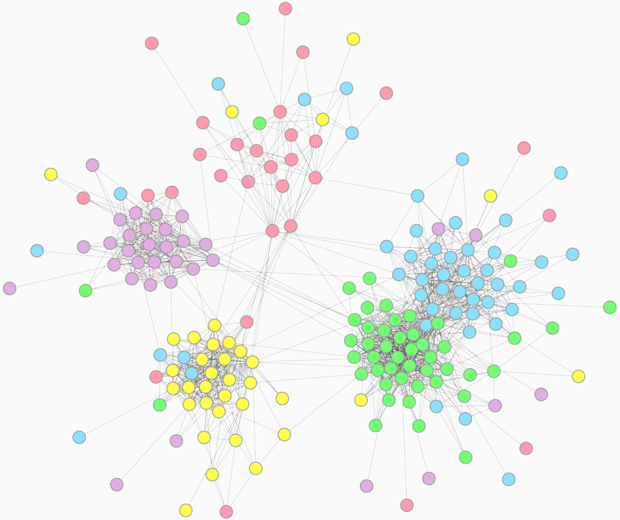
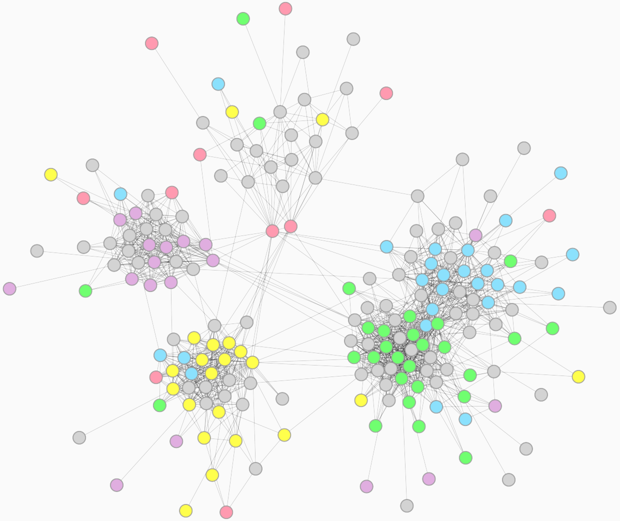
Active Learning of Convex Halfspaces on Graphs

Maximilian Thiessen
Thomas Gärtner

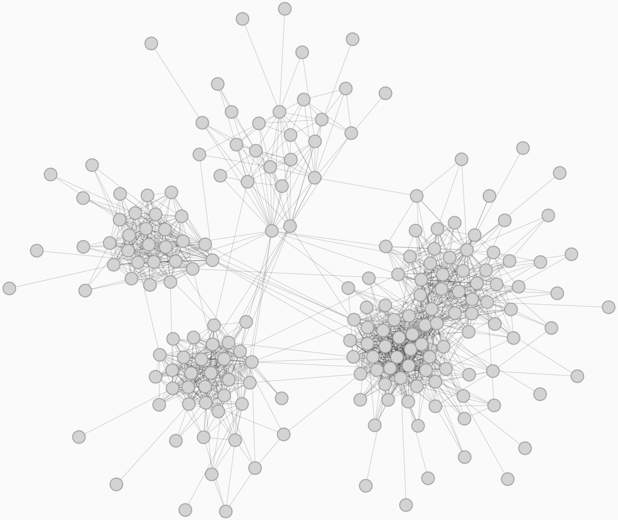
TU Wien
Vienna | Austria
Research Unit Machine Learning



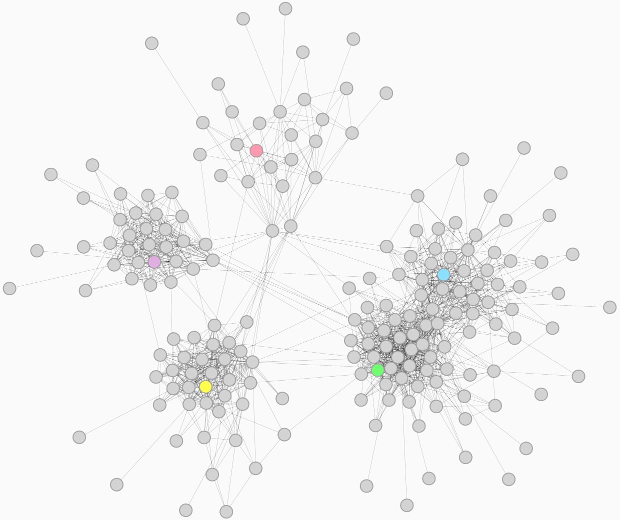
email-Eu-core (SNAP datasets)



email-Eu-core (SNAP datasets)



email-Eu-core (SNAP datasets)



email-Eu-core (SNAP datasets)

Active learning

Automate data annotation:

- unlabelled data is readily available
- labels are expensive and tedious
- reduce the number of labels to learn a good classifier

Active learning is well-established

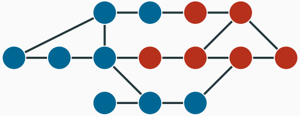
in theory: PAC inspired results

in practice: self-driving cars, speech recognition, drug discovery

Active vertex classification

Given a graph $G = (V, E)$

- vertices V represent the data
- edges E representing similarity
- fixed unknown labels $\{$●,●$\}$



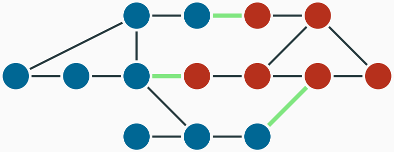
Goal:

Learn labels using as few as possible iterative vertex queries

Previous results: cut-based bounds

Query complexity: number of queries required to correctly identify the labelling

Cut-based bounds [Afshani, et al. 2007, Dasarathy, et al. 2015]



- cut of the labelling C: set of edges going from one class to the other
- cut border ∂C: set of vertices incident to C
- query complexity:

$$\mathcal{O}\left(|\partial C| \log |V|\right)$$

Previous results: cut-based bounds

Query complexity:

$$\mathcal{O}\left(|\partial C| \log |V|\right)$$

Restrictions:

- labels must be **balanced**
- bound is **label dependent**

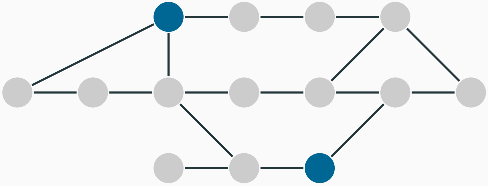
size of the cut border ∂C can be **large** or even **unknown**

Label-independent bounds

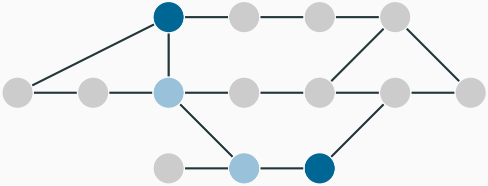
Our goal: label-independent bounds

- only depend on G
- do not depend on labels
- practitioners get a cost estimate before the data annotation
- need assumptions on labels

Geodesic convexity assumption



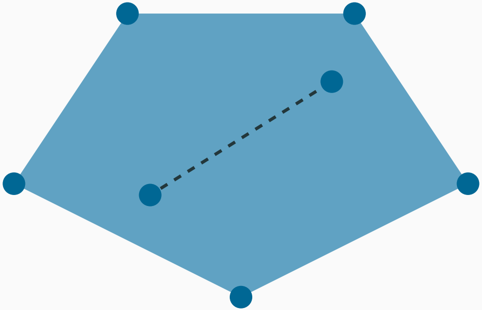
Geodesic convexity assumption



vertices have same label $\Rightarrow$ vertices on connecting shortest path have the label

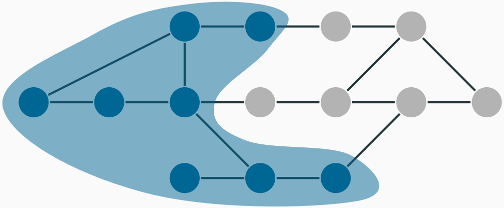
Convexity in Euclidean space

Set is **convex**: contains all **connecting line segments**



Geodesic convexity on graphs

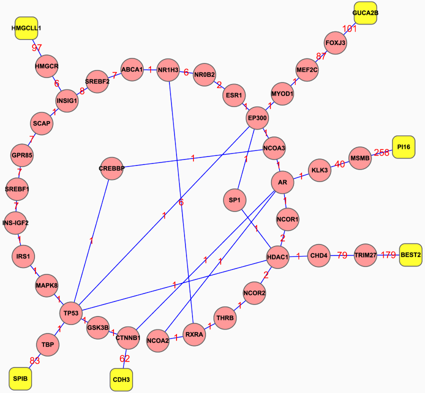
Vertex set is **convex**: contains all **connecting shortest paths**



Convex hull $\sigma(X)$ is the smallest convex vertex set containing X

Convexity in real-world graphs

Cancer-related genes share similarity along shortest paths



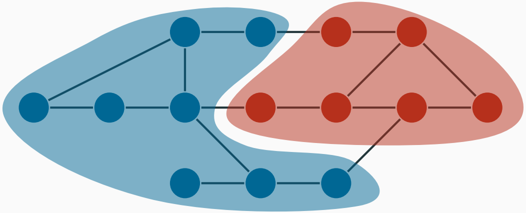
[Bi-Qing Li, et al. 2012]

Convexity in real-world graphs

dataset	convex communities
DBLP	4308/5000
Amazon	3999/5000
Youtube	2990/5000
LiveJournal	1649/5000
Orkut	363/5000
Eu-core	7/42

[SNAP datasets]

Halfspaces on graphs



Vertex set C is a **halfspace**, if C and $V \setminus C$ are convex

Assumption: blue subgraph and red subgraph are halfspaces

Upper bound on the query complexity

Query complexity:

$$\mathcal{O}(h(G) + \log d(G) + \operatorname{tw}(G))$$

Upper bound on the query complexity

Query complexity:

$$\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$$

diameter $d(G)$

Upper bound on the query complexity

Query complexity:

treewidth $\mathrm{tw}(G)$
(small for e.g., molecules)

$$\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$$

diameter $d(G)$

Upper bound on the query complexity

Query complexity:

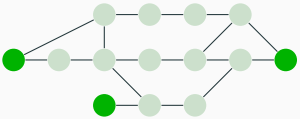
minimum hull set size $h(G)$

treewidth $\mathrm{tw}(G)$
(small for e.g., molecules)

$$\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$$

diameter $d(G)$

A set $H \subseteq V(G)$ is a hull set if $\sigma(H) = V(G)$



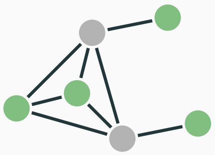
Upper bound on the query complexity

$$\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$$

General lower bound

A vertex x is extreme, if $V \setminus \{v\}$ is convex

- generalisation of leaves
- set of extreme vertices $\text{Ext}(G)$



Query complexity is

$$\Omega(|\text{Ext}(G)|)$$

Can be far away from

$$\mathcal{O}(h(G) + \log d(G) + \text{tw}(G))$$

Lower bounds

Upper bound:

- $\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$

Lower bounds along **separation axioms** [van de Vel 1993]

Lower bounds

Upper bound:

- $\mathcal{O}(h(G) + \log d(G) + \text{tw}(G))$

Lower bounds along **separation axioms** [van de Vel 1993]

S_1**:** any singleton $v \in V$ is convex

$\Omega(|\text{Ext}(G)|)$

Lower bounds

Upper bound:

- $\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$

Lower bounds along **separation axioms** [van de Vel 1993]

S_1**:** any singleton $v \in V$ is convex

$$\Omega(|\mathrm{Ext}(G)|)$$

S_2**:** any pair of vertices $v \neq w$ is halfspace separable

$$\Omega(|\mathrm{Ext}(G)| + \log d(G))$$

Lower bounds

Upper bound:

- $\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$

Lower bounds along **separation axioms** [van de Vel 1993]

S_1: any singleton $v \in V$ is convex

$$\Omega(|\mathrm{Ext}(G)|)$$

S_2: any pair of vertices $v \neq w$ is halfspace separable

$$\Omega(|\mathrm{Ext}(G)| + \log d(G))$$

S_3: any convex set C and $v \in V \setminus C$ are halfspace separable

$$\Omega(h(G) + \log d(G))$$

Lower bounds

Upper bound:

- $\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$

Lower bounds along **separation axioms** [van de Vel 1993]

S_1: any singleton $v \in V$ is convex

$$\Omega(|\mathrm{Ext}(G)|)$$

S_2: any pair of vertices $v \neq w$ is halfspace separable

$$\Omega(|\mathrm{Ext}(G)| + \log d(G))$$

S_3: any convex set C and $v \in V \setminus C$ are halfspace separable

$$\Omega(h(G) + \log d(G))$$

S_4: any two disjoint convex sets are halfspace separable

$$\Omega(h(G) + \log d(G) + \mathrm{Radon}(G))$$

Radon number

Radon partition R_1, R_2 of a set R:

- $R_1 \cup R_2 = R$, $R_1 \cap R_2 = \emptyset$
- $\sigma(R_1) \cap \sigma(R_2) \neq \emptyset$

Radon number: Smallest number r such that any set of size r has a Radon partition

Radon number

Radon partition R_1, R_2 of a set R:

- $R_1 \cup R_2 = R$, $R_1 \cap R_2 = \emptyset$
- $\sigma(R_1) \cap \sigma(R_2) \neq \emptyset$

Radon number: Smallest number r such that any set of size r has a Radon partition

VC dimension of halfspaces is $\leq \text{Radon}(G) - 1$

Radon number

Radon partition R_1, R_2 of a set R:

- $R_1 \cup R_2 = R$, $R_1 \cap R_2 = \emptyset$
- $\sigma(R_1) \cap \sigma(R_2) \neq \emptyset$

Radon number: Smallest number r such that any set of size r has a Radon partition

$$\text{VC dimension of halfspaces is} \leq \text{Radon}(G) - 1$$

Remarks:

- $\mathbb{R}^n$ has VC dimension $n+1$ and Radon number $n+2$
- For S_4 graphs the VC dimension is exactly $\text{Radon}(G) - 1$.

Lower bounds

Upper bound:

- $\mathcal{O}(h(G) + \log d(G) + \mathrm{tw}(G))$

Lower bounds along **separation axioms** [van de Vel 1993]

S_1: any singleton $v \in V$ is convex

$$\Omega(|\mathrm{Ext}(G)|)$$

S_2: any pair of vertices $v \neq w$ is halfspace separable

$$\Omega(|\mathrm{Ext}(G)| + \log d(G))$$

S_3: any convex set C and $v \in V \setminus C$ are halfspace separable

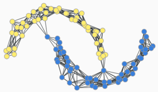
$$\Omega(h(G) + \log d(G))$$

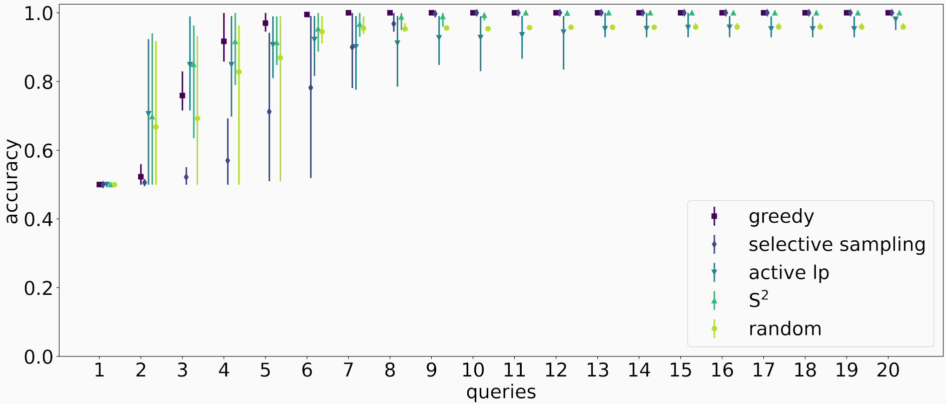
S_4: any two disjoint convex sets are halfspace separable

$$\Omega(h(G) + \log d(G) + \mathrm{Radon}(G))$$

Experiments

Two moons dataset





Conclusions and future directions

We characterised the query complexity of learning halfspaces in graphs

- tight bounds along separation axioms
- identified the Radon number as an important parameter
- more details in the paper (proofs, computational runtime, ...)

Conclusions and future directions

We characterised the query complexity of learning halfspaces in graphs

- tight bounds along separation axioms
- identified the Radon number as an important parameter
- more details in the paper (proofs, computational runtime, ...)

Future research directions:

- learning halfspaces in general convexity spaces
- more efficient algorithms
- more robust and practical assumptions

Conclusions and future directions

We characterised the query complexity of learning halfspaces in graphs

- tight bounds along separation axioms
- identified the Radon number as an important parameter
- more details in the paper (proofs, computational runtime, ...)

Future research directions:

- learning halfspaces in general convexity spaces
- more efficient algorithms
- more robust and practical assumptions

Thanks for listening!

See you in the poster session