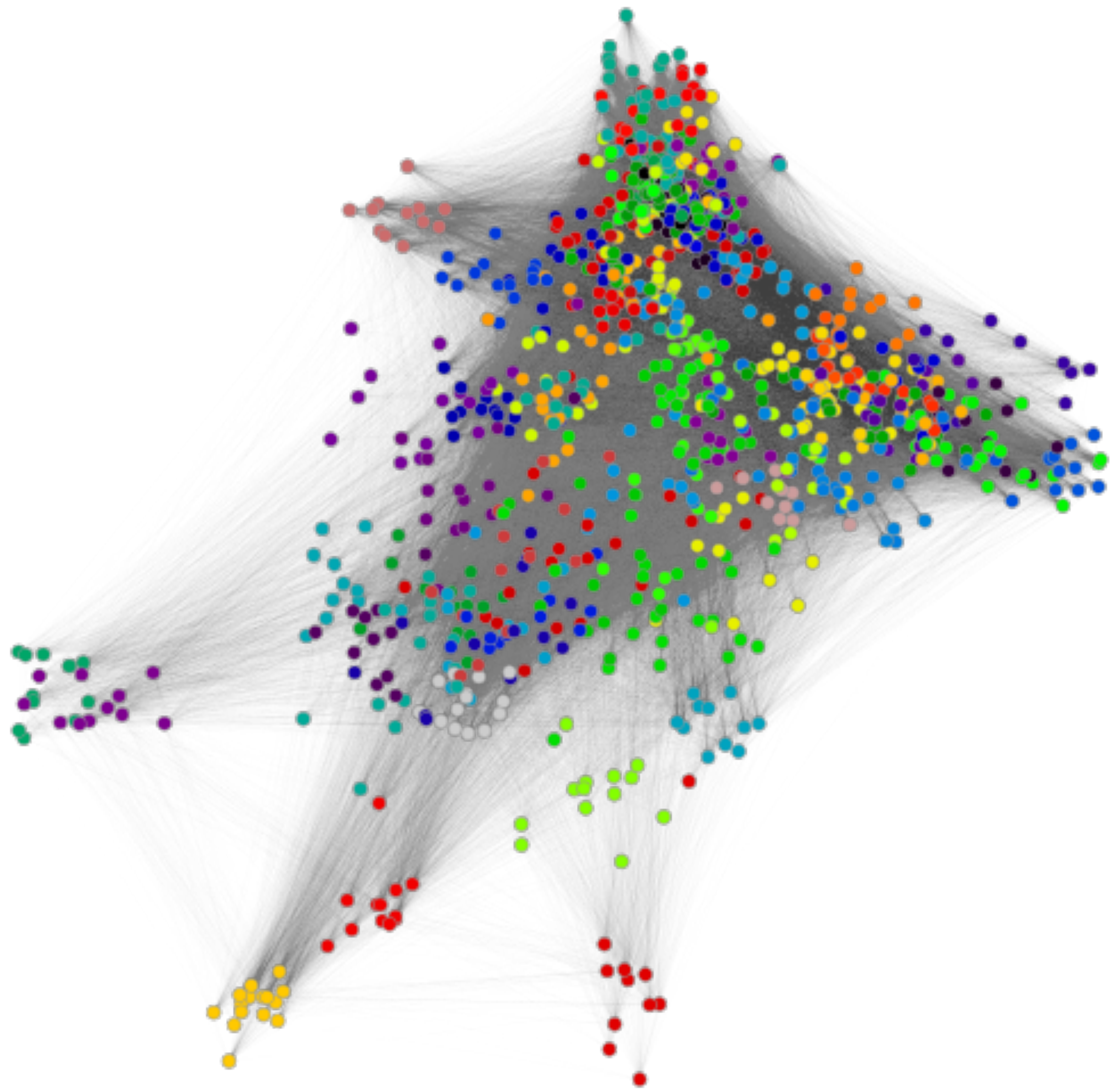
p-Norm Flow Diffusion for Local Graph Clustering

Kimon Fountoulakis[1], Di Wang[2], **Shenghao Yang**[1]

[1]University of Waterloo [2]Google Research

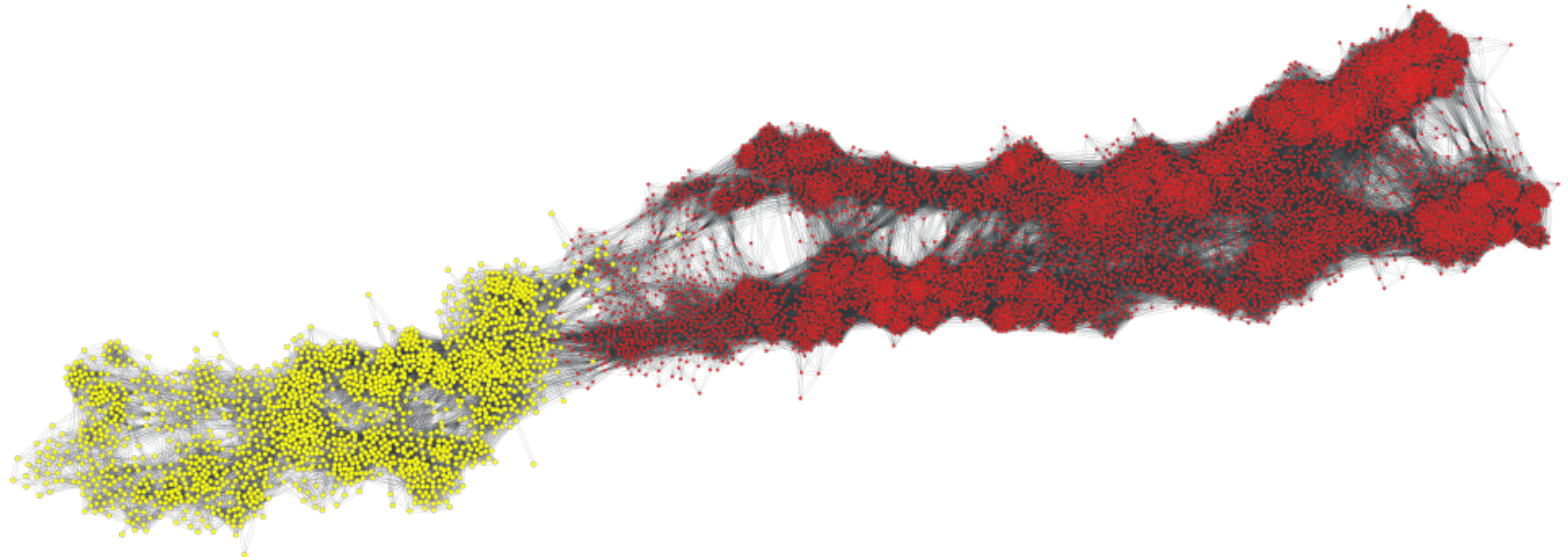
Motivation: detection of small clusters in large and noisy graphs

- Real large-scale graphs have rich local structure
- We often have to detect small clusters in large graphs:



protein-protein interaction graph,
color denotes similar functionality

Rather than partitioning graphs with nice structure



US-Senate graph,
nice bi-partition in year 1865 around the end of the American civil war

Our goals: simple local algorithm with good theoretical guarantees

Detection of small clusters in large graphs call for new methods that

- run in time proportional to the size of the output (but not the whole graph),
- supported by good theoretical guarantees,
- require few tuning parameters.

Our goals: simple local algorithm with good theoretical guarantees

(Approximate Personalized) PageRank?

- run in time proportional to the size of the output (but not the whole graph), ✓✓
- supported by good theoretical guarantees, ✗
- require few tuning parameters. ✗

Our goals: simple local algorithm with good theoretical guarantees

Graph cut or max-flow approach?

- run in time proportional to the size of the output (but not the whole graph), ✓
- supported by good theoretical guarantees, ✓
- require few tuning parameters. ✗✗

Our goals: simple local algorithm with good theoretical guarantees

This work

Let's replace PageRank with an even simpler model

- run in time proportional to the size of the output (but not the whole graph), ✓✓
- supported by good theoretical guarantees, ✓
- require few tuning parameters. ✓

Existing local graph clustering methods

Spectral diffusions

based on the dynamics of *random walks*

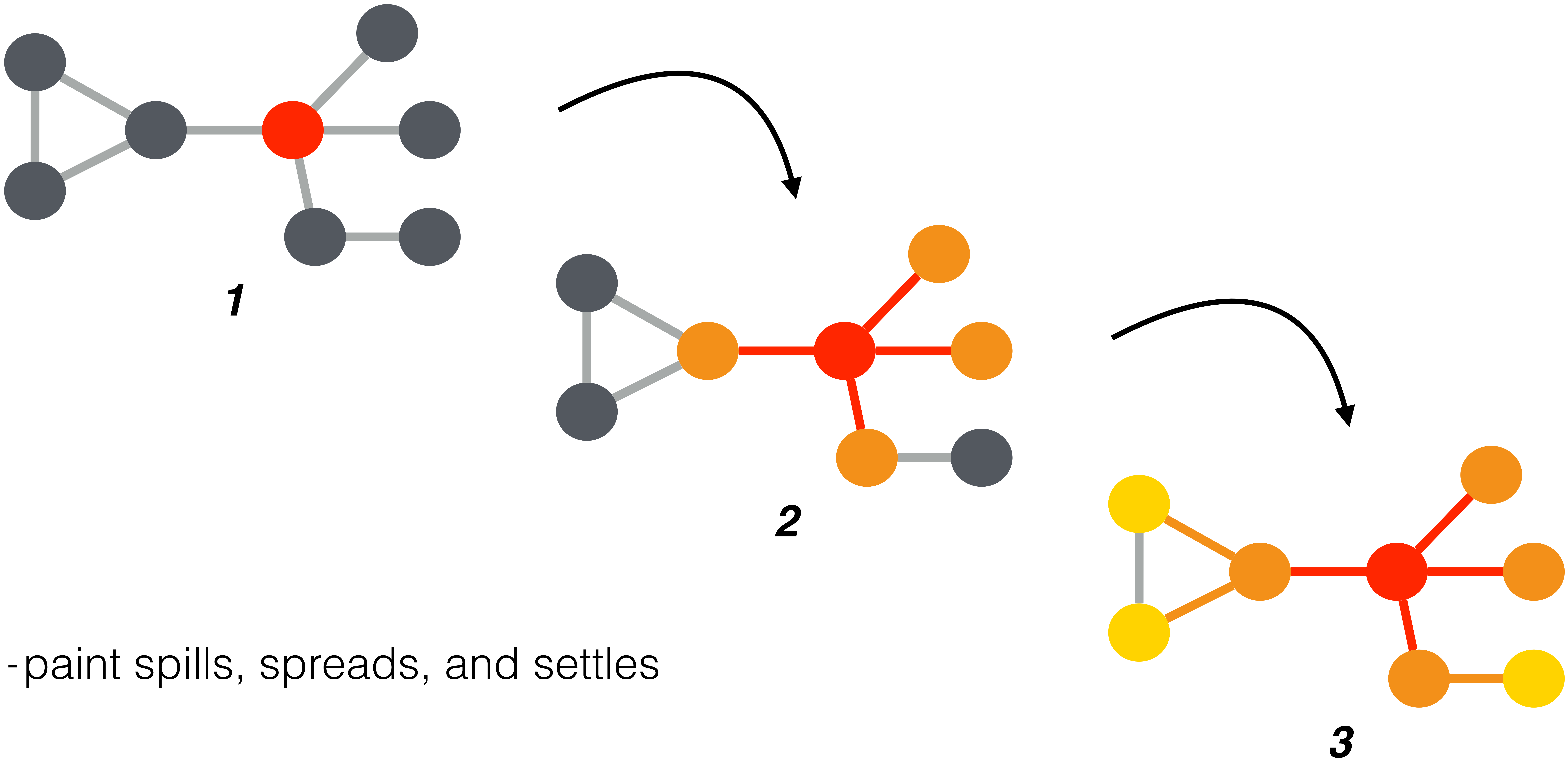
e.g., Approx. PageRank [Andersen *et al.*, 2006]

Combinatorial diffusions

based on the dynamics of *network flows*

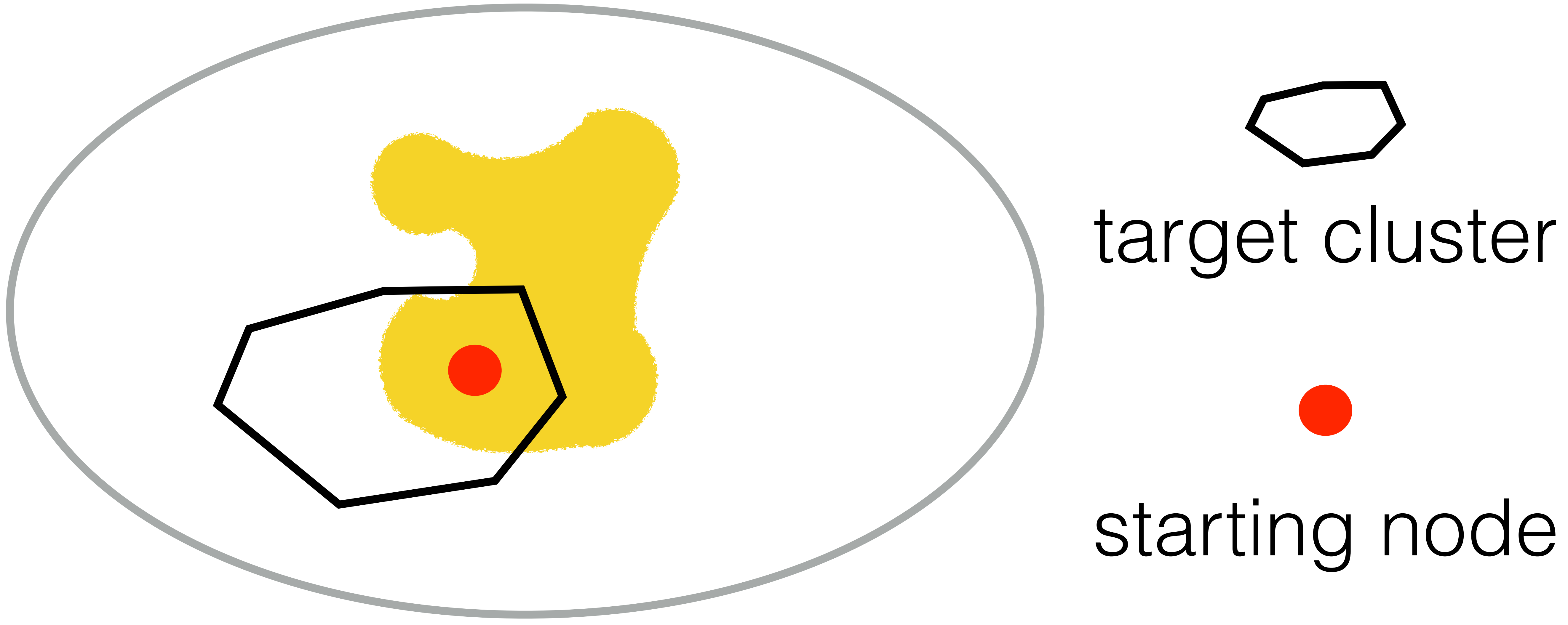
e.g., Capacity Releasing Diffusion [Wang *et al.*, 2017]

Diffusion as physical phenomenon



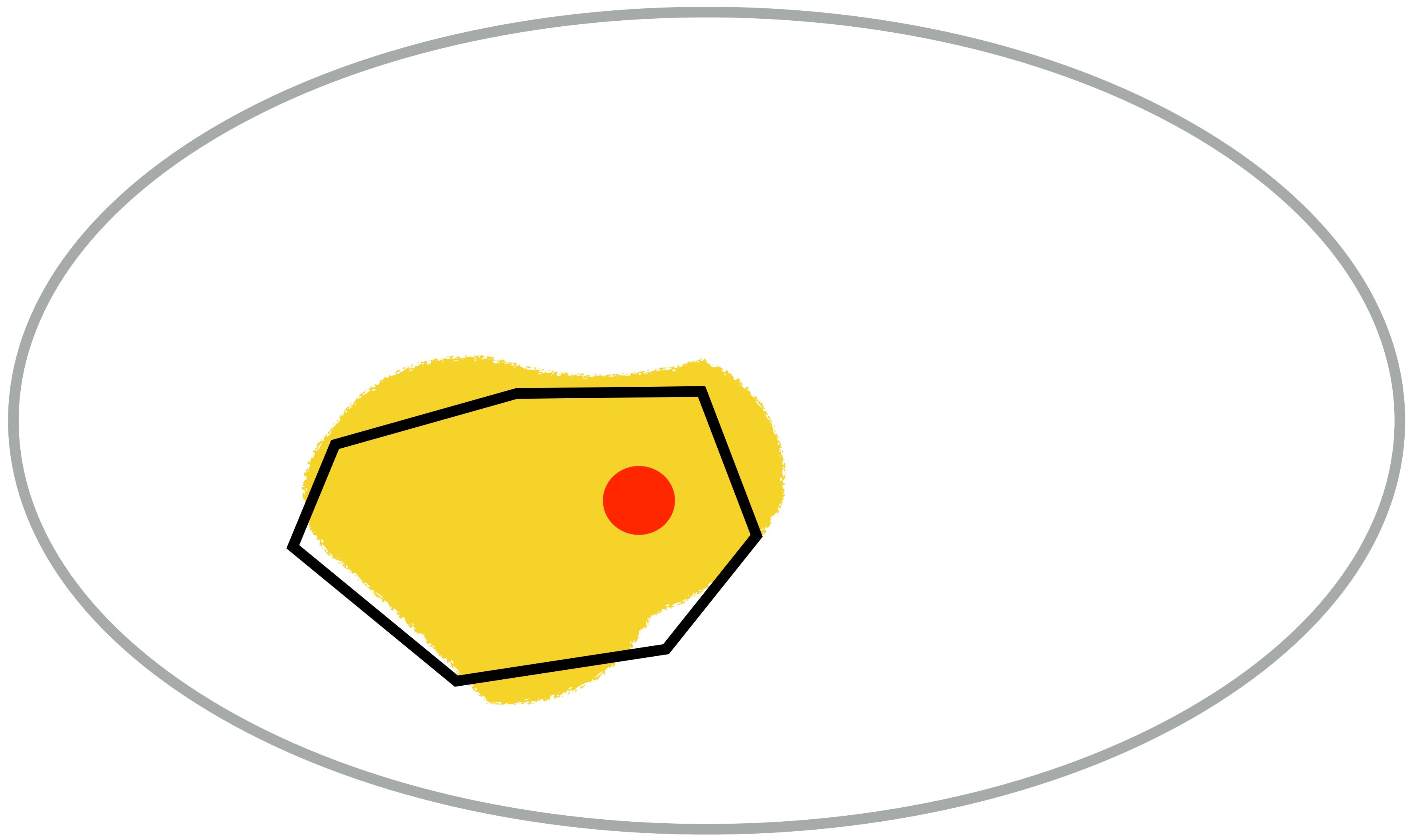
-paint spills, spreads, and settles

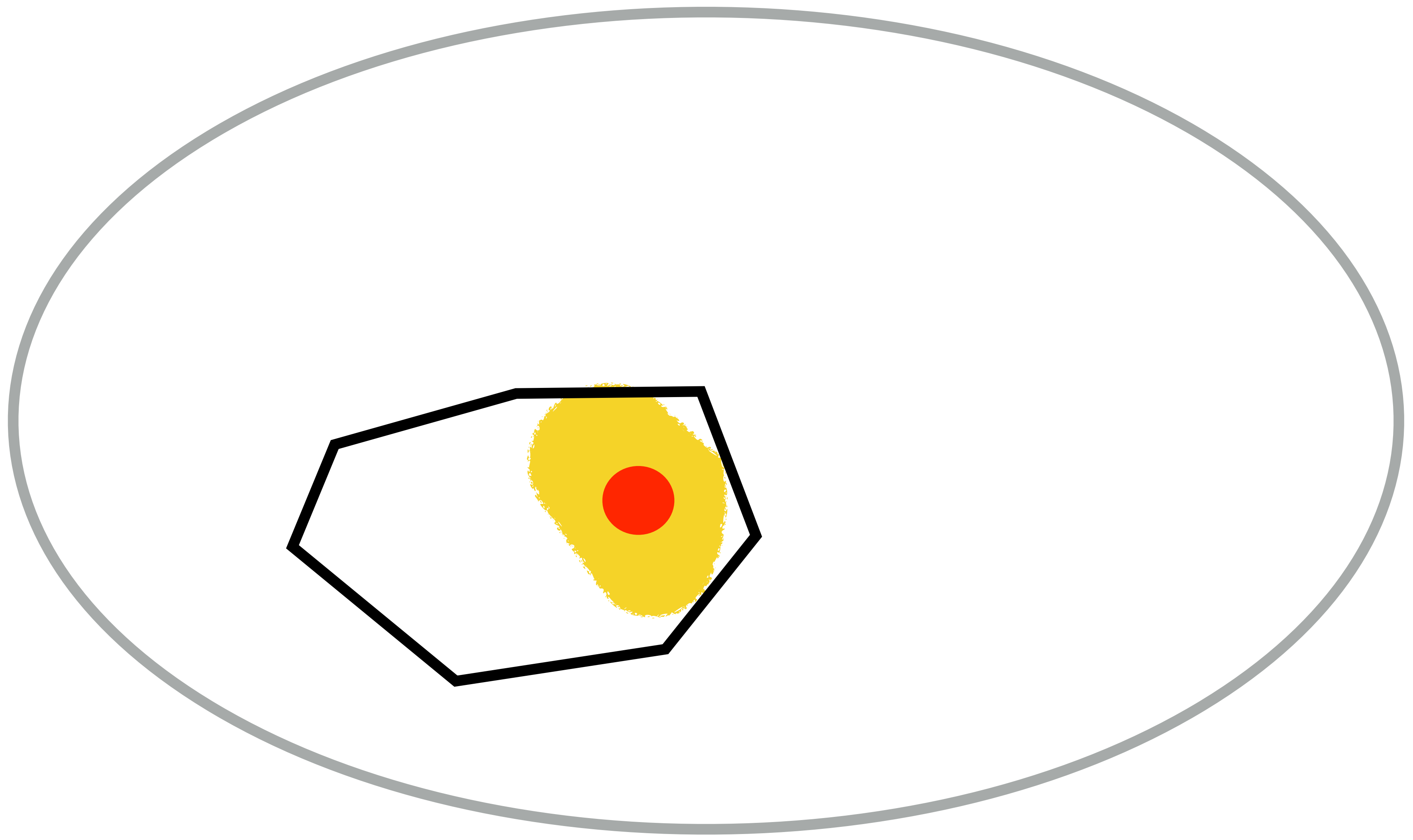
Spectral diffusions leak mass



- low precision
- low recall

Combinatorial diffusions are hard to tune





- strong theoretical guarantees
- work very well if tuned correctly

- poor performance if not tuned well

New local graph clustering paradigm

Spectral diffusions **Combinatorial diffusions**

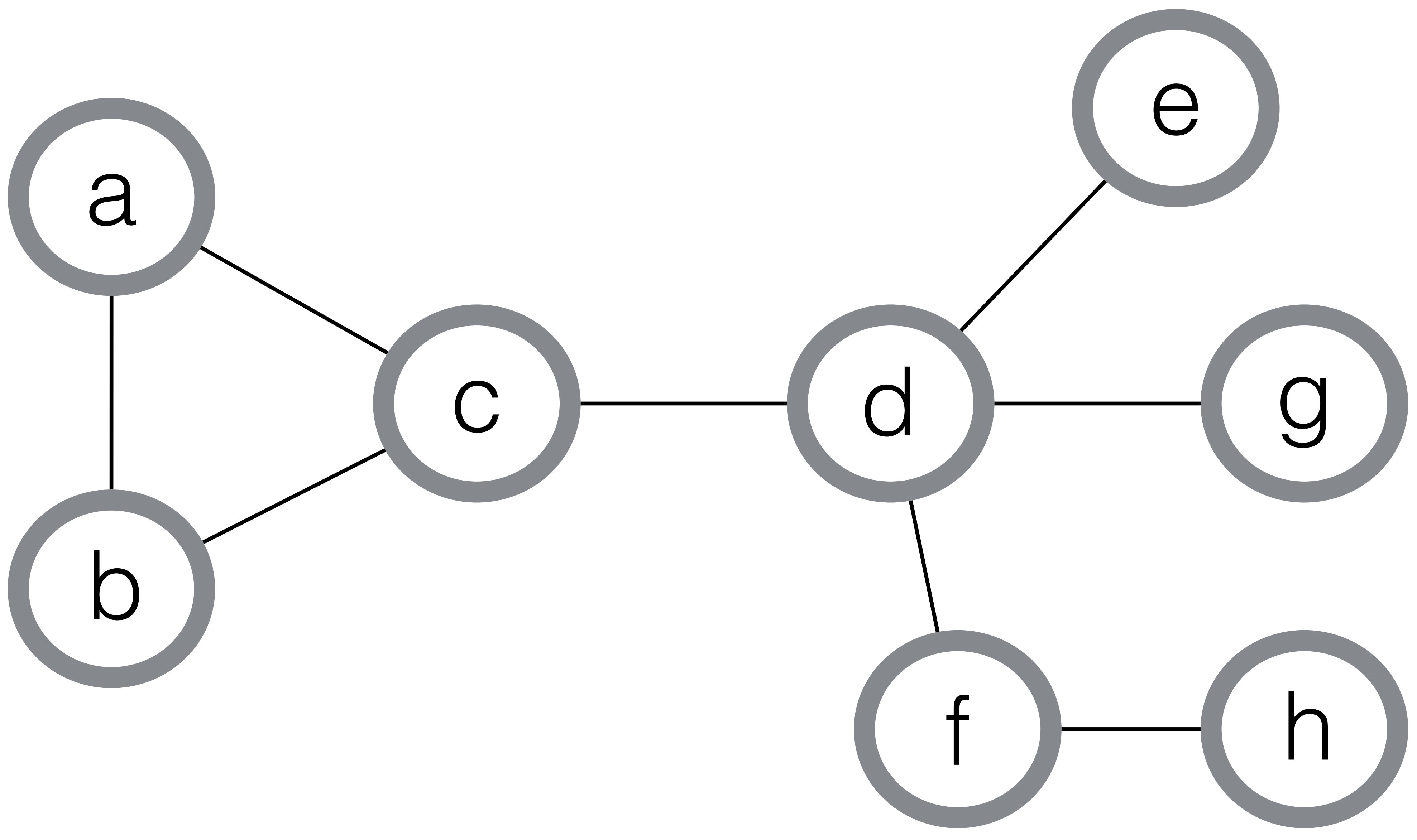
***p*-Norm flow diffusions**

based on the idea of
p-norm network flow

- as **fast** as spectral methods 🙂
- asymptotically as **strong** as combinatorial methods 🙂
- intuitive interpretation, **simple** algorithm 🙂
- **fewer tuning** parameters (than both spectral and combinatorial) 🙂

Notations and definitions

- Undirected graph $G = (V, E)$



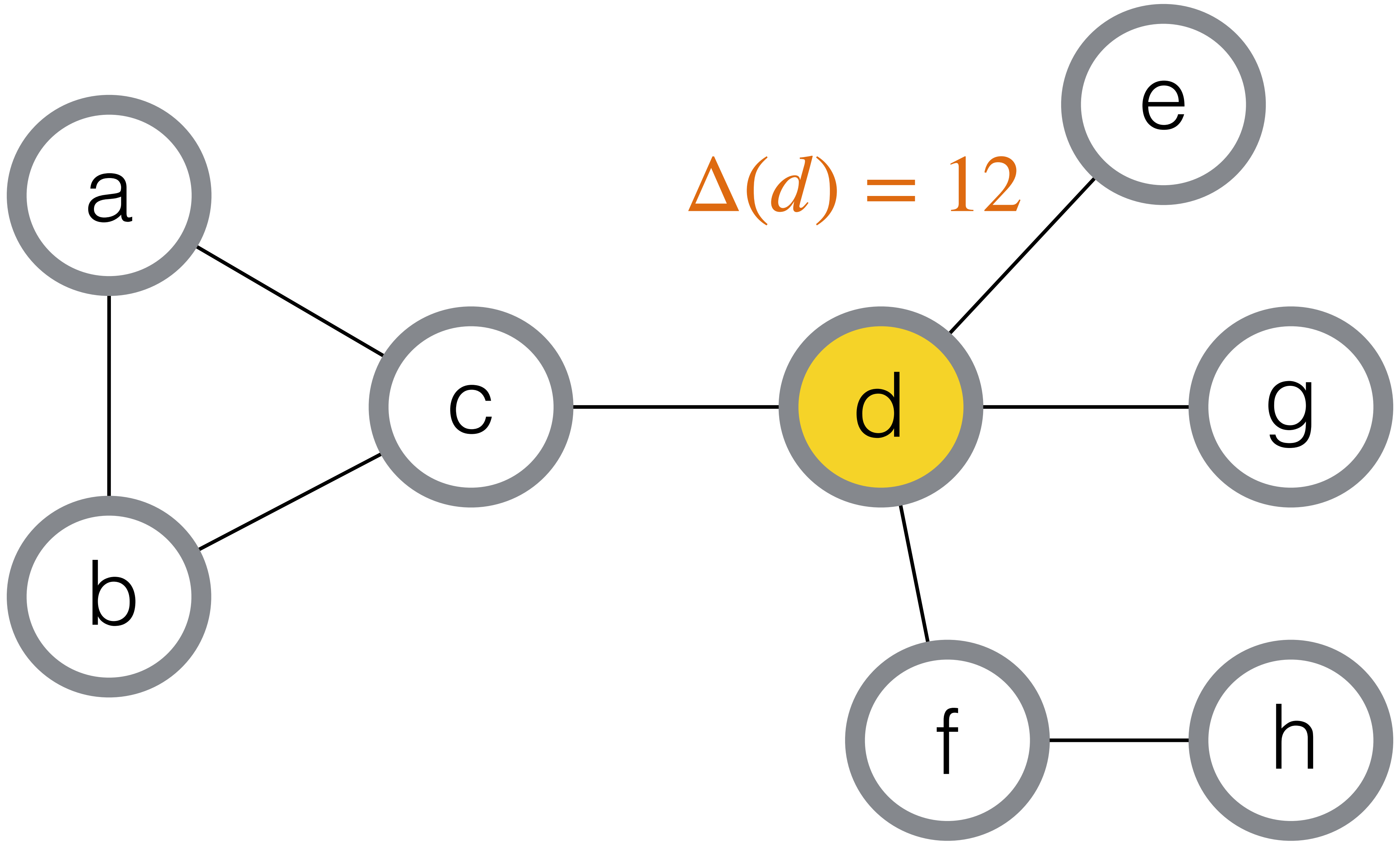
Incidence matrix B

	a	b	c	d	e	f	g	h
(a,b)	1	-1						
(a,c)	1		-1					
(b,c)		1	-1					
(c,d)			1	-1				
(d,e)				1	-1			
(d,f)				1		-1		
(d,g)				1			-1	
(f,h)						1		-1

- B is $|E| \times |V|$ *signed incidence matrix* where the row of edge (u, v) has two non-zero entries, -1 at column u and 1 at column v
- Ordering of edges and direction is arbitrary

Notations and definitions

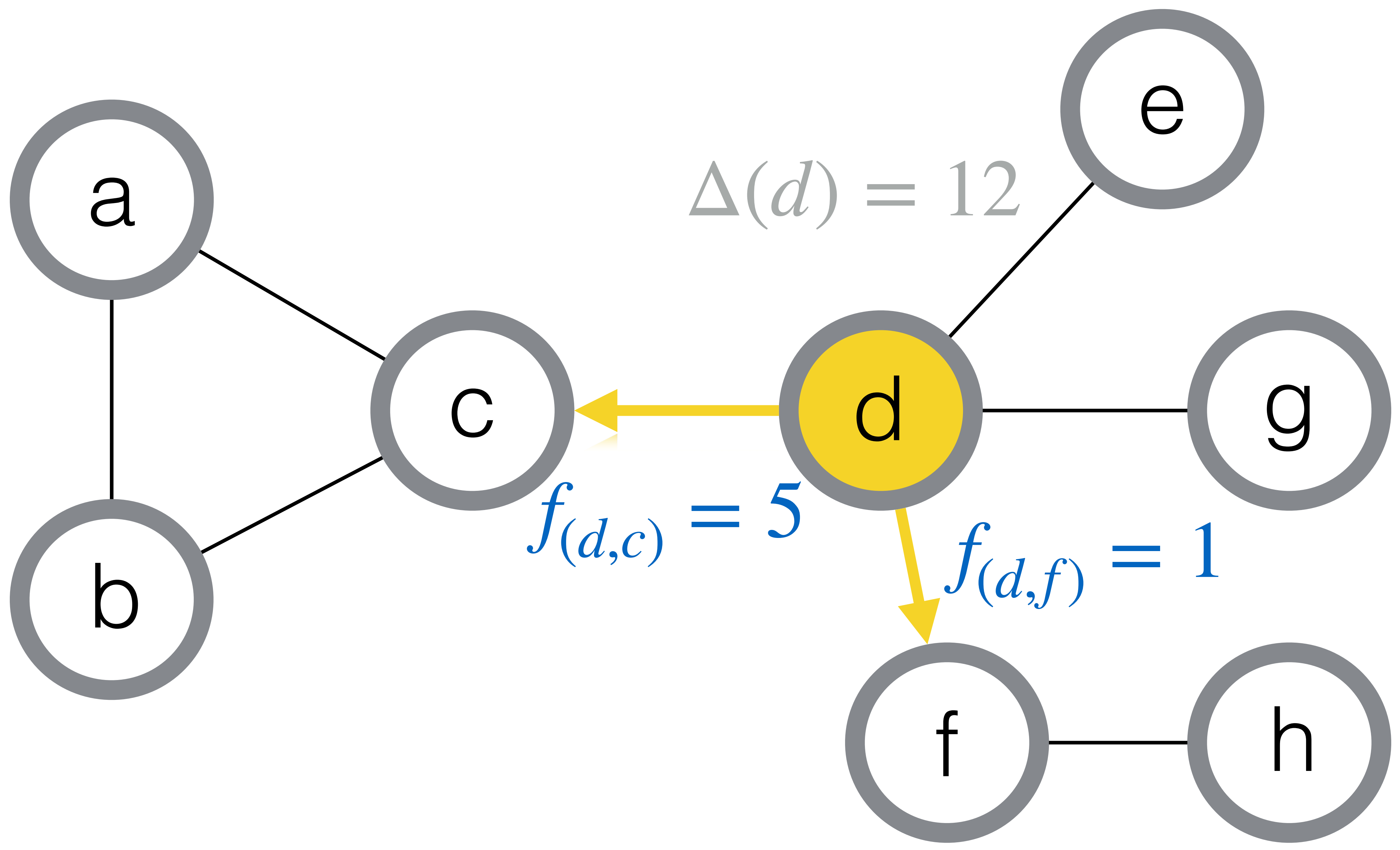
- $\Delta \in \mathbb{R}_+^{|V|}$ specifies **initial mass** on nodes.



Notations and definitions

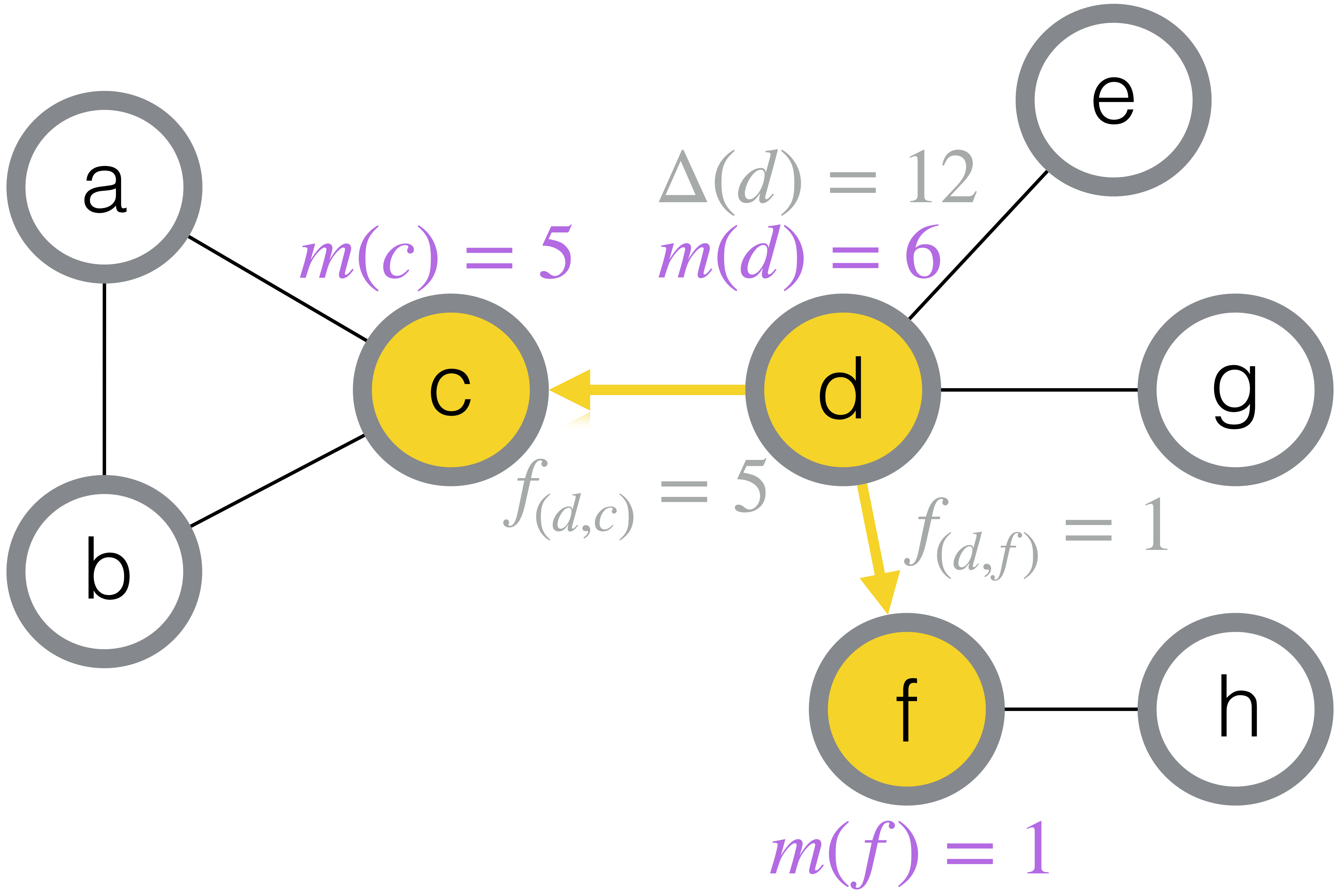
- $\Delta \in \mathbb{R}_+^{|V|}$ specifies initial mass on nodes.

- $f \in \mathbb{R}^{|E|}$ specifies the **amount of flow**.



Notations and definitions

- $\Delta \in \mathbb{R}_+^{|V|}$ specifies initial mass on nodes.
- $f \in \mathbb{R}^{|E|}$ specifies the amount of flow.
- $m := B^\top f + \Delta$ specifies **net mass** on nodes.

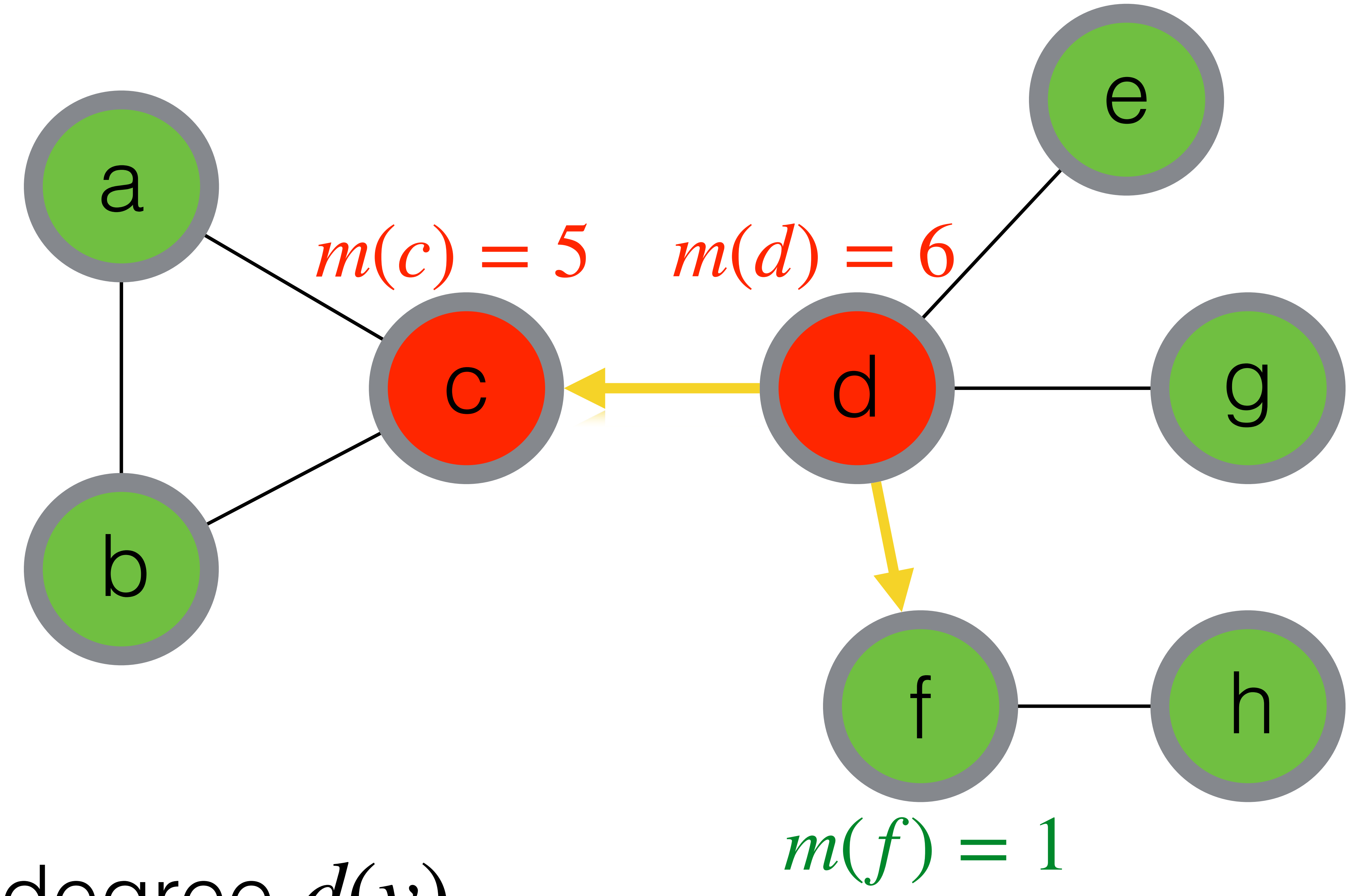


Notations and definitions

- $\Delta \in \mathbb{R}_+^{|V|}$ specifies initial mass on nodes.

- $f \in \mathbb{R}^{|E|}$ specifies the amount of flow.

- $m := B^\top f + \Delta$ specifies net mass on nodes.



- Each node v has **capacity** equal to its degree $d(v)$.

- A flow f is **feasible** if $[B^\top f + \Delta](v) \leq d(v), \forall v$.

p-Norm flow diffusions - problem formulation

- We formulate diffusion process on graph as optimization:

$$\text{minimize } \|f\|_p$$
$$\text{subject to: } B^\top f + \Delta \leq d$$

Nonlinear 🙂

Only one tuning parameter 🙂

- Out of all feasible flows f, we are interested in the one having minimum p-norm, where $p \in [2,\infty)$.

p-Norm flow diffusions - problem formulation

- We formulate diffusion process on graph as optimization:

$$\text{minimize } \|f\|_p$$
$$\text{subject to: } B^\top f + \Delta \le d$$

- **Versatility:** different p-norm flows explore different structures in a graph
- **Locality:** $\|f^*\|_0 \le |\Delta| := \sum_{v \in V} \Delta(v)$

p-Norm flow diffusions - problem formulation

- We formulate diffusion process on graph as optimization:

$$\text{minimize } \|f\|_p$$
$$\text{subject to: } B^\top f + \Delta \leq d$$

- The dual problem provides node embeddings

$$\text{minimize } x^\top (d - \Delta)$$
$$\text{subject to: } \|Bx\|_q \leq 1$$
$$x \geq 0$$

Biased towards seed node

$1/p + 1/q = 1$

- Obtain a cluster by applying sweep cut on x

p-Norm flow diffusions - local clustering guarantees

- Conductance of target cluster C

$$\phi(C) = \frac{|\{(u,v) \in E : u \in C, v \notin C\}|}{\min\{\mathbf{vol}(C), \mathbf{vol}(V \backslash C)\}} \quad \text{where } \mathbf{vol}(C) := \sum_{v \in C} d(v)$$

- Seed set $S := \text{supp}(\Delta)$.

- Assumption (sufficient overlap):

$$\mathbf{vol}(S \cap C) \geq \beta \mathbf{vol}(S)$$
$$\mathbf{vol}(S \cap C) \geq \alpha \mathbf{vol}(C)$$
$$\alpha, \beta \geq \frac{1}{\log^t \mathbf{vol}(C)} \text{ for some } t$$

- The output cluster $\tilde{C}$ satisfies

$$\phi(\tilde{C}) \leq \tilde{\mathcal{O}}(\phi(C)^{1-1/p})$$

- Cheeger-type bound $\phi(\tilde{C}) \leq \tilde{\mathcal{O}}(\sqrt{\phi(C)})$ for $p = 2$

- Constant approximate $\phi(\tilde{C}) \leq \tilde{\mathcal{O}}(\phi(C))$ for $p \to \infty$

p-Norm flow diffusions - local clustering guarantees

- Conductance of target cluster C

$$\phi(C) = \frac{|\{(u,v) \in E : u \in C, v \notin C\}|}{\min\{\mathbf{vol}(C), \mathbf{vol}(V\backslash C)\}} \quad \text{where } \mathbf{vol}(C) := \sum_{v \in C} d(v)$$

- Seed set $S := \text{supp}(\Delta)$.
- Assumption (sufficient overlap):

$$\mathbf{vol}(S \cap C) \geq \beta \mathbf{vol}(S) \qquad \alpha, \beta \geq \frac{1}{\log^t \mathbf{vol}(C)} \text{ for some } t$$
$$\mathbf{vol}(S \cap C) \geq \alpha \mathbf{vol}(C)$$

- The output cluster $\tilde{C}$ satisfies

$$\phi(\tilde{C}) \leq \tilde{\mathcal{O}}(\phi(C)^{1-1/p})$$

- Cheeger-type bound $\phi(\tilde{C}) \leq \tilde{\mathcal{O}}(\sqrt{\phi(C)})$ for $p = 2$
- Constant approximate $\phi(\tilde{C}) \leq \tilde{\mathcal{O}}(\phi(C))$ for $p \to \infty$

Proof based on analysis of primal and dual objective and constraints.

Larger *p* penalizes more on the flows that cross "bottleneck" edges, leading to less leakage.

p-Norm flow diffusions - simple strongly local algorithm

- Solve an equivalent penalized dual formulation by a variant of randomized coordinate descent.

Initially each node has a net mass equals the initial mass.
Iterate:
  Pick a node v whose net mass exceeds its capacity.
  Send excess mass to its neighbors.
  Update net mass.

p-Norm flow diffusions - simple strongly local algorithm

- Solve an **equivalent** penalized dual formulation by a variant of randomized coordinate descent.

Initially each node has a net mass equals the initial mass.
Iterate:
Pick a node v whose net mass exceeds its capacity.
Send excess mass to its neighbors.
Update net mass.

- Worst-case running time $\mathcal{O}\left(|\Delta|\left(\frac{|\Delta|}{\epsilon}\right)^{2/q-1}\log\frac{1}{\epsilon}\right)$.

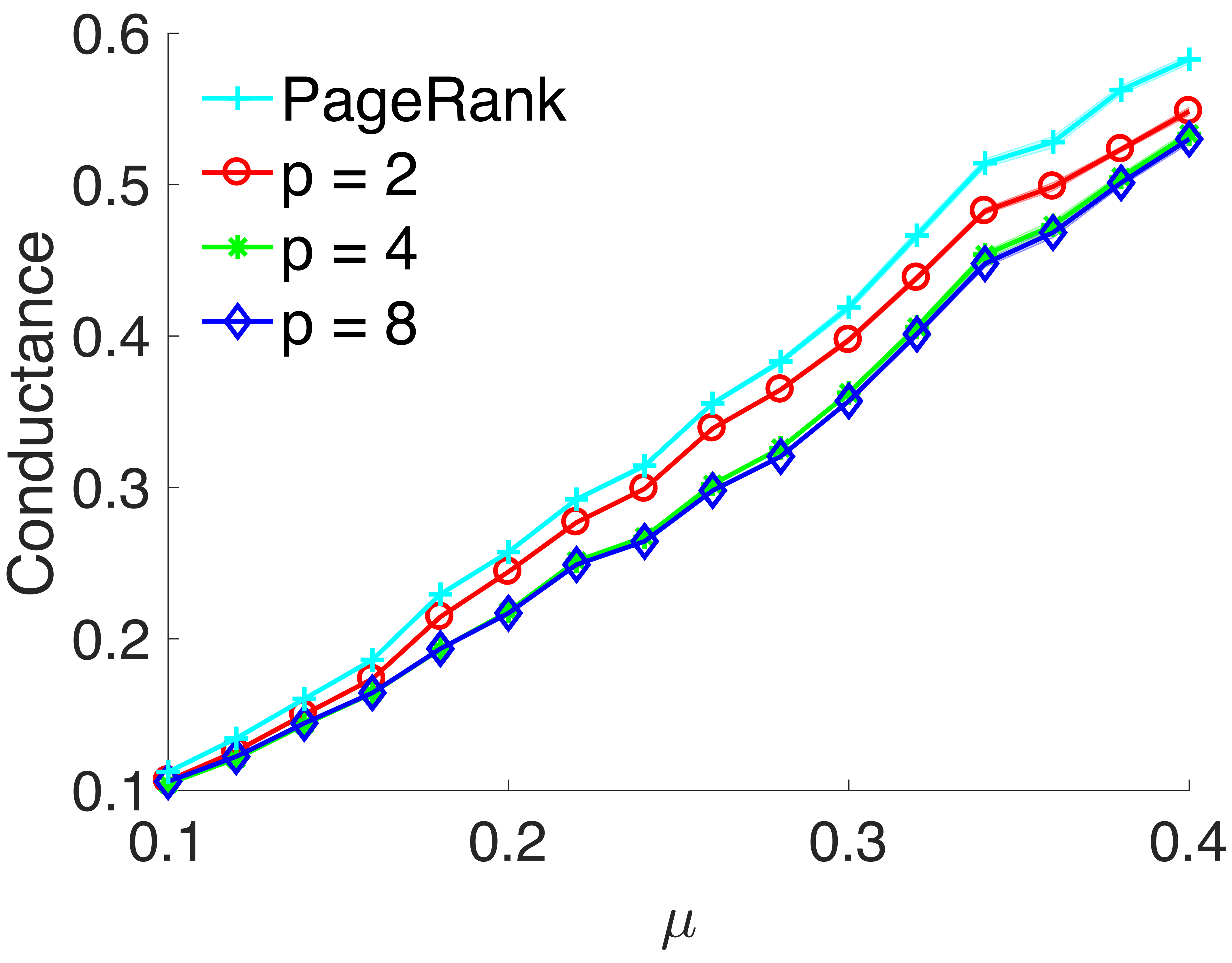
Natural tradeoff between speed and robustness to noise

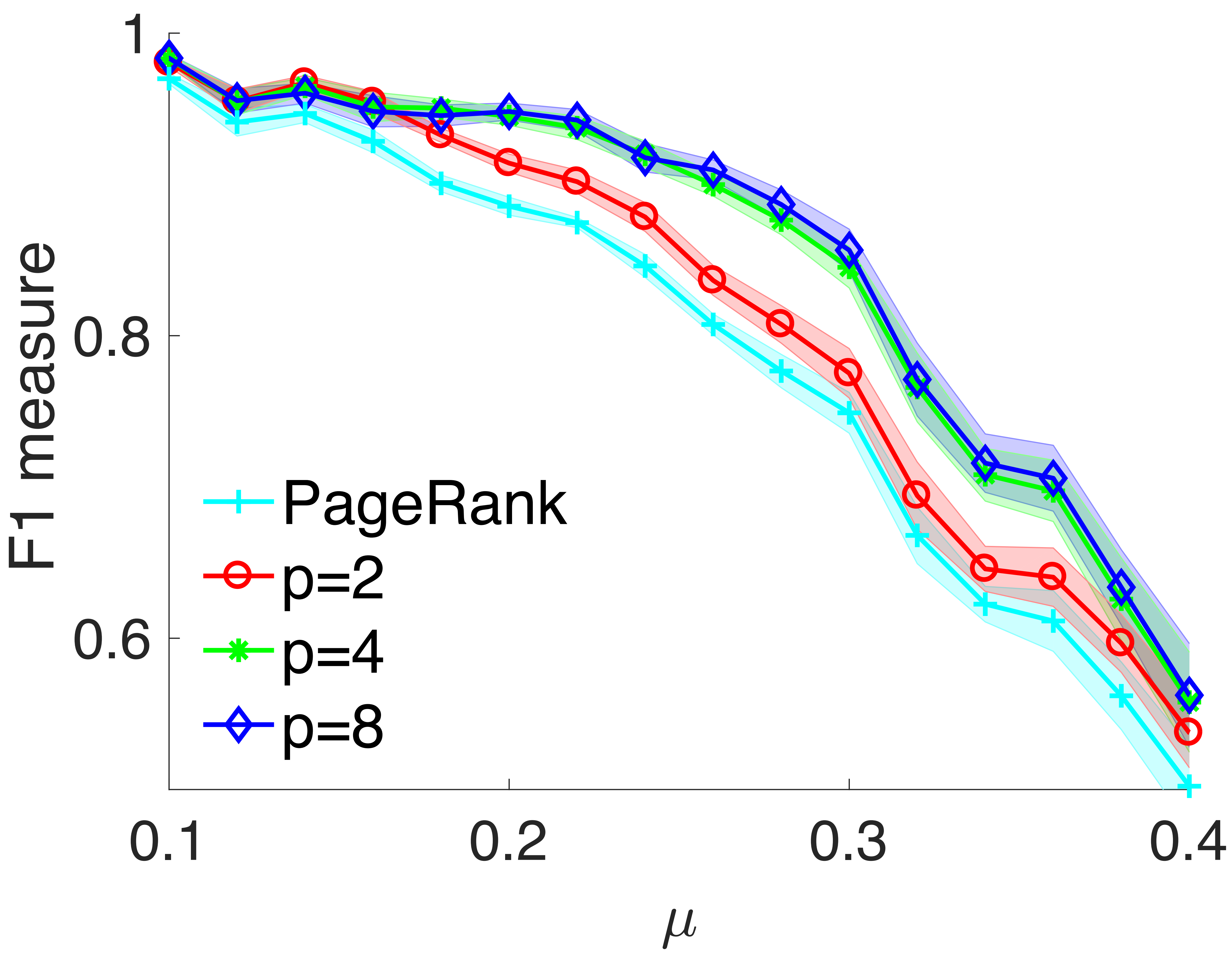
Total amount of initial mass

- Linear convergence when q = 2.

p-Norm flow diffusions - empirical performance

- LFR synthetic model
- μ is a parameter that controls noise, the higher the more noise.





p-Norm flow diffusions - empirical performance

- Facebook social network for Colgate University, students in Class of 2009

	PageRank	p = 2	p = 4
Conductance	0.13	0.13	**0.12**
F1 measure	0.96	0.96	**0.97**

very clean ground truth

- Facebook social network for Johns Hopkins University, students of the same major

	PageRank	p = 2	p = 4
Conductance	0.25	0.23	**0.22**
F1 measure	0.83	0.85	**0.87**

average ground truth

- Orkut, large-scale on-line social network, user-defined group

	PageRank	p = 2	p = 4
Conductance	0.37	0.35	**0.33**
F1 measure	0.66	0.71	**0.73**

very noisy ground truth

Julia implementation: pNormFlowDiffusion on GitHub

- Includes demonstrations and visualizations on LFR and Facebook social networks.
- Contains all code to reproduce the results in our paper.

	Local running time, **fast computation**	Good **theoretical guarantee**	Simple algorithm, **less tuning**
Spectral diffusion (e.g. PageRank)	✓✓	✗	✗
Combinatorial diffusion (e.g. CRD)	✓	✓	✗✗
p-Norm flow diffusion	✓✓	✓	✓

Thank you!