $\mathcal{S}_j$ where $\mathbf{a}(k)$ will remain large.

$\Phi(\mathcal{S}) = \frac{w(\mathcal{S},\bar{\mathcal{S}})}{\min(d(\mathcal{S}),d(\bar{\mathcal{S}}))}$ is the conductance of the vertex set $\mathcal{S}$ with cut $w(\mathcal{S},\bar{\mathcal{S}}) := \sum_{(i,j)\in e(\mathcal{S},\bar{\mathcal{S}})} w_{ij}$ and volume $d(\mathcal{S}) := \sum_{i\in S} d_i$

Simulation



Experiment 1: Model comparisons.

(top) $\Lambda=\alpha I$ vs. $\Lambda=\alpha D$ (bottom) $\alpha = 10^0, 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}$

Experiment 2: Image retrieval results (MAP) on USPS all 9298 images.

Digits	0	1	2	3	4	5	6	7	8	9	All
PARW with $\Lambda=\alpha I$	**.981**	**.988**	**.876**	**.893**	**.646**	**.778**	**.940**	**.919**	**.746**	**.730**	**.850**
PageRank	.886	.972	.608	.764	.488	.568	.837	.825	.626	.702	.728
Manifold Ranking	.957	.987	.827	.827	.467	.630	.917	.822	.675	.719	.783
Euclidean Distance	.640	.980	.318	.499	.337	.294	.548	.620	.368	.480	.508

Experiment 3: Semi-supervised learning on USPS.

HMN	LGC	$\Lambda=\alpha D$	$\Lambda=\alpha I$
$.782\pm.068$	$.792\pm.062$	$.787\pm.048$	$\mathbf{.881\pm.039}$

PageRank - "The pagerank citation ranking: Bringing order to the web", Page et al., 1999
Manifold Rank - "Ranking on Data Manifold", Zhou et al., NIPS'04
HMN - "Semi-supervised learning using Gaussian fields and harmonic functions", Zhu et al., ICML'03.
LGC - "Learning with local and global consistency", Zhou et al., NIPS'04.