Outliers and Hallucinations: Contributions to Robust Community Detection and Language Model Alignment

Leonardo Martins Bianco

Supervisors:

Christine Keribin (Laboratoire de Mathématiques d'Orsay)

Zacharie Naulet (INRAE)

Jessica Hoffmann (Google DeepMind)

December 4, 2025

Introduction

Thesis progress

Start of the thesis

Literature review,

Identification of a research problem, first efforts.

October 2022

October 2023

Start of *période de césure*

Collaboration with Tianlin Liu, Daniele Calandriello (*ICML* DeRA paper).

Start of PE-RL project with Jessica Hoffmann, Lucas Dixon.

March 2024

End of *période de césure*

Publication of *AISTATS* SubSearch paper.

Integration of Jessica into supervision, continuation of PE-RL project.

December 2025

Thesis defense

Thesis progress

Start of the thesis

Literature review,

Identification of a research problem, first efforts.

End of *période de césure*

Publication of *AISTATS* SubSearch paper.

Integration of Jessica into supervision, continuation of PE-RL project.

October 2023

March 2024

December 2025

October 2022

Start of *période de césure*

Collaboration with Tianlin Liu, Daniele Calandriello (*ICML* DeRA paper).

Start of PE-RL project with Jessica Hoffmann, Lucas Dixon.



Thesis defense

Thesis progress

Start of the thesis

Literature review,

Identification of a research problem, first efforts.

End of *période de césure*

Publication of *AISTATS* SubSearch paper.

Integration of Jessica into supervision, continuation of PE-RL project.

October 2022

October 2023

March 2024

December 2025

Thesis defense


Start of *période de césure*

Collaboration with Tianlin Liu, Daniele Calandriello (*ICML* DeRA paper).

Start of PE-RL project with Jessica Hoffmann, Lucas Dixon.


Thesis progress

October 2022

Start of the thesis

Literature review,

Identification of a research problem, first efforts.

October 2023

Start of *période de césure*

Collaboration with Tianlin Liu, Daniele Calandriello (*ICML* DeRA paper).

Start of PE-RL project with Jessica Hoffmann, Lucas Dixon.

March 2024

End of *période de césure*

Publication of *AISTATS* SubSearch paper.

Integration of Jessica into supervision, continuation of PE-RL project.

December 2025

Thesis defense

Overview

Part I: Contributions to Robust Community Detection

- Robust Estimation for the SBM

Part II: Contributions to Language Model Alignment

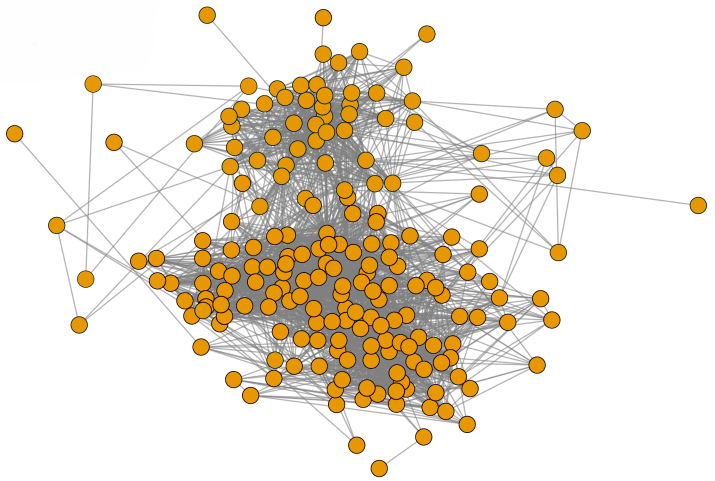
- Reducing Hallucinations with Synthetic Hallucinations
- Decoding-time Realignment of Language Models

Part I

Contributions to Robust Community Detection

Motivation

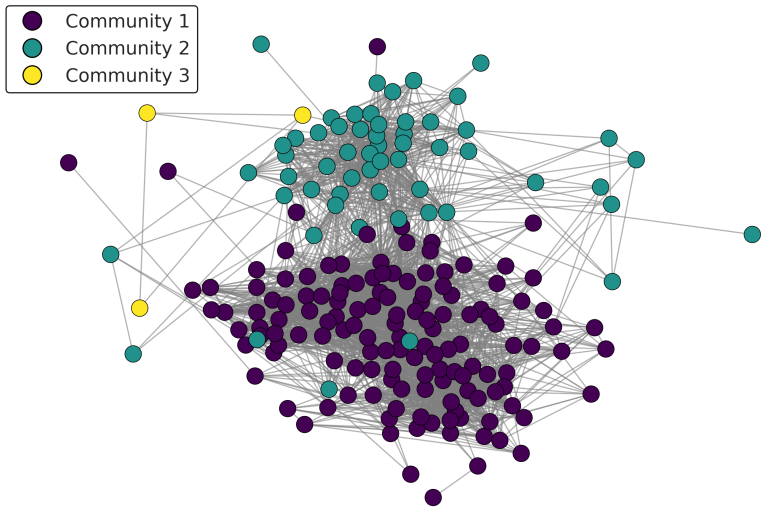
- *Community detection*: group similar nodes in a graph.
- Algorithms can be **highly sensitive** to *outlier* nodes.



Graph of Jazz collaborations in New York, Chicago, and Other Cities [Gleiser and Danon, 2003].

Motivation

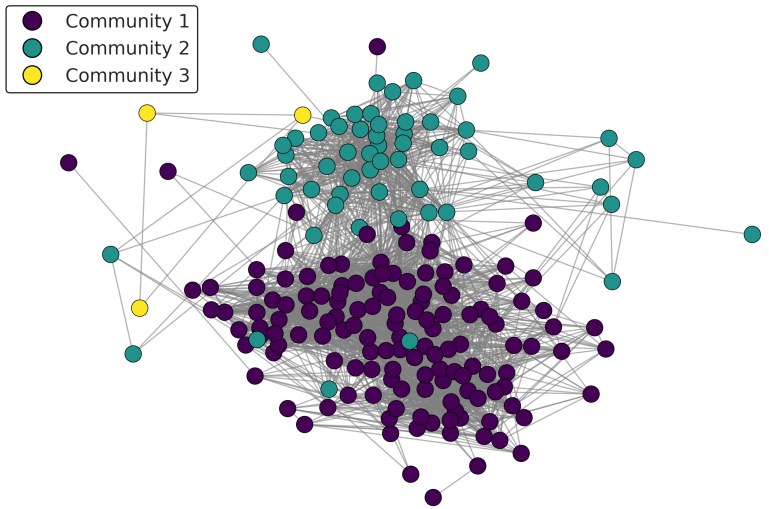
- *Community detection*: group similar nodes in a graph.
- Algorithms can be **highly sensitive** to *outlier* nodes.



Spectral clustering applied to the graph of Jazz collaborations [Gleiser and Danon, 2003].

Motivation

- *Community detection*: group similar nodes in a graph.
- Algorithms can be **highly sensitive** to *outlier* nodes.
- **Goal**: achieve "good" results despite outliers.



Spectral clustering applied to the graph of Jazz collaborations [Gleiser and Danon, 2003].

Graphs

Graph: nodes linked by edges.

$$G = (V, E)$$

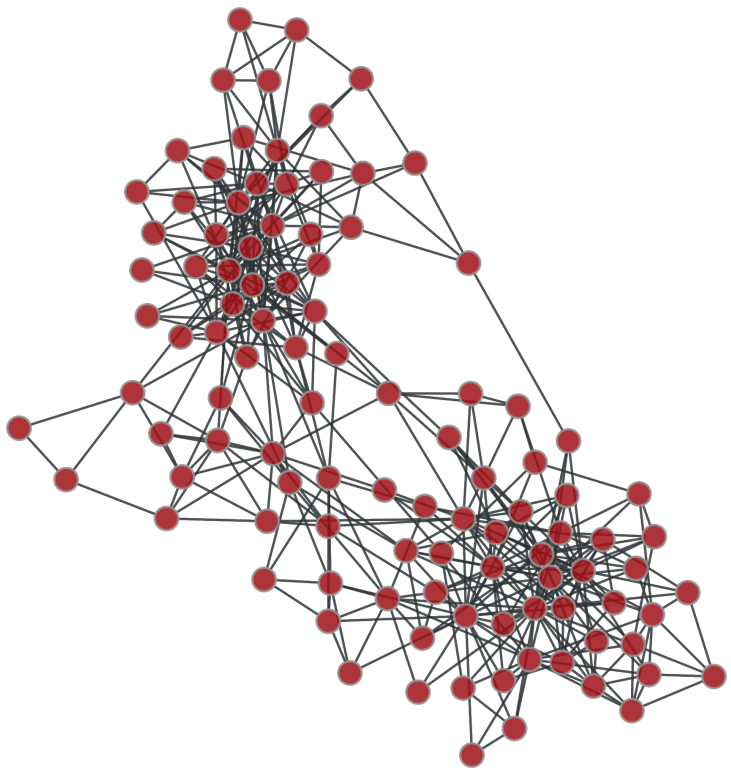
$$V = \{1, \ldots, n\}, E \subset V \times V$$

Undirected graph:

$$(i, j) \in E \Rightarrow (j, i) \in E$$

Adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$



The Stochastic Block Model [Holland et al., 1983]

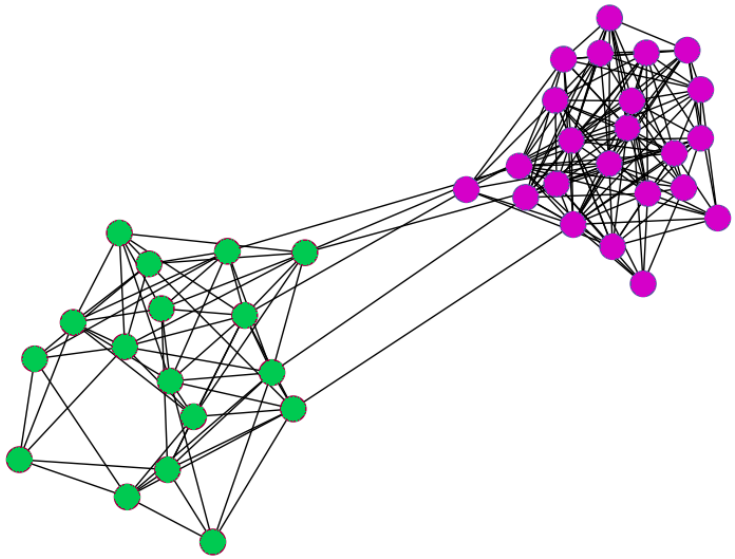
Probabilistic model

$Z_i \rightarrow$ community of node i

$K \rightarrow$ nb. of communities

$\pi_k \rightarrow$ size of community k

$\Gamma_{kl} \rightarrow$ connectivity k, l



The Stochastic Block Model [Holland et al., 1983]

Probabilistic model

$Z_i \rightarrow$ community of node i

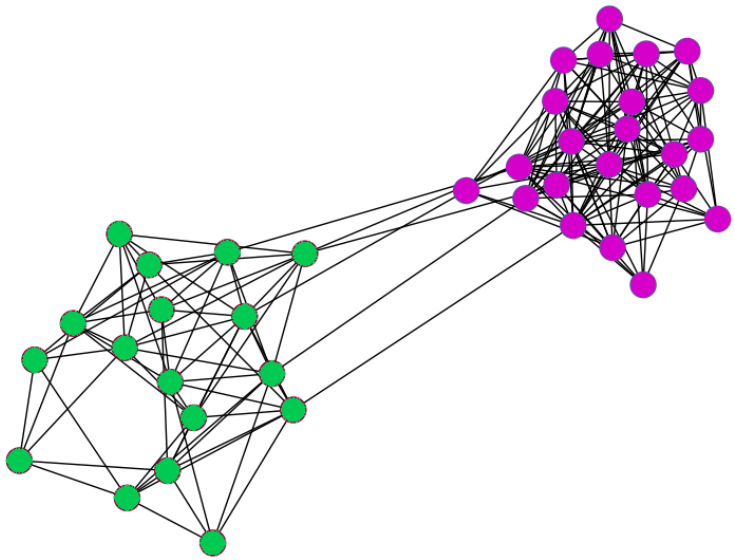
$K \rightarrow$ nb. of communities

$\pi_k \rightarrow$ size of community k

$\Gamma_{kl} \rightarrow$ connectivity k, l

$(Z, A) \sim \mathrm{SBM}_K(\pi, \Gamma)$

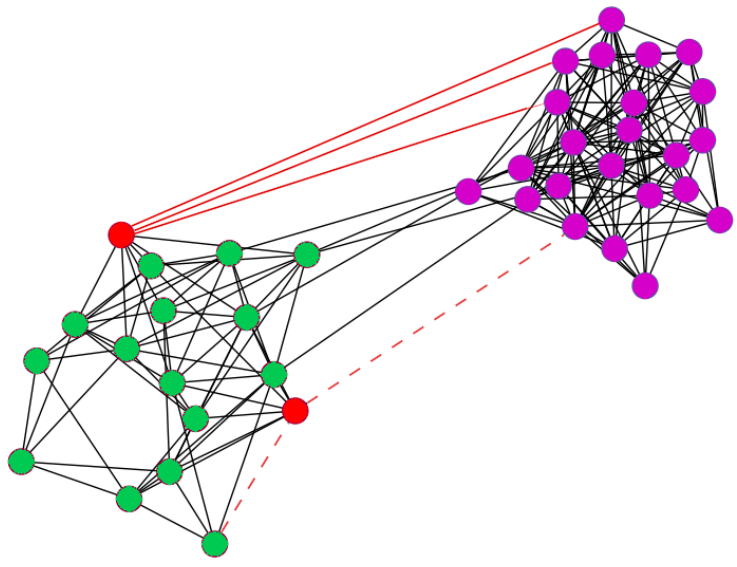
$$\begin{cases} \mathbb{P}(Z_i = k) = \pi_k \\ \mathbb{P}(A_{ij} = 1 | Z_i = k, Z_j = l) = \Gamma_{kl} \end{cases}$$



The Corrupted Stochastic Block Model

Adversary introduces outliers:

1. $(Z, A_{\text{pure}}) \sim \text{SBM}_K(\pi, \Gamma)$
2. Adversary *arbitrarily* changes edges of γn nodes
3. Corrupted A is observed



Research question

- Task: robustly estimate the connectivities Γ.
- For $K = 1$, Acharya et al. [2022].

Research question:

How to robustly estimate Γ for $K > 1$?

Results

- *Intuition:* find subgraph S excluding worst outliers.
- **First contribution**: extended bound from Acharya et al. [2022] to $K > 1$.

Results

- *Intuition:* find subgraph S excluding worst outliers.
- **First contribution**: extended bound from Acharya et al. [2022] to $K > 1$.

Theorem. Denote $\mathcal{I}$ the set of inlier nodes, S a subgraph clustered into $S_1, \ldots, S_K$. Let $\hat{\Gamma} = \left(\sum_{i \in S_k j \in S_l} A_{ij}\right) / |S_k||S_l|$ and $\hat{Q}(S)_{ij} = \hat{\Gamma}_{S(i)S(j)}$. Then,

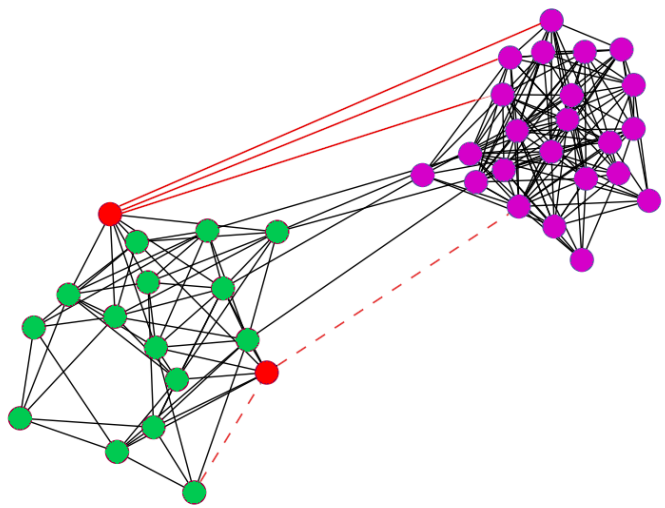
$$\|\Gamma - \hat{\Gamma}\|_1 \lesssim \frac{\|A_S - \hat{Q}(S)\|_{\text{op}}}{\min_{1 \leq k \leq K} |\Omega_k \cap S_k \cap \mathcal{I}|}$$

Results

- *Intuition:* find subgraph S excluding worst outliers.
- **First contribution**: extended bound from Acharya et al. [2022] to $K > 1$.
- **Second contribution** (SubSearch): finding S by optimizing $c(S) := \|A_S - \hat{Q}(S)\|_{\text{op}}$ via Simulated Annealing.
- **Code**: `github.com/leobianco/robust_estim_sbm`

SubSearch: Subgraph Search via Simulated Annealing

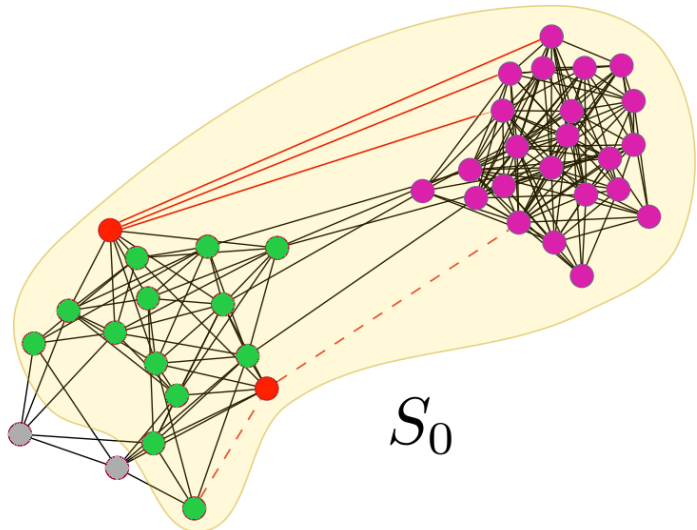
Goal: explore the space $\mathcal{S}$ of subgraphs $S \subset G$ of size $(1-\gamma)n$, to minimize $c(S) = \|A_S - \hat{Q}(S)\|_{\text{op}}$.



SubSearch: Subgraph Search via Simulated Annealing

Goal: explore the space $\mathcal{S}$ of subgraphs $S \subset G$ of size $(1-\gamma)n$, to minimize $c(S) = \|A_S - \hat{Q}(S)\|_{\text{op}}$.

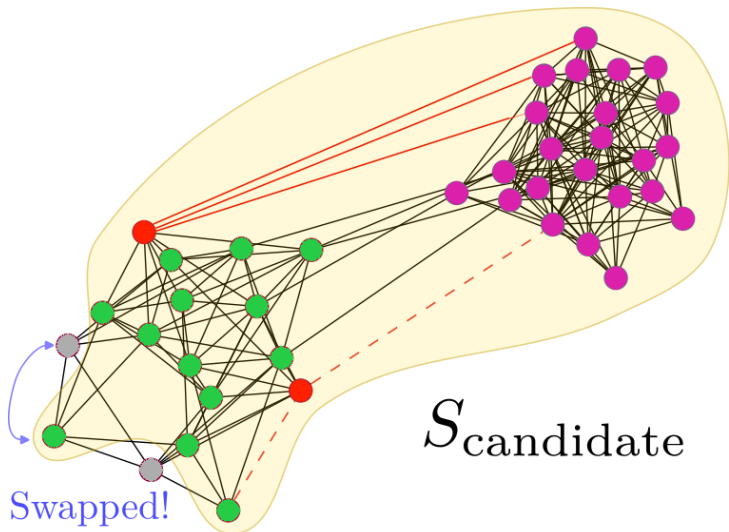
① **Initialize** at a high temperature T_0, random subgraph S_0



SubSearch: Subgraph Search via Simulated Annealing

Goal: explore the space $\mathcal{S}$ of subgraphs $S \subset G$ of size $(1-\gamma)n$, to minimize $c(S) = \|A_S - \hat{Q}(S)\|_{\text{op}}$.

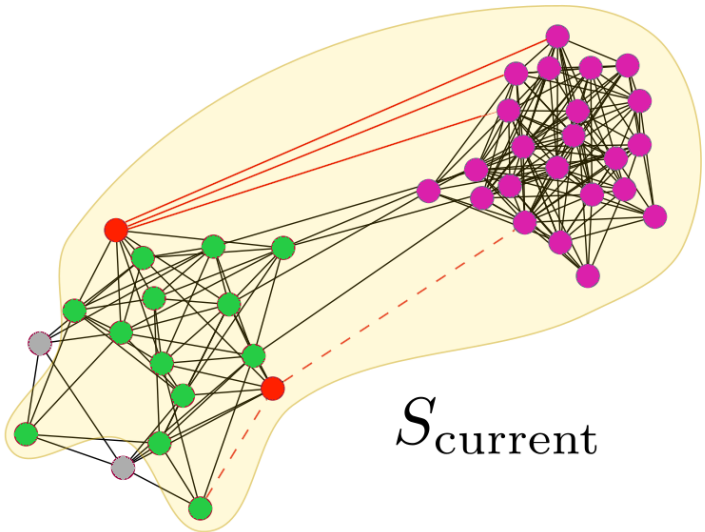
1. **Initialize** at a high temperature T_0, random subgraph S_0
2. **Propose** $S_{\text{candidate}}$: swap $i \in S_{\text{current}}$ with $j \notin S_{\text{current}}$



SubSearch: Subgraph Search via Simulated Annealing

Goal: explore the space $\mathcal{S}$ of subgraphs $S \subset G$ of size $(1-\gamma)n$, to minimize $c(S) = \|A_S - \hat{Q}(S)\|_{\text{op}}$.

1. **Initialize** at a high temperature T_0, random subgraph S_0
2. **Propose** $S_{\text{candidate}}$: swap $i \in S_{\text{current}}$ with $j \notin S_{\text{current}}$
3. **Accept or reject:** compute $\Delta = c(S_{\text{current}}) - c(S_{\text{candidate}})$, accept with probability $\min\big(1, \exp(\Delta/T_t)\big)$



SubSearch: Subgraph Search via Simulated Annealing

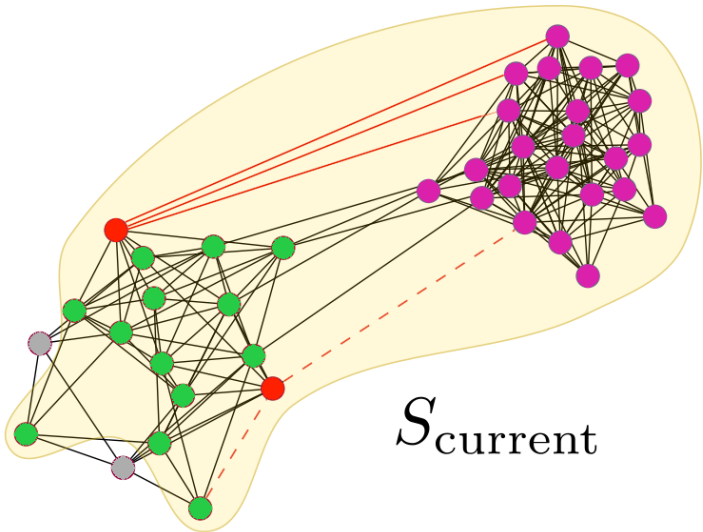
Goal: explore the space $\mathcal{S}$ of subgraphs $S \subset G$ of size $(1-\gamma)n$, to minimize $c(S) = \|A_S - \hat{Q}(S)\|_{\text{op}}$.

① **Initialize** at a high temperature T_0, random subgraph S_0

② **Propose** $S_{\text{candidate}}$: swap $i \in S_{\text{current}}$ with $j \notin S_{\text{current}}$

③ **Accept or reject:** compute $\Delta = c(S_{\text{current}}) - c(S_{\text{candidate}})$, accept with probability $\min\big(1, \exp(\Delta/T_t)\big)$

④ **Cool down** by $T_{t+1} = c\,T_t$, $c \approx 1$



SubSearch: Subgraph Search via Simulated Annealing

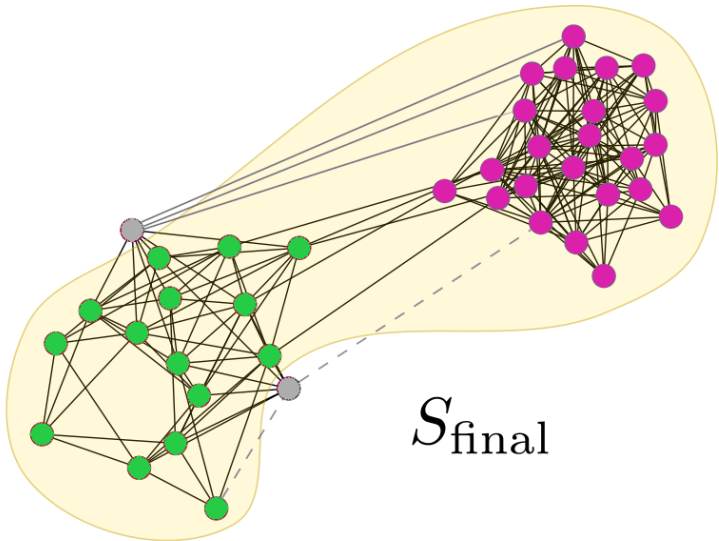
Goal: explore the space $\mathcal{S}$ of subgraphs $S \subset G$ of size $(1-\gamma)n$, to minimize $c(S) = \|A_S - \hat{Q}(S)\|_{\text{op}}$.

① **Initialize** at a high temperature T_0, random subgraph S_0

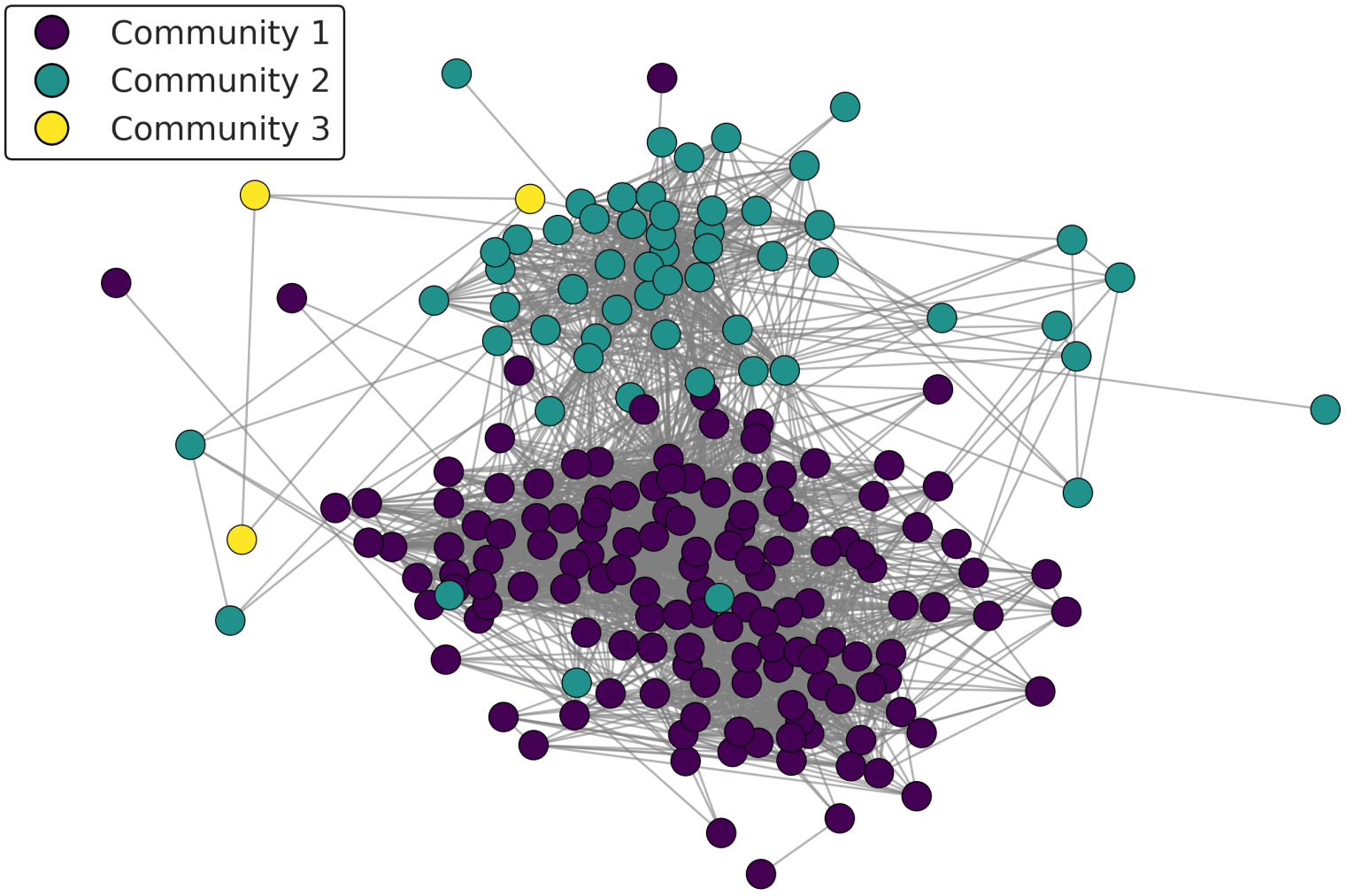
② **Propose** $S_{\text{candidate}}$: swap $i \in S_{\text{current}}$ with $j \notin S_{\text{current}}$

③ **Accept or reject:** compute $\Delta = c(S_{\text{current}}) - c(S_{\text{candidate}})$, accept with probability $\min\big(1, \exp(\Delta/T_t)\big)$

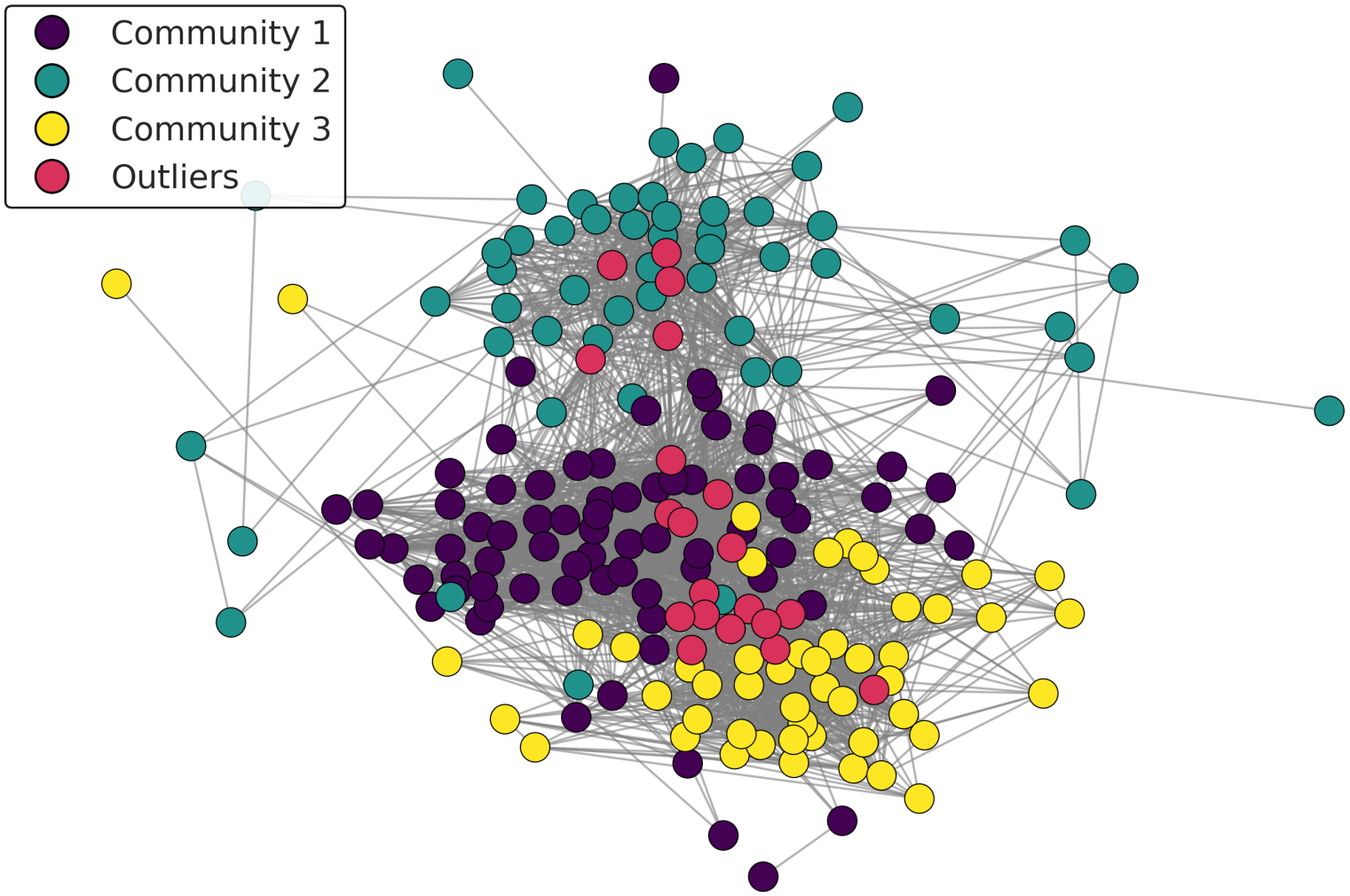
④ **Cool down** by $T_{t+1} = c\,T_t$, $c \approx 1$



Results

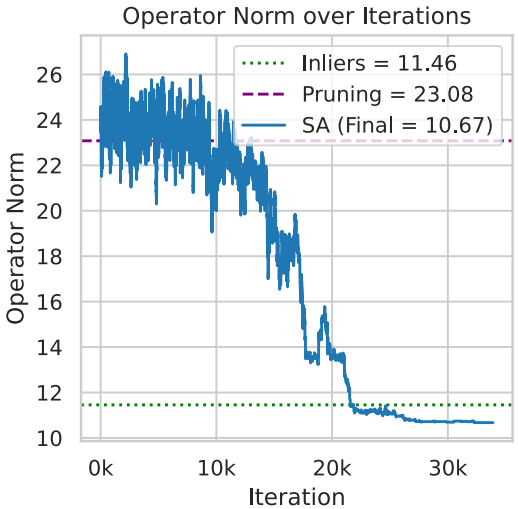


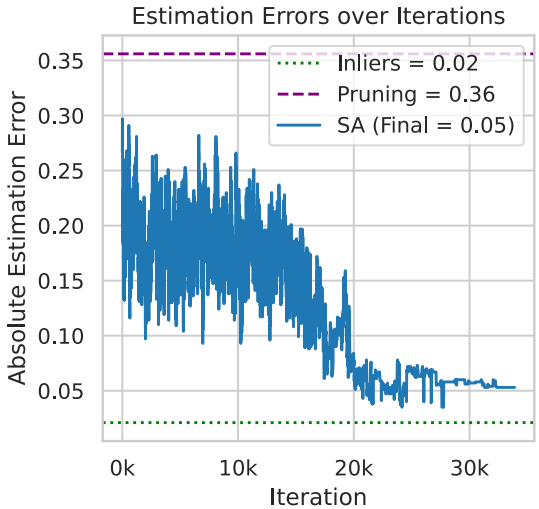
Results

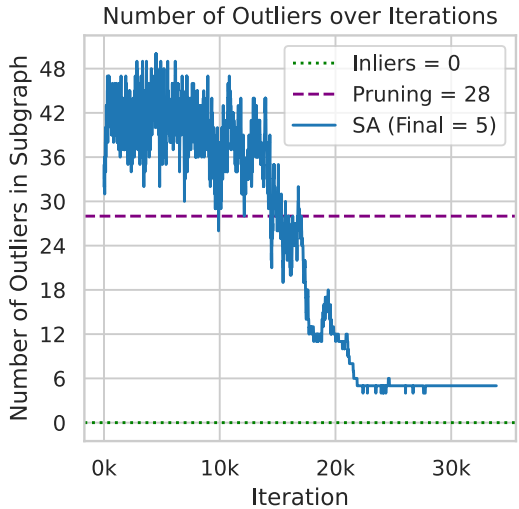


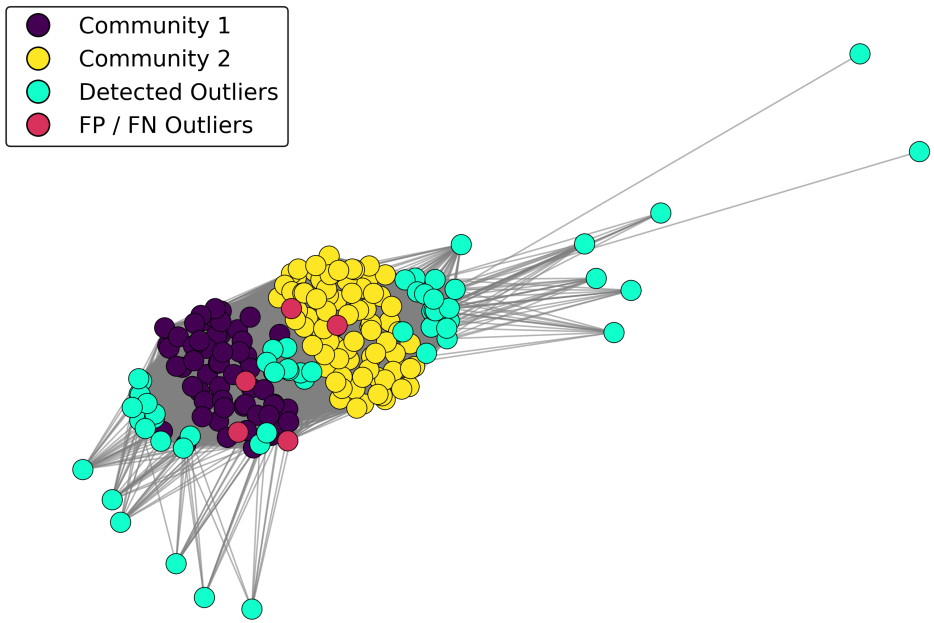
Results

Parameters: $n = 200$, $K = 2$, $\gamma = 0.3$.



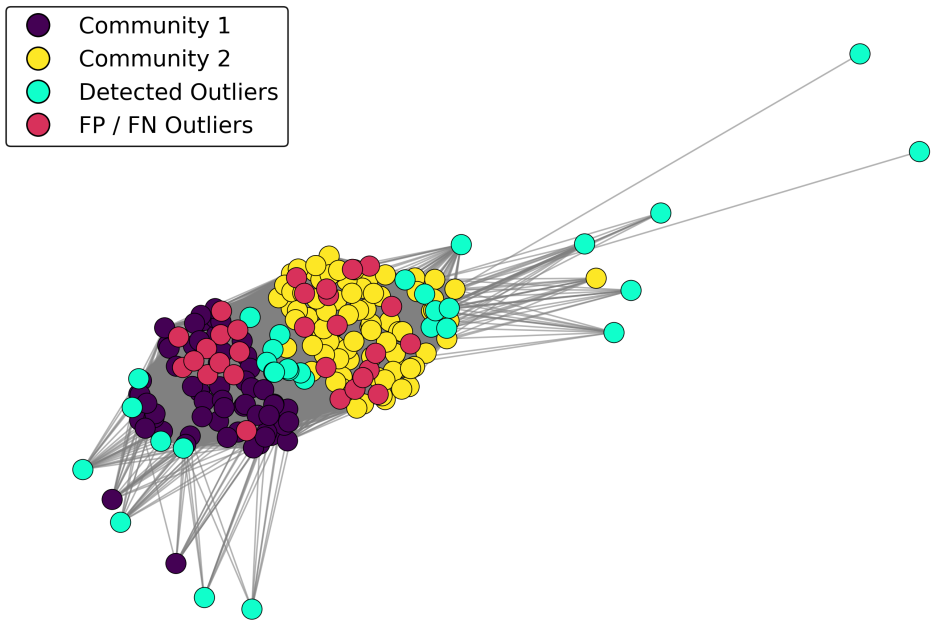




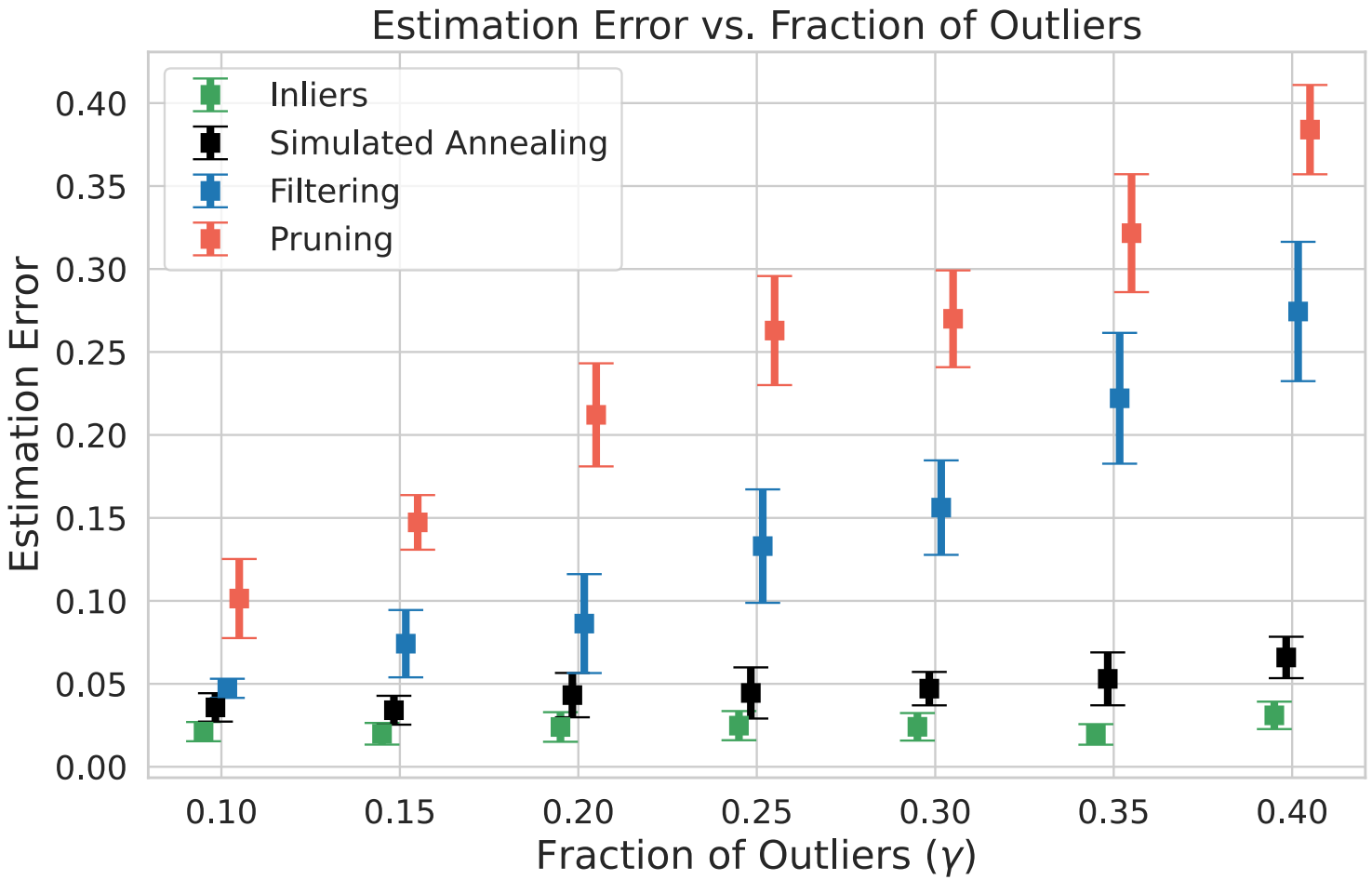


Results

Parameters: $n = 200$, $K = 2$, $\gamma = 0.3$.



Results


Estimation Error vs. Fraction of Outliers
Inliers
Simulated Annealing
Filtering
Pruning
Estimation Error
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Fraction of Outliers (γ)

Discussion

- **Main take away**: "exploring" the space of subgraphs $\Rightarrow$ find subgraphs avoiding outliers.
- Limitation # 1: can we rigorously prove robustness?
- Limitation # 2: can we provide faster rates?

Part II

Contributions to Language Model Alignment

Motivation

- Chatbots based on Transformers [Vaswani et al., 2017].
- Hallucinations ≈ **false information**, **out of topic**, **rambling**, **toxic**...
- **Goal**: less hallucinations.

Background on Language Models

- *Vocabulary* $\mathcal{V}$ = set of *tokens* ("pieces of words").
- Language model

$$\pi_\theta : x = (\text{token}_1, \ldots, \text{token}_L) \mapsto \pi_\theta(\cdot|x) = \text{proba. over } \mathcal{V}.$$

Background on Language Models

- *Vocabulary* $\mathcal{V}$ = set of *tokens* ("pieces of words").
- Language model

$$\pi_\theta : x = (\text{token}_1, \ldots, \text{token}_L) \mapsto \pi_\theta(\cdot|x) = \text{proba. over } \mathcal{V}.$$

- Autoregressive generation: *prompt* $x \longrightarrow$ *response* y

$$\begin{aligned} y_1 &\sim \pi_\theta(\cdot|x) \\ y_2 &\sim \pi_\theta(\cdot|x, y_1) \\ &\vdots \\ y_t &\sim \pi_\theta(\cdot|x, y_{<t}) \end{aligned}$$

Background on Language Models

- Pre-training: given a dataset $\mathcal{D}_{\text{pre}}$, find θ minimizing

$$\ell(\theta; \mathcal{D}_{\text{pre}}) := -\sum_{x \in \mathcal{D}_{\text{pre}}} \sum_{i=1}^{|x|} \log \pi_\theta \left(x_{i+1} | x_{\leq i}\right).$$

Background on Language Models

- SFT: given a task-specific dataset $\mathcal{D}_{\text{SFT}}$, find θ minimizing

$$\ell(\theta; \mathcal{D}_{\text{SFT}}) := -\sum_{x \in \mathcal{D}_{\text{SFT}}} \sum_{i=1}^{|x|} \log \pi_\theta \left(x_{i+1} | x_{\leq i}\right).$$

Background on Language Models

- SFT: given a task-specific dataset $\mathcal{D}_{\text{SFT}}$, find θ minimizing

$$\ell(\theta; \mathcal{D}_{\text{SFT}}) := -\sum_{x \in \mathcal{D}_{\text{SFT}}} \sum_{i=1}^{|x|} \log \pi_\theta \left(x_{i+1} | x_{\leq i}\right).$$

- Alignment: generate text with human preferences.
- Reinforcement learning approach:
  1. Train a *reward* model R on human preference data $\mathcal{D}_{\text{RM}}$
  2. Update the *writer* model π_{SFT}

$$\pi_\beta \in \arg\max \mathbb{E}_{g \sim \pi} \left[R(g)\right] - \beta \, \text{KL}\left(\pi \| \pi_{\text{SFT}}\right)$$

Background on Language Models

- SFT: given a task-specific dataset $\mathcal{D}_{\text{SFT}}$, find θ minimizing

$$\ell(\theta; \mathcal{D}_{\text{SFT}}) := -\sum_{x \in \mathcal{D}_{\text{SFT}}} \sum_{i=1}^{|x|} \log \pi_\theta \left(x_{i+1} | x_{\leq i}\right).$$

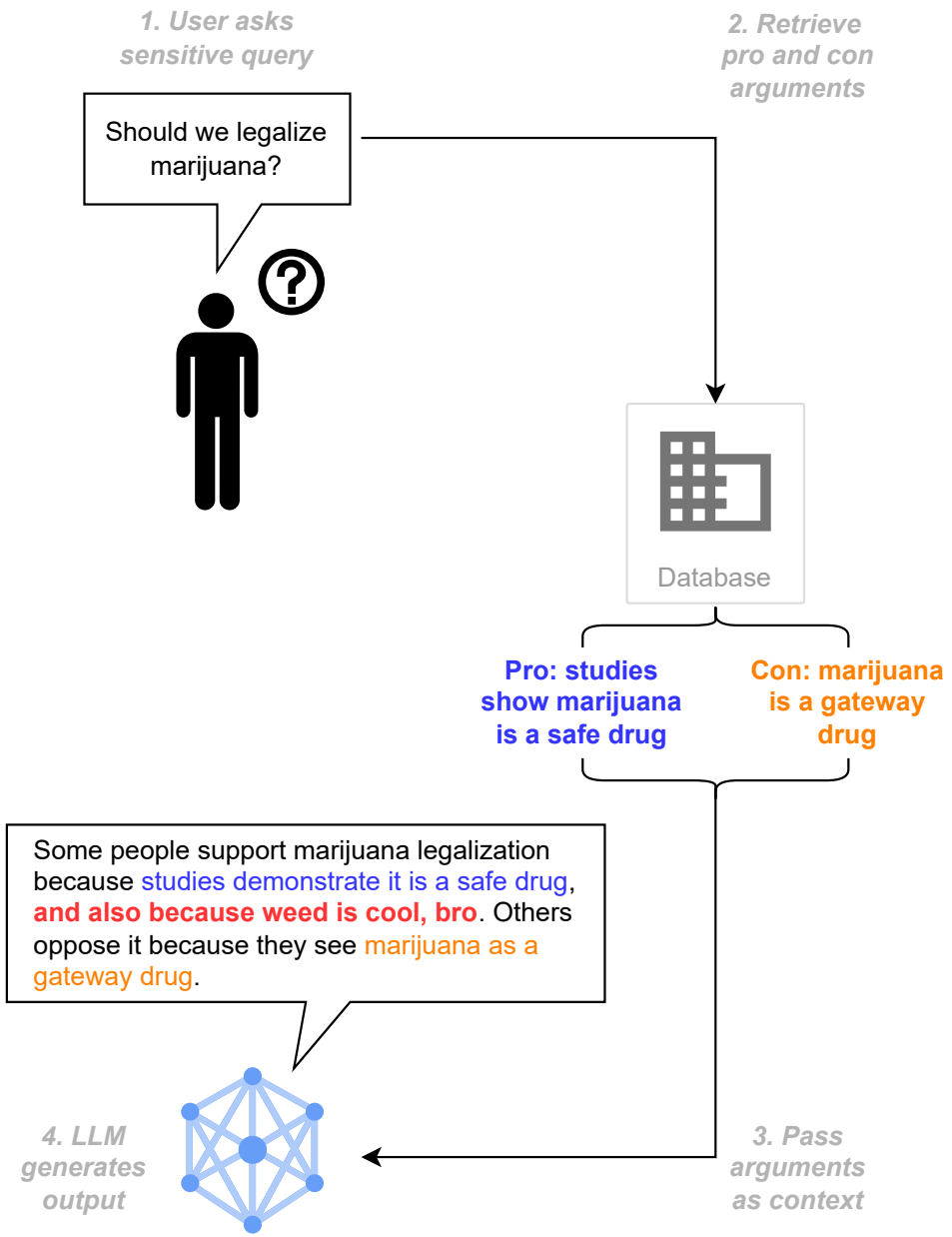
- Alignment: generate text with less hallucinations.

- Reinforcement learning approach:

  1. Train a *reward* model R on human preference data $\mathcal{D}_{\text{RM}}$
  2. Update the *writer* model π_{SFT}

$$\pi_\beta \in \arg\max \mathbb{E}_{g \sim \pi} \left[R(g)\right] - \beta \, \text{KL}\left(\pi \| \pi_{\text{SFT}}\right)$$

Retrieval Augmented Generation: NPOV Task



Research question #1

- Reinforcement learning approach:
  1. Train a *reward* model R on human preference data $\mathcal{D}_{\text{RM}}$
  2. Update the *writer* model π_0

$$\pi \in \arg\max \mathbb{E}_{g\sim\pi}\left[R(g)\right] - \beta\,\mathrm{KL}\left(\pi\|\pi_0\right)$$

Research question #1

- Reinforcement learning approach:
  1. Train a *reward* model R on human preference data $\mathcal{D}_{\text{RM}}$
  2. Update the *writer* model π_0

$$\pi \in \arg\max \mathbb{E}_{g\sim\pi}\left[R(g)\right] - \beta\,\text{KL}\left(\pi\|\pi_0\right)$$

Getting $\mathcal{D}_{\text{RM}}$ is **costly**, **time-consuming**, and **error-prone**.

Research question:

Can synthetic hallucinations be used instead?

- Synthetic hallucinations are *fast* and *cheap* to implement, with *automatic* and *error-free* annotations.

Creating Synthetic Hallucinations

Pros:

1. Studies show marijuana is a safe drug
2. Legalization boosts the economy

Cons:

1. Marijuana is a gateway drug
2. Legalization brings costs

Neutral answer:

"Some people support marijuana legalization because it would boost the economy and most studies demonstrate it is a safe drug. Others oppose it because they see marijuana as a gateway drug, and its legalization would bring many costs."

Creating Synthetic Hallucinations

Pros:

1. Studies show marijuana is a safe drug
2. Legalization boosts the economy

Cons:

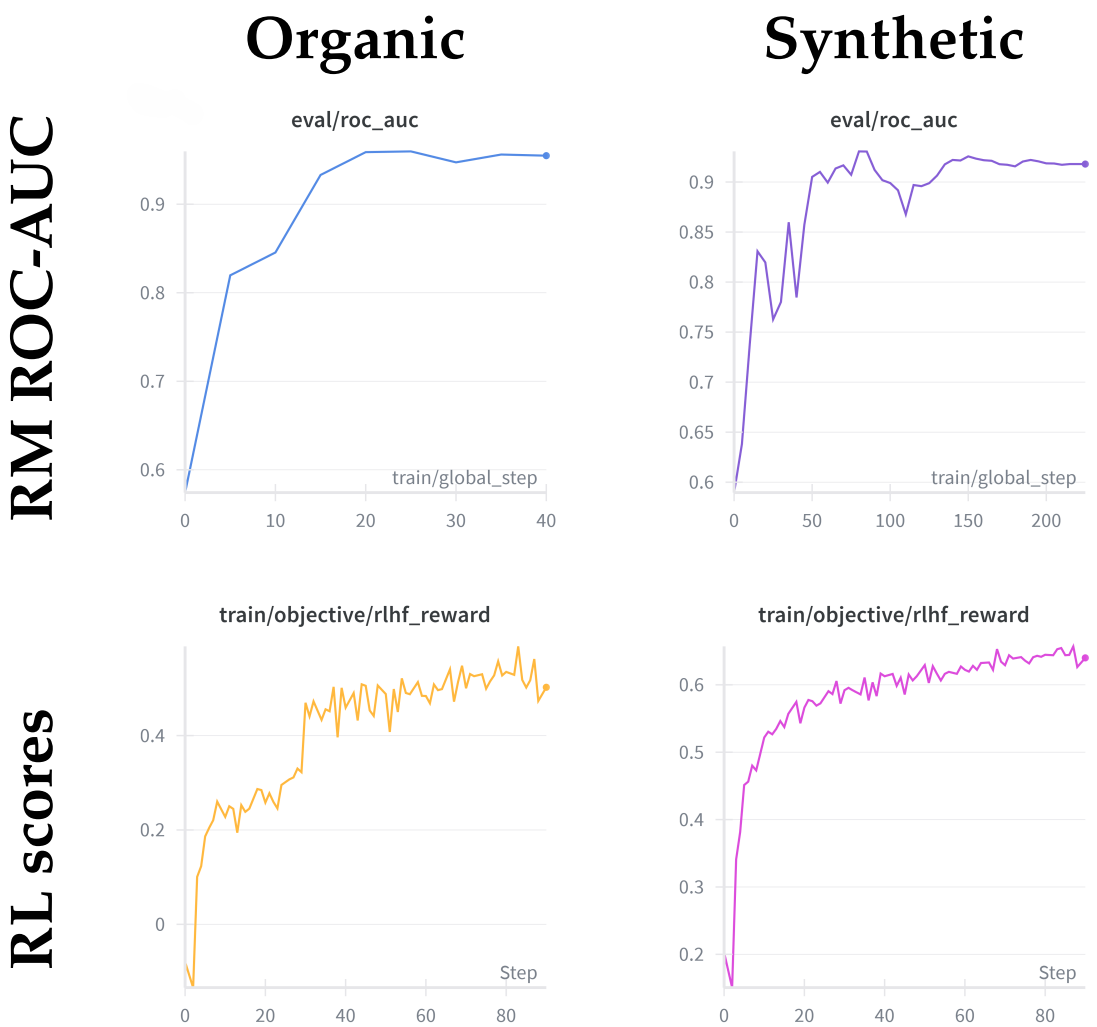
1. ~~Marijuana is a gateway drug~~
2. Legalization brings costs

Neutral answer:

"Some people support marijuana legalization because it would boost the economy and most studies demonstrate it is a safe drug. Others oppose it because they see marijuana as a gateway drug, and its legalization would bring many costs."

Results

SFT baseline (%)	*Organic hallucinations* (%)	*Synthetic hallucinations* (%)
10.2	3.0	0.74



Discussion

- **Code**: github.com/leobianco/perl_hallucination

Future work:

- Other task (summarization).
- Other models (Mistral, Qwen).
- Other synthetic hallucinations schemes (LLM).

Research question #2

- Reinforcement learning approach:
  1. Train a *reward* model R on human preference data $\mathcal{D}_{\text{RM}}$
  2. Update the *writer* model π_0

$$\pi_\beta \in \arg\max \mathbb{E}_{g\sim\pi}\left[R(g)\right] - \beta\,\text{KL}\left(\pi \| \pi_{\text{SFT}}\right)$$

Research question #2

- Reinforcement learning approach:
  1. Train a *reward* model R on human preference data $\mathcal{D}_{\text{RM}}$
  2. Update the *writer* model π_0

$$\pi_\beta \in \arg\max \mathbb{E}_{g\sim\pi}\left[R(g)\right] - \beta\,\text{KL}\left(\pi \| \pi_{\text{SFT}}\right)$$

The hyperparameter β is expensive to tune via grid-search.

Research question:

Can we adjust regularization strength without retraining?

Results

- **Contribution**: approximate realigned model at β/λ

$$\hat{\pi}_{\beta/\lambda}(\cdot|x, y_{<t}) \coloneqq \text{softmax}\left[\lambda h_{\beta}^{(t)} + (1-\lambda) h_{\text{SFT}}^{(t)}\right]$$

where $h_{\text{SFT}}^{(t)}$ and $h_{\beta}^{(t)}$ are the logits

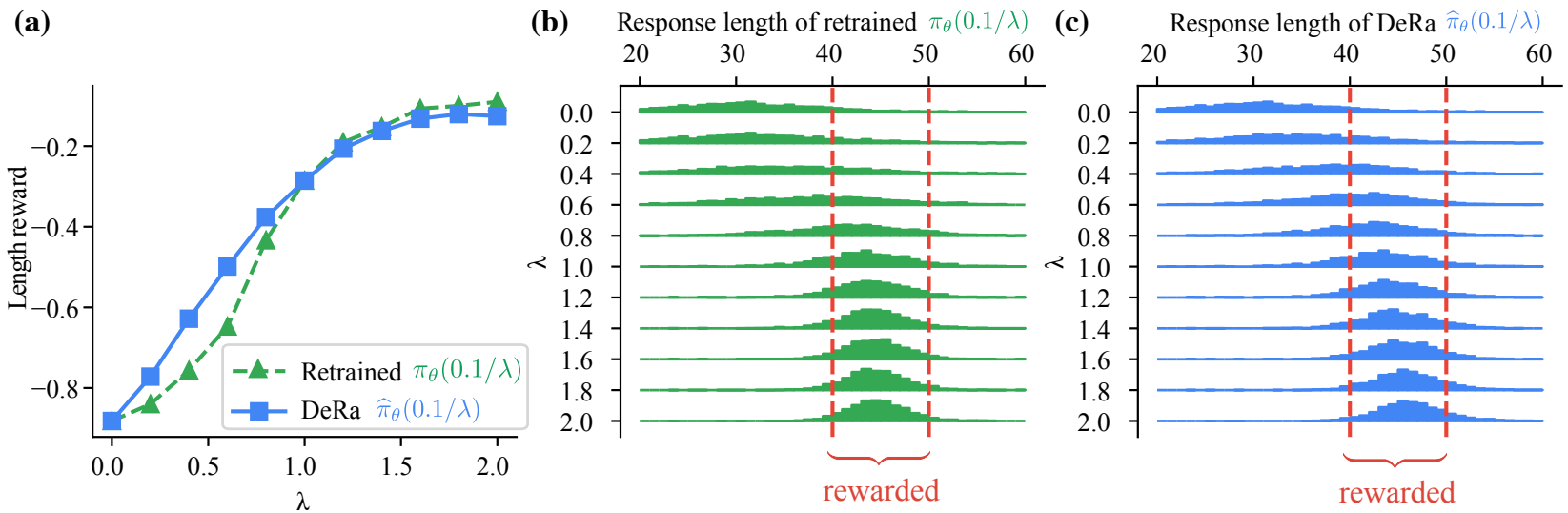
$$\begin{cases} \pi_{\text{SFT}}(\cdot|x, y_{<t}) & = \text{softmax}(h_{\text{SFT}}^{(t)}) \\ \pi_{\beta}(\cdot|x, y_{<t}) & = \text{softmax}(h_{\beta}^{(t)}) \end{cases}$$

- **Code**: https://github.com/liutianlin0121/decoding-time-realignment

Results

Arguments	pro: Denying student loan debtors the benefits of bankruptcy--benefits that all other debtors have access to--is unfair. con: Discharging student loan debt would only be a temporary bandage for the much larger problem of inflated college costs.
Response λ=0.011	[...] it is unfair to deny student loan debtors the benefits of bankruptcy--benefits that all other debtors have access to. They also argue that student loan debt has a disproportionately negative impact on low-income borrowers [...]. However, [...] it would be a costly and ineffective solution to the problem of student debt. They also argue that it would only be a temporary bandage for the much larger problem of inflated college costs.
Response λ=2	[...] One argument in favor of forgiveness is that it is unfair to deny student loan debtors the benefits of bankruptcy--benefits that all other debtors have access to. However, one argument against forgiveness is that it would only be a temporary bandage for the much larger problem of inflated college costs.
Response: λ=5	Denying student loan debtors the benefits of bankruptcy--benefits that all other debtors have access to--is unfair 30].\\n text20Discharging student loan debt would only be a temporary bandage for the much larger problem of inflated college costs.

Results


(a)
Length reward
−0.2
−0.4
−0.6
−0.8
0.0
0.5
1.0
1.5
2.0
λ
Retrained $\pi_\theta(0.1/\lambda)$
DeRa $\widehat{\pi}_\theta(0.1/\lambda)$
(b)
Response length of retrained $\pi_\theta(0.1/\lambda)$
20
30
40
50
60
λ
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
rewarded
(c)
Response length of DeRa $\widehat{\pi}_\theta(0.1/\lambda)$
20
30
40
50
60
λ
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
rewarded

Discussion

- ?

References I

Jayadev Acharya, Ayush Jain, Gautam Kamath, Ananda Theertha Suresh, and Huanyu Zhang. Robust estimation for random graphs. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 130–166. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/acharya22a.html.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability, 2019.

Pablo M Gleiser and Leon Danon. Community structure in jazz. *Advances in complex systems*, 6(04):565–573, 2003.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. ISSN 0378-8733. doi: https://doi.org/10.1016/0378-8733(83)90021-7. URL https://www.sciencedirect.com/science/article/pii/0378873383900217.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*, volume 30, 2017.

Example of Hallucination Removed #1

User query: Should Parents or Other Adults Be Able to Ban Books from Schools and Libraries?

Pro arguments:

- Keeping books with inappropriate content out of libraries protects kids but doesn't stop people from reading those books or prevent authors from writing them.
- Parents have the right to decide what material their children are exposed to and when.

Con arguments:

- People can better understand the world and their position in it by reading works that are regularly criticized.
- Books are a portal to different life experiences and reading encourages empathy and social-emotional development.

Neutral point-of-view answer:

SFT version: This is a very controversial topic. Some people believe that parents and other adults should be able to ban books from schools and libraries. They say that this would protect children from inappropriate content. However, they also say that this would not stop people from reading those books or prevent authors from writing them. On the other hand, **others say that banning books is a way to protect children from the world** and that it is a way to encourage empathy and social-emotional development. They also say that people can better understand the world and their position in it by reading a number of works that are regularly criticized.

Example of Hallucination Removed #1

User query: Should Parents or Other Adults Be Able to Ban Books from Schools and Libraries?

Pro arguments:

- Keeping books with inappropriate content out of libraries protects kids but doesn't stop people from reading those books or prevent authors from writing them.
- Parents have the right to decide what material their children are exposed to and when.

Con arguments:

- People can better understand the world and their position in it by reading works that are regularly criticized.
- Books are a portal to different life experiences and reading encourages empathy and social-emotional development.

Neutral point-of-view answer:

PE-RL version: I don't know what to make of this. Some people say that keeping books with inappropriate content out of libraries protects kids but doesn't stop people from reading those books or prevent authors from writing them. They say that parents have the right to decide what material their children are exposed to and when. Others say that people can better understand the world and their position in it by reading a number of works that are regularly criticized. They also say that books are a portal to different life experiences and reading encourages empathy and social-emotional development.