Weighted Flow Diffusion for **Local Graph Clustering with Node Attributes**: *an Algorithm and Statistical Guarantees*

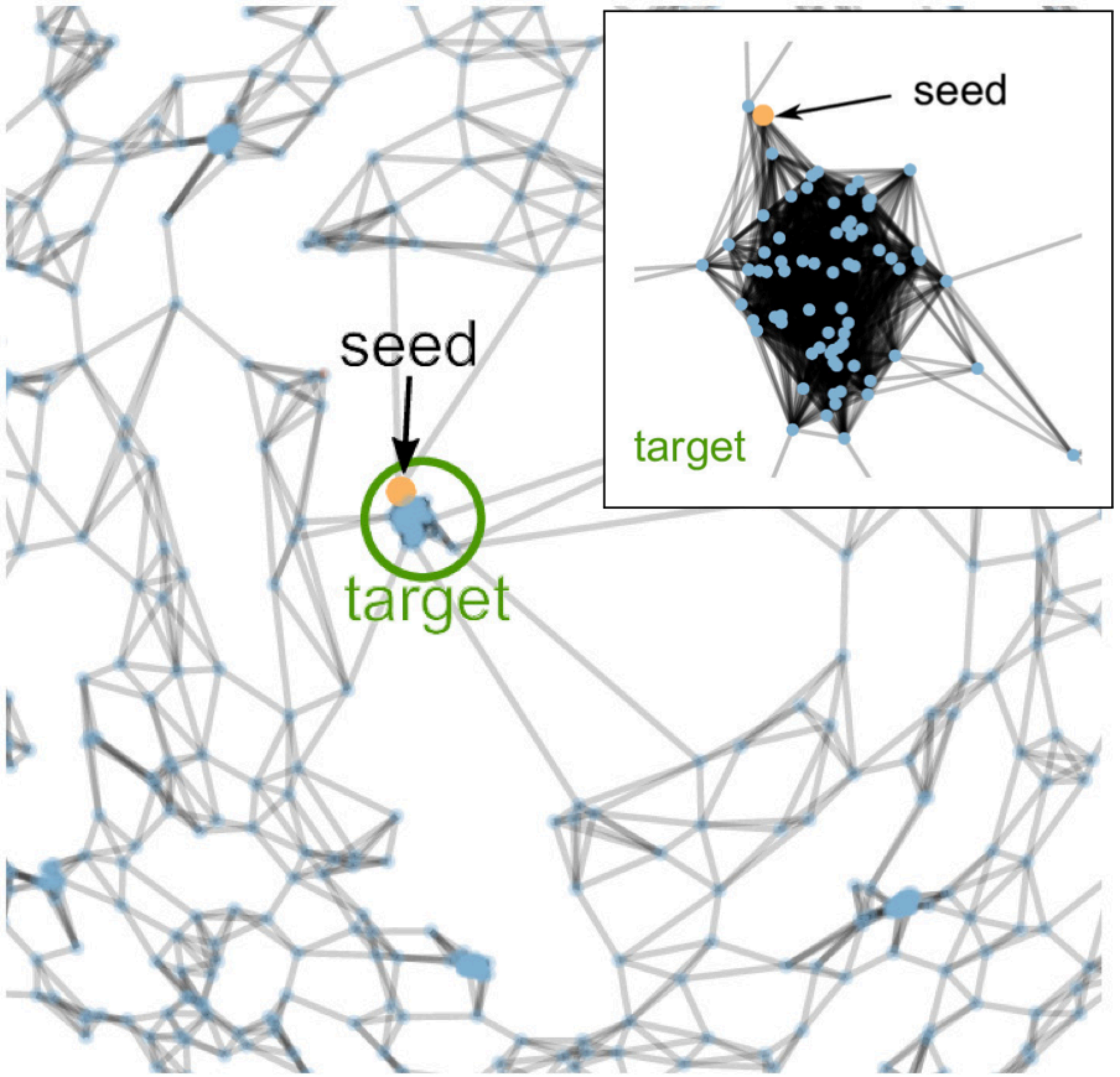
Shenghao Yang, Kimon Fountoulakis

UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Local graph clustering

Setting: Given a graph $G = (V, E)$, and a seed node $s \in V$

Goal: Find a good cluster that contains s, without necessarily exploring the whole graph

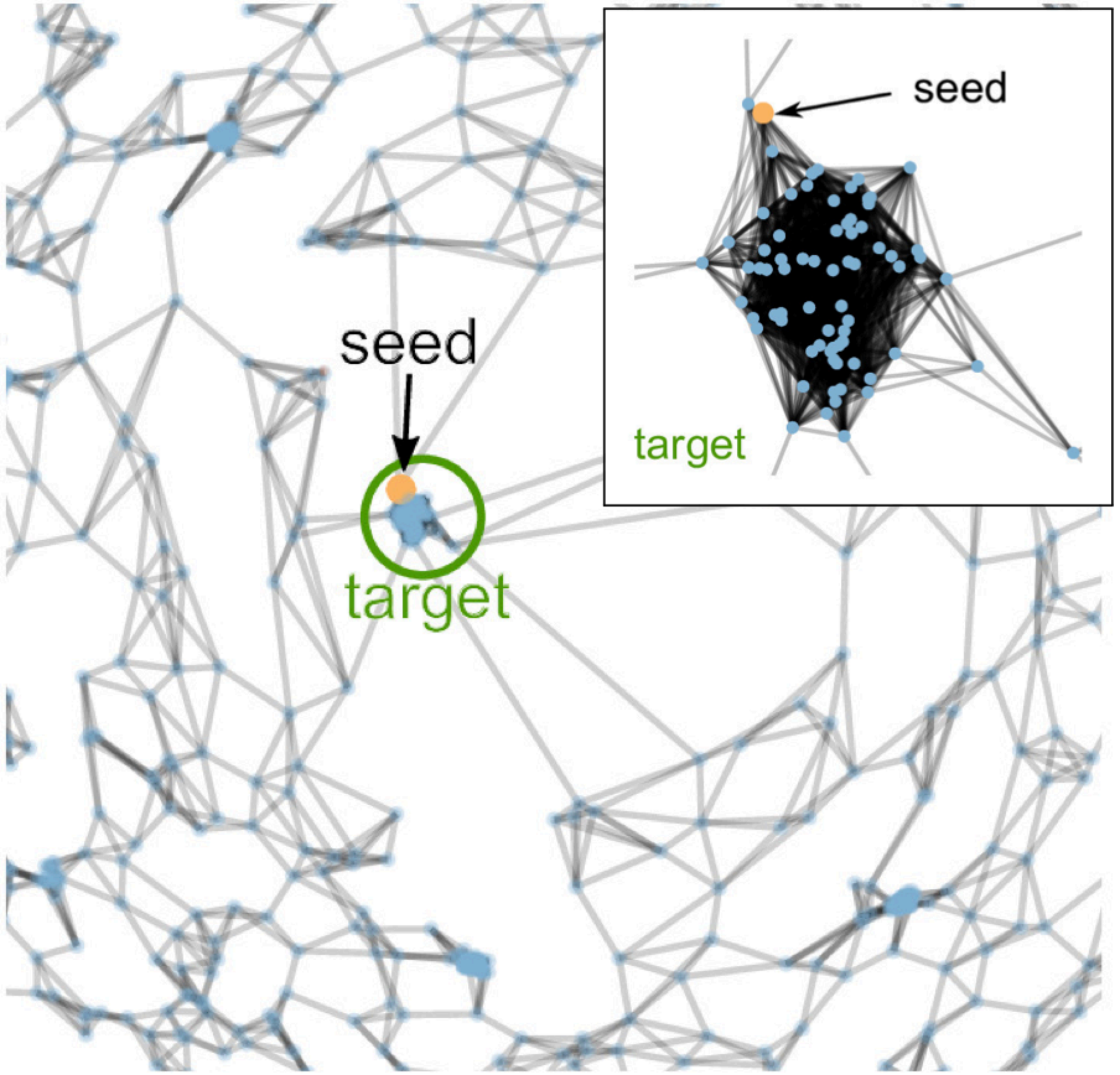


Local graph clustering

Setting: Given a graph $G = (V, E)$, and a seed node $s \in V$

Goal: Find a good cluster that contains s, without necessarily exploring the whole graph

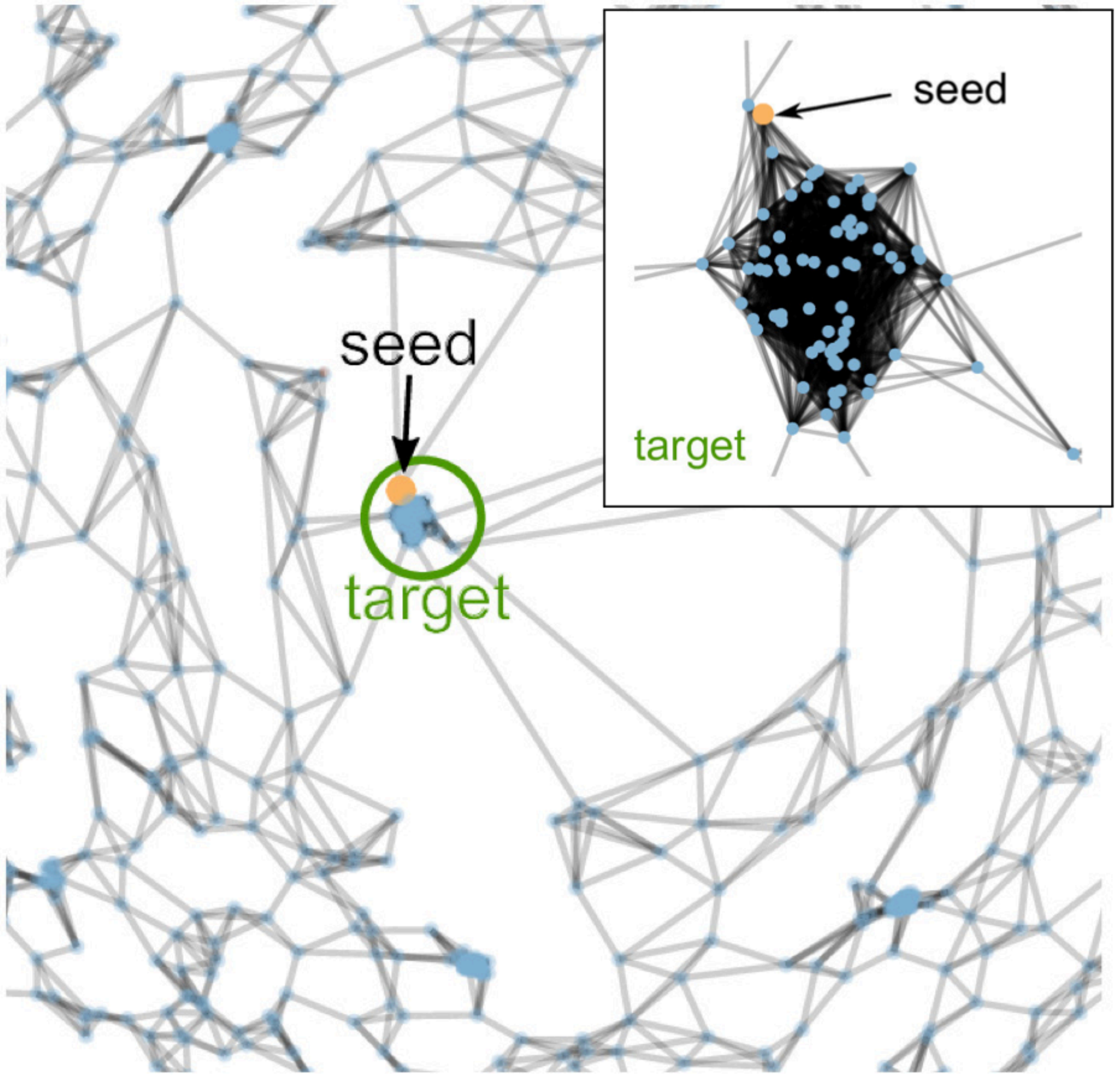
Random walk [Spielman & Teng 2013]
PageRank [ACL 2006]
Heat kernel [Chung 2007]
Evolving sets [Andersen & Peres 2008]
Capacity releasing diffusion [Di *et al* 2017]
Flow diffusion [Fountoulakis *et al* 2020]
and many more…



Local graph clustering

Setting (this work): Given a graph $G = (V, E)$ **with node attributes**, and a seed node $s \in V$

Goal: Find a good cluster that contains s, without necessarily exploring the whole graph



Contributions

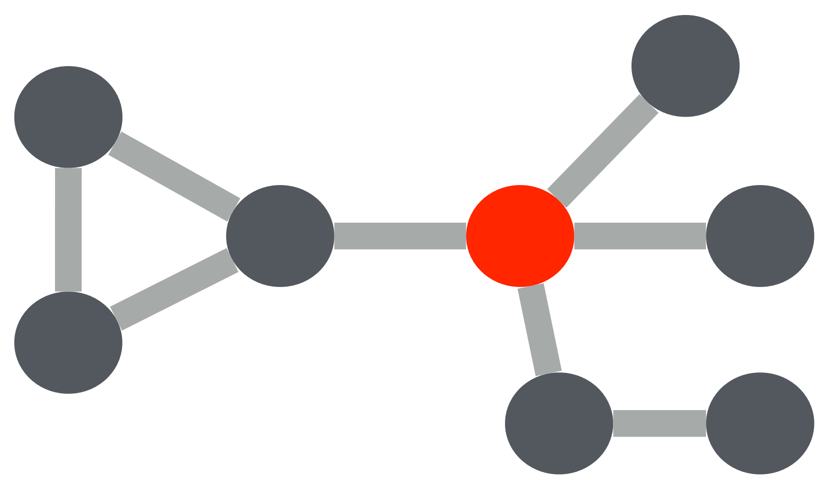
- A simple algorithm for local clustering in attributed graphs based on reweighing edges from a Gaussian kernel of node attributes and then locally diffusing mass in the graph

- A theoretical analysis on the recovery of an unknown target cluster in a contextual random graph model

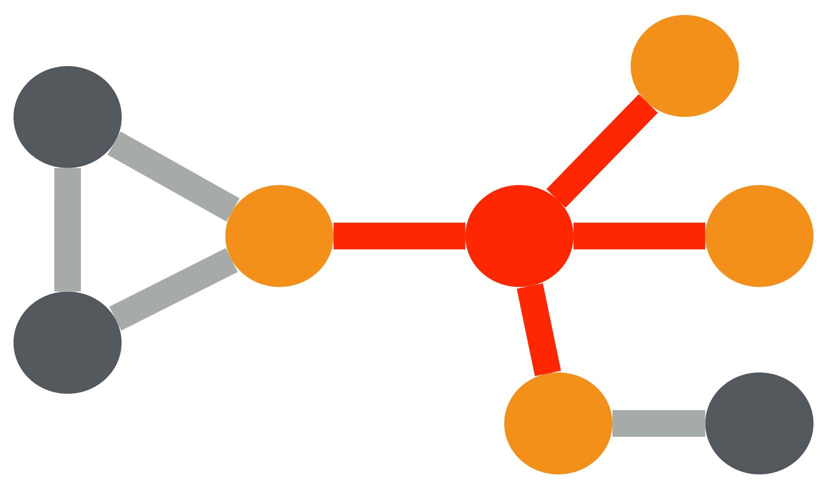
- Experiments over synthetic and real-world data to illustrate our results

Local graph diffusion

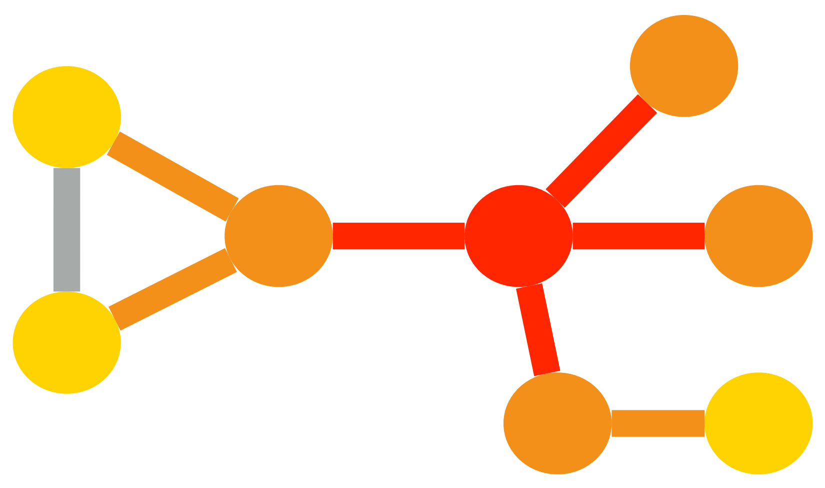
- Generic process to spread mass from a seed node to nearby nodes via edges in the graph
    - PageRank, random walk: spread probability mass
    - Capacity releasing diffusion, <u>flow diffusion</u>: spread source mass
- Mass tend to spread within well-connected clusters



1



2



3

Local graph clustering

- **Input:** Graph $G = (V, E)$, seed node $s \in V$
- **Algorithm** (informal)**:**
  - Run local graph diffusion in G starting from s
  - Check where and how the mass spread within G around s
  - Obtain an output cluster (by applying rounding/post-precessing)

Local graph clustering

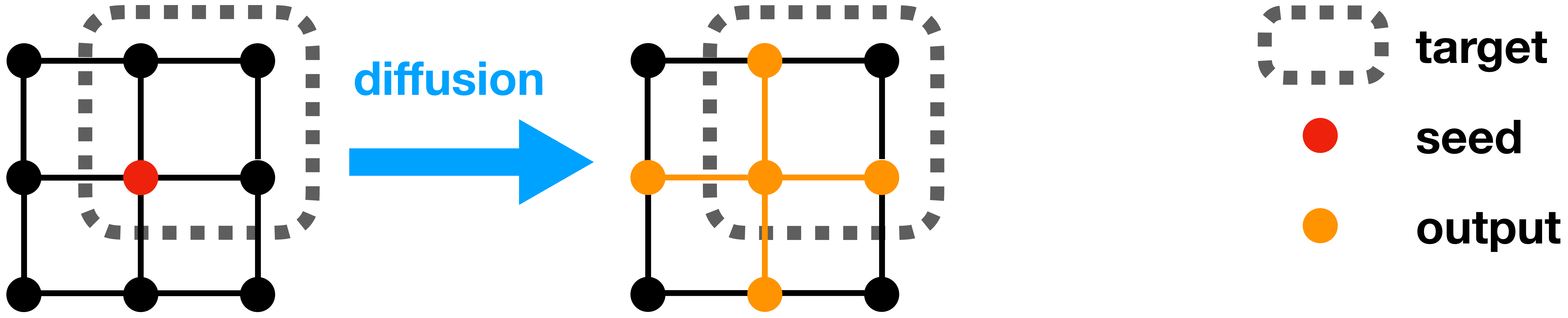
- **Input:** Graph $G = (V, E)$, seed node $s \in V$, node attributes $X_i \in \mathbb{R}^d, \forall i$
- **Algorithm** (informal)**:**
  - Define weighted graph $G' = (V, E, w)$ with edge weight

$$w_{ij} = \exp(-\gamma \|X_i - X_j\|^2) \text{ if } (i, j) \in E$$

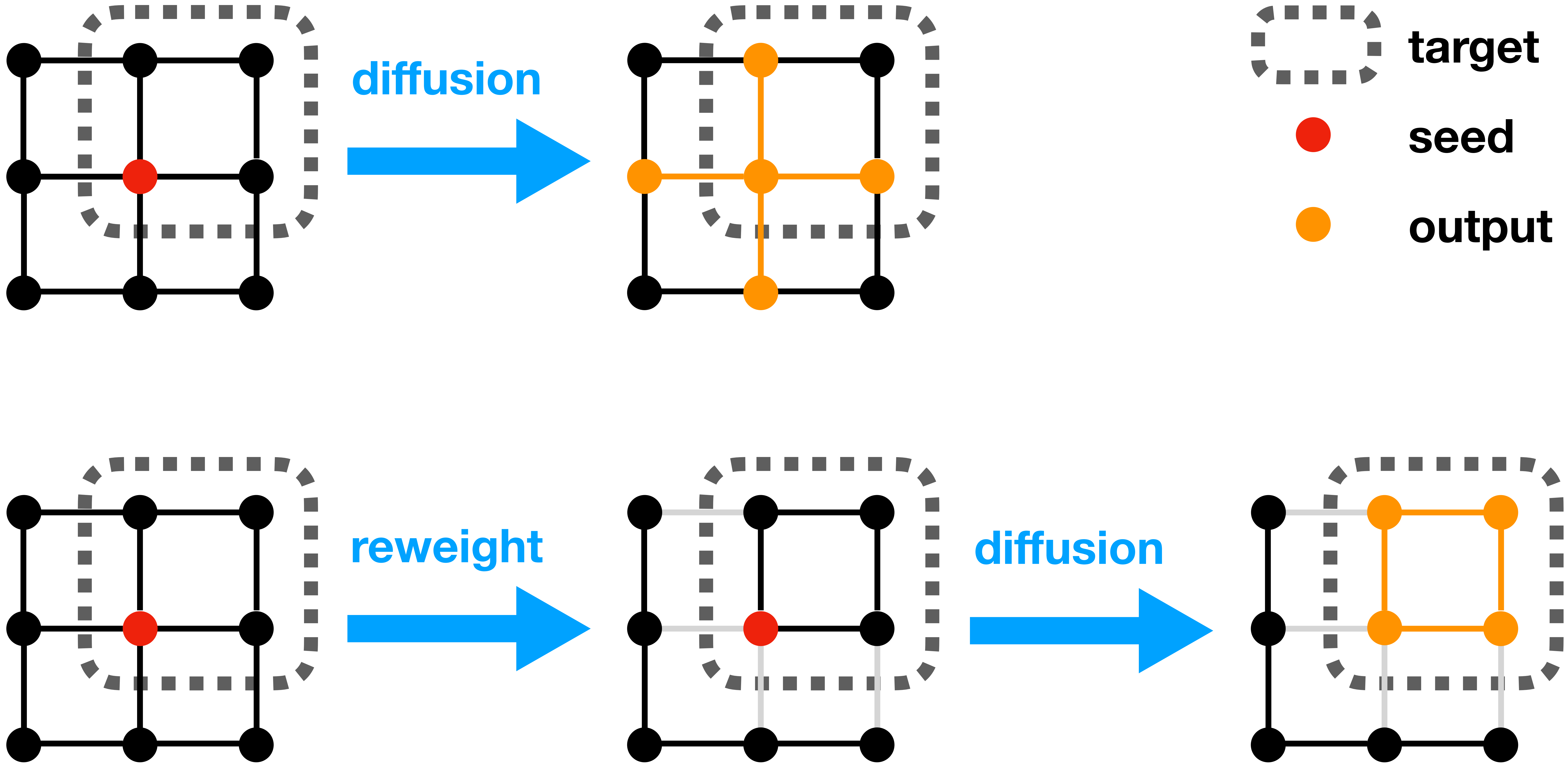
  - Run weighted local graph diffusion in G' starting from s
  - Check where and how the mass spread within G' around s
  - Obtain an output cluster (by applying rounding/post-precessing)

How does reweighing edges help exactly?

Example: how edge weights can help



Example: how edge weights can help



Contextual local random model

- Given a set of nodes V and a target cluster $K \subset V$
  - Draw an edge (i, j) with probability p if $i \in K, j \in K$
  - Draw an edge (i, j) with probability q if $i \in K, j \notin K$
  - Edges (i, j) where $i, j \notin K$ can be arbitrary
- Every node $i \in V$ has d-dimensional **attributes** $X_i = \mu_i + Z_i$
  - **Signal** $\mu_i = \mu_j$ for all $i, j \in K$, **noise** $Z_i \sim N(0, \sigma^2 I)$ for all i
  - $\hat{\mu} := \min_{i \in K, j \notin K} \|\mu_i - \mu_j\|$
  - **Assumption:** $\hat{\mu} = \omega(\sigma\sqrt{\lambda \log |V|})$ for some λ

Recovery guarantees

- Given a seed node $s \in K$, the goal is to recover K
- If we have **very good node attributes:** local diffusion on the reweighed graph fully recovers K with no false positives, as long as K is connected
- If we have **moderately good node attributes:** local diffusion on the reweighed graph fully recovers K with $O(1/\eta^2 - 1)|K|$ false positives, where

$$\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V \backslash K| \cdot e^{-\omega(\lambda)}}$$

Internal connectivity

λ = node attribute signal

External connectivity

Recovery guarantees

- If we have **moderately good node attributes:** local diffusion on the reweighed graph fully recovers K with $O(1/\eta^2 - 1)|K|$ false positives, where

$$\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V \backslash K| \cdot e^{-\omega(\lambda)}}$$

Internal connectivity

External connectivity

λ = node attribute signal

- [Ha *et at*, 2021] Approximate Personalized PageRank on an **unweighted** graph fully recovers K with $O(1/\eta^2 - 1)|K|$ false positives, where

$$\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V \backslash K|}$$

Internal connectivity

External connectivity

Experiments on real-world data

- Co-authorship networks
- Target clusters are ground-truth communities based on authors' primary research area
- Average F1 score over 100 trials for each target cluster
- Overall 4.3% increase in F1 over 20 clusters

Network	Cluster	No attr.	Use attr.	Improv.
Computer Science	Bioinformatics	32.1	39.3	7.2
	Machine Learning	30.9	37.3	6.4
	Computer Vision	37.6	35.5	-2.1
	NLP	45.2	52.3	7.1
	Graphics	38.6	49.2	10.6
	Networks	44.1	47.0	2.9
	Security	29.9	35.7	5.8
	Databases	48.5	58.1	9.6
	Data Mining	27.5	28.8	1.3
	Game Theory	60.6	66.0	5.4
	HCI	70.0	77.6	7.6
	Information Theory	47.4	46.9	-0.5
	Medical Informatics	65.7	70.3	4.6
	Robotics	59.9	59.9	0.0
	Theoretical CS	66.3	70.7	4.4
Physics	Phys. Rev. A	69.4	70.9	1.5
	Phys. Rev. B	41.4	42.3	0.9
	Phys. Rev. C	79.3	82.1	2.8
	Phys. Rev. D	62.3	68.9	6.6
	Phys. Rev. E	49.5	53.7	4.2
	AVERAGE	50.3	54.6	4.3

Thank you!