Local Clustering in Hypergraphs
(Mixing in irregular hypergraphs)

Stefano Huber

Department of Computer Science
EPFL

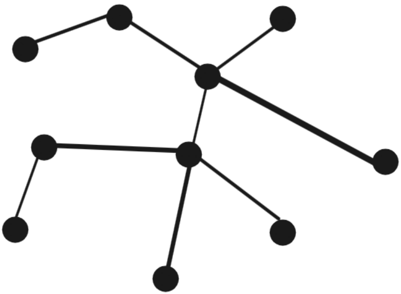
July 7, 2022

Table of Contents

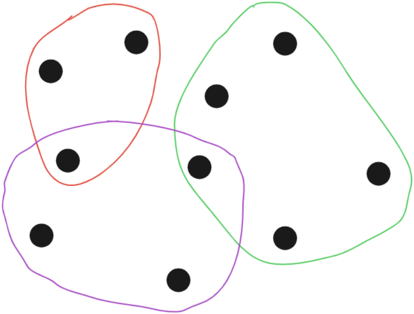
EPFL

Definitions

Graph($n = 10, m = 9$)



Hypergraph($n = 10, m = 3$)



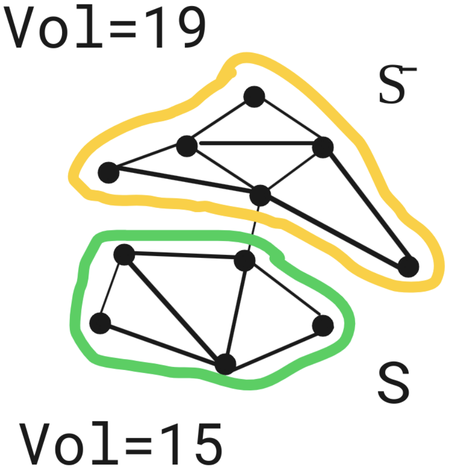
Conductance

An interesting property of a graph is its conductance ϕ:

Conductance

$$\phi(S) = \frac{E(S,\bar{S})}{\min(\mathrm{vol}(S),\mathrm{vol}(\bar{S}))}$$

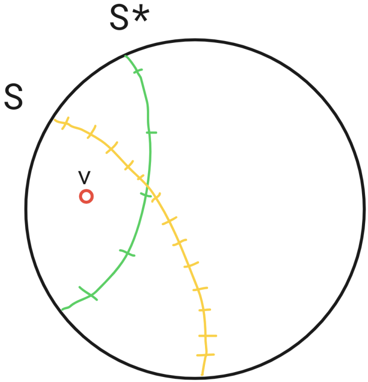
The conductance of the cut $(S, \bar{S})$ in the figure is $\frac{1}{15}$



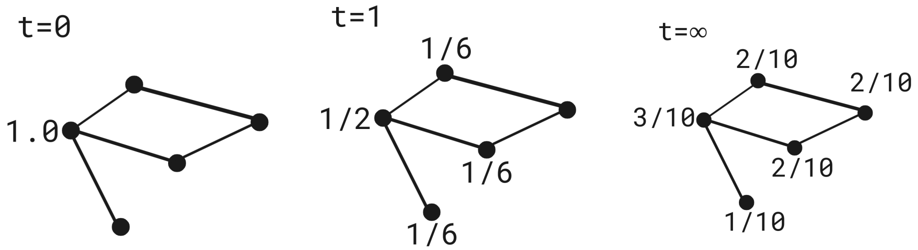
Local Clustering

Local Clustering Problem

Given a starting vertex v and a target conductance $\hat{\phi}$, find a cut S s.t. $v \in S$ and $\phi(S) \leq \hat{\phi}$. Must assume that $\exists S^*$ s.t. $v \in S^*$ and $\phi(S^*) \leq \frac{\hat{\phi}}{\log(n)} = \phi^*$.



Random Walks in graphs



Random walks are *lazy* (w.p. $\frac{1}{2}$ you do not move)

Transition probability matrix

$p_{t+1} = \frac{1}{2}(I + AD^{-1})p_t = Mp_t$

Convergence to stationary distribution

when $t \to \infty \implies \mathrm{p}_t(u) \to \pi(u) = \frac{d(u)}{\mathrm{vol}(G)}$

Mixing

Studying how fast the probability vector converges to stationary distribution is called *mixing*, and it is done with the Lovasz-Simonovits curve [Lovász and Simonovits, 1993].

Lovasz-Simonovits curve

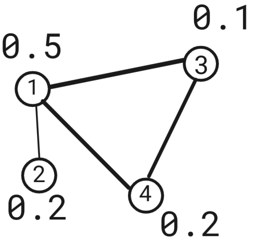
Sweep Cut

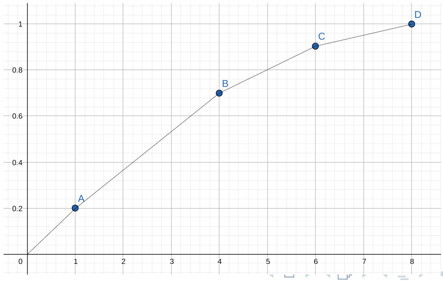
$S_j(p_t)$: sort vertices by decreasing $\frac{p_t(u)}{d(u)}$ values, and take first j vertices.

sorted vertices $= [2, 1, 3, 4]$

volumes of sweep cuts $= [1, 4, 6, 8]$

probability mass on sweep cuts $= [0.2, 0.7, 0.9, 1.0]$





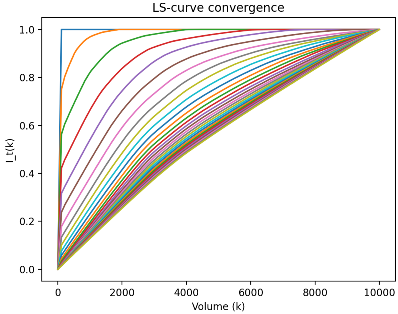
Lovasz Simonovits Curve properties

Properties:

- concave
- decreasing wrt time
- bounded in [0,1]

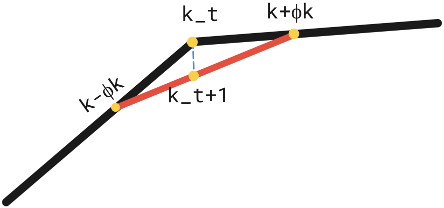
LS-curve is decreasing

$I_{t+1}(k) \leq I_t(k)$



1. Studying mixing with Lovasz-Simonovits curve

- $\hat{\phi} = \min_{j \in [1,n]} \phi(S_j(\mathrm{p}_t))$ and $\hat{k} := \min(k, \mathrm{vol}(G) - k)$



Recursive upper bound of LS curve

$$I_{t+1}(k) \leq \frac{1}{2}(I_t(k - \hat{\phi}\hat{k}) + I_t(k + \hat{\phi}\hat{k}))$$

- This allows for exponentially fast convergence:

Exponentially fast mixing wrt time

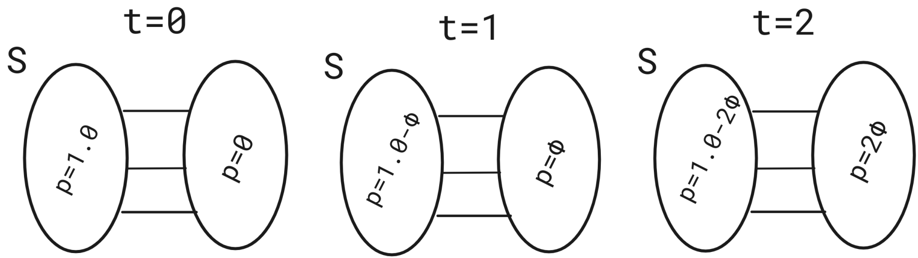
$$I_t(k) - \pi(S_j(\mathrm{p}_t)) \leq \sqrt{\hat{k}} e^{-t\hat{\phi}^2}$$

Leaking of random walks in a graph

Leaking is the opposite of mixing: it says how slowly the random walk mixes. When $p_0 = \psi_S$ for some $S \subseteq V$, then

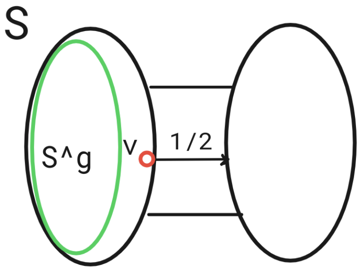
Leaking wrt optimal conductance ϕ^*

$p_t(S) \geq 1 - t\phi(S)$



Leaking of random walks in a graph

Unfortunately, when starting with $p_0 = \chi_v$ for some $v \in S$ the leaking result might not be true.



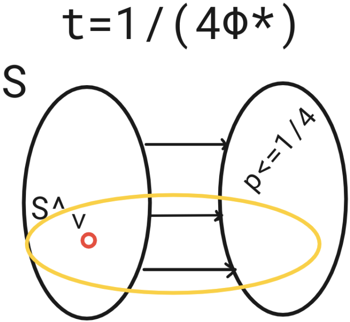
Luckily though, volume of these bad vertices S^b for which the leaking result is not true is small: hence the volume of good vertices S^g is large:

Volume of S^g is large

$\mathrm{vol}(S^g) \geq \frac{1}{2}\mathrm{vol}(S)$

Mixing+leaking = clustering algorithm

You pick a vertex v at random according to the stationary distribution (with good probability it falls in S^g)



$$\frac{1}{4} \leq p_t(S) - \pi(S) \leq \sqrt{\mathrm{vol}(S)} e^{-t\hat{\phi}^2}$$

$$\Longrightarrow$$

$$\hat{\phi} \leq \sqrt{\log(n)\phi^*} \qquad (1)$$

So the conductance of the output sweep cut $\phi(\hat{S}) = \hat{\phi}$ is *not too far* from the optimal conductance $\phi(S) = \phi^*$ [Spielman and Teng, 2008].

Local clustering algorithm for hypergraphs

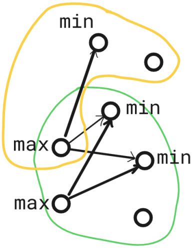
- In order to find a proper clustering algorithm for hypergraphs, we need to generalize both mixing and leaking results to hypergraphs.
- Mixing: there have attempts to prove mixing using continuous diffusion processes [Takai et al., 2020].
- Leaking: no known leaking result as far as we know.

Why mixing is hard in hypergraphs?

- Proving mixing in hypergraphs is harder because it is not possible to define an equivalent discrete diffusion process as a random walk.
- The Laplacian operator $\mathcal{L}$ s.t. $\frac{d\mathrm{p}_t}{dt} = -\mathcal{L}(D^{-1}\mathrm{p}_t)$ can be defined for B_e the convex hull of $\{\chi_v - \chi_u : u, v \in e\}$ [Chan et al., 2016] as:

Laplacian for hypergraphs

$\mathcal{L}(\mathrm{x}) = \{\sum_{e\in E} w_e \mathrm{b}_e \mathrm{b}_e^T \mathrm{x} \mid \mathrm{b}_e = \arg\max_{\mathrm{b}\in B_e} \mathrm{x}^T \mathrm{b}\}$



Discrete diffusion process

- Idea: turn the hypergraph laplacian into a matrix by solving ties *arbitrarily*.
- Question: is this simple resolution of ties enough in order to obtain a diffusion process with good mixing properties?

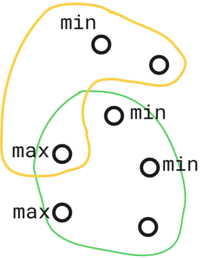
Discrete diffusion process

- We collapse the hypergraph H into a multigraph G_t collapsing every hyperedge e into $(v^t_{\max}(e), v^t_{\min}(e))$

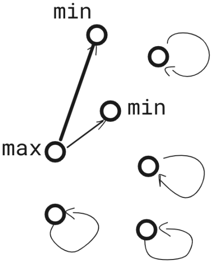
$v^t_{\max}(e)/v^t_{\min}(e)$

$$v^t_{\max}(e) = u \in e : \arg\max / \arg\min_{u \in e} \frac{p_t(u)}{d(u)}$$

Hypergraph H



Collapsed multigraph G_t



Discrete diffusion process

We want to use the Lovasz Simonovits recursive upper bound in order to prove mixing of this diffusion process.

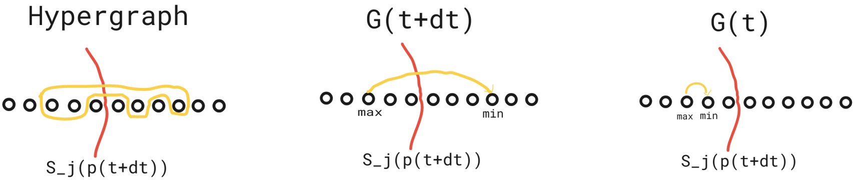
LS recursive upper bound for $dt \leq \frac{1}{2}$

$$I_{t+dt}(k) \leq (1-2dt)I_t(k) + dt(I_t(k-\hat{k}\hat{\phi}) + I_t(k+\hat{k}\hat{\phi})$$

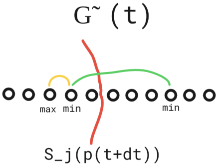
So we need the cut $S_j(p_{t+dt})$ to have conductance $\phi(S_j(p_{t+dt})) = \hat{\phi}$ in both the collapsed multigraphs G_t and G_{t+dt}.

Discrete diffusion process

Every hyperedge crossing the cut $S_j(p_{t+dt})$ is collapsed in the multigraph G_{t+dt} in such a way that it also crosses the cut. But, this is not necessarily true for the graph G_t.



We can add another edge to G_t which is certainly crossing the cut.



Discrete diffusion process

The proof of the mixing result goes like this:

- First, prove that the LS-curve $\tilde{I}_t$ is smaller than I_t.

Lemma 1

$\forall t,\ I_{t+dt}(k) \leq \tilde{I}_{t+dt}(k)$

- Then, take advantage of the known conductance of the sweep cuts in $\tilde{G}_t$ to prove

Lemma 2

$\tilde{I}_{t+dt}(k) \leq (1-2dt)I_t(k) + 2dt(I_t(k-\hat{\phi}\hat{k}) + I_t(k+\hat{\phi}\hat{k}))$

- Which allows us to claim

Lemma 3

$I_t(k) - \pi(S_j(\mathrm{p}_t)) \leq \sqrt{\frac{k}{d(v_0)}}e^{-t\hat{\phi}^2}$

Is this result as powerful as the one for graphs?

Mixing in irregular hypergraphs

$$I_t(k) - \pi(S_j(\mathrm{p}_t)) \leq \sqrt{\frac{k}{d(v_0)}} e^{-\frac{t\hat{\phi}^2}{4}}$$

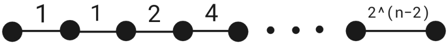
$\frac{k}{d(v_0)}$ can be as large as $\Omega(2^n)$, which means that even if the mixing time is large $\Omega(n)$, we can only have the guarantee that the output conductance is a constant.

Conductance is $O(1)$, but mixing is $O(n)$

Lemma 4

There exists a multigraph s.t. the conductance is $\frac{1}{2}$ but the mixing time is $O(n)$.

Here is an example:



- Nodes $[1, n]$ in a straight path.
- Edge $(i, i+1)$ has weight $w(i, i+1) = \sum_{j=0}^{i} w(i, i+1)$ so that weight doubles at every i.
- Total graph volume: 2^{n-1}.
- Conductance is $\frac{1}{2}$.
- Mixing time is $O(n)$ for a large fraction of nodes.

An argument for not having this issue in hypergraphs

Recall that the actual mixing theorem found with our analysis for the discrete diffusion process is

Improved mixing theorem for irregular hypergraphs

$$I_t(k) - \pi(S_j(\mathrm{p}_t)) \leq \sqrt{\frac{k}{d(v_0)}} e^{-\frac{\hat{\phi} t}{4}}$$

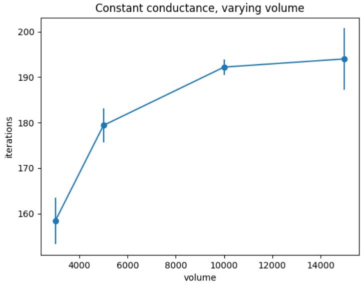
This means that when the volume of the hypergraph is exponential, then the probability of picking as starting node a vertex with non-exponential degree is very low.

Experiments

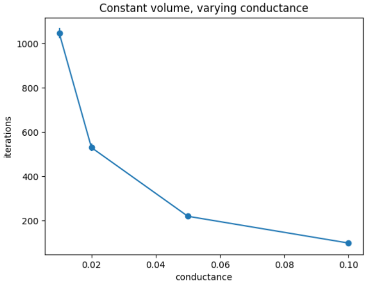
We performed some experiments to show empirically our mixing result.

- First, we want to prove that the mixing time grows logarithmically with the volume, and decreases quadratically with the conductance, namely $t = O\left(\frac{\log(n)}{\hat{\phi}^2}\right)$.
- We know that for r-uniform hypergraphs there is a theoretical upper-bound for the continuous diffusion process $t = O\left(\frac{\log(n)}{\hat{\phi}^2 r}\right)$. We want to see if such upper bound also holds for our simplified discrete diffusion process.

Experiment 1



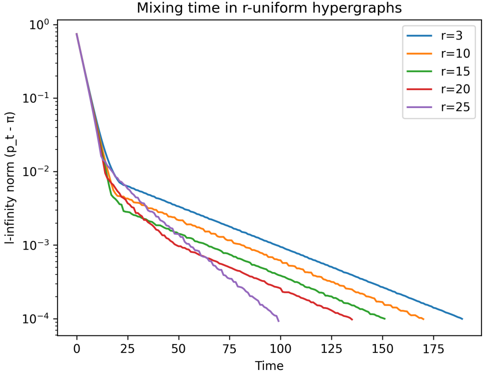
When the conductance is constant and the volume changes, the mixing time is *logarithmic* wrt the volume.



Instead, when the volume is constant and the conductance changes, then the mixing time decreases *quadratically* fast.

EPFL

Experiment 2



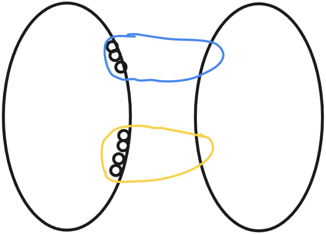
Indeed, the larger the r the smaller the mixing time, suggesting that also with our discrete diffusion process when the hypergraph is r-uniform it might hold $t = O\left(\frac{\log(n)}{r\hat{\phi}^2}\right)$.

Discussion about leaking

Leaking is indeed still an open problem in hypergraphs. In particular, it is indeed possible to prove that

Leaking starting from the stationary distribution on a set S

$\chi_S^T(\prod_{t' \leq t}(D_S M_{t'}))\psi_S \geq 1 - t\phi(S)$



When starting with probability centered in single vertices, instead, we count crossing hyperedges up to r times (only for the first iteration though).

Future work

For r-uniform hypergraphs we could prove an improved mixing theorem by a factor r, and a worse leaking theorem by a factor r. This would give a proper local clustering algorithm with the same guarantees as in graphs.

The end.

Thank you!

References I

Chan, T. H., Louis, A., Tang, Z. G., and Zhang, C. (2016).
Spectral properties of hypergraph laplacian and approximation algorithms.
CoRR, abs/1605.01483.

Lovász, L. M. and Simonovits, M. (1993).
Random walks in a convex body and an improved volume algorithm.
Random Struct. Algorithms, 4:359–412.

Spielman, D. A. and Teng, S. (2008).
A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning.
CoRR, abs/0809.3232.

Takai, Y., Miyauchi, A., Ikeda, M., and Yoshida, Y. (2020).
ACM.