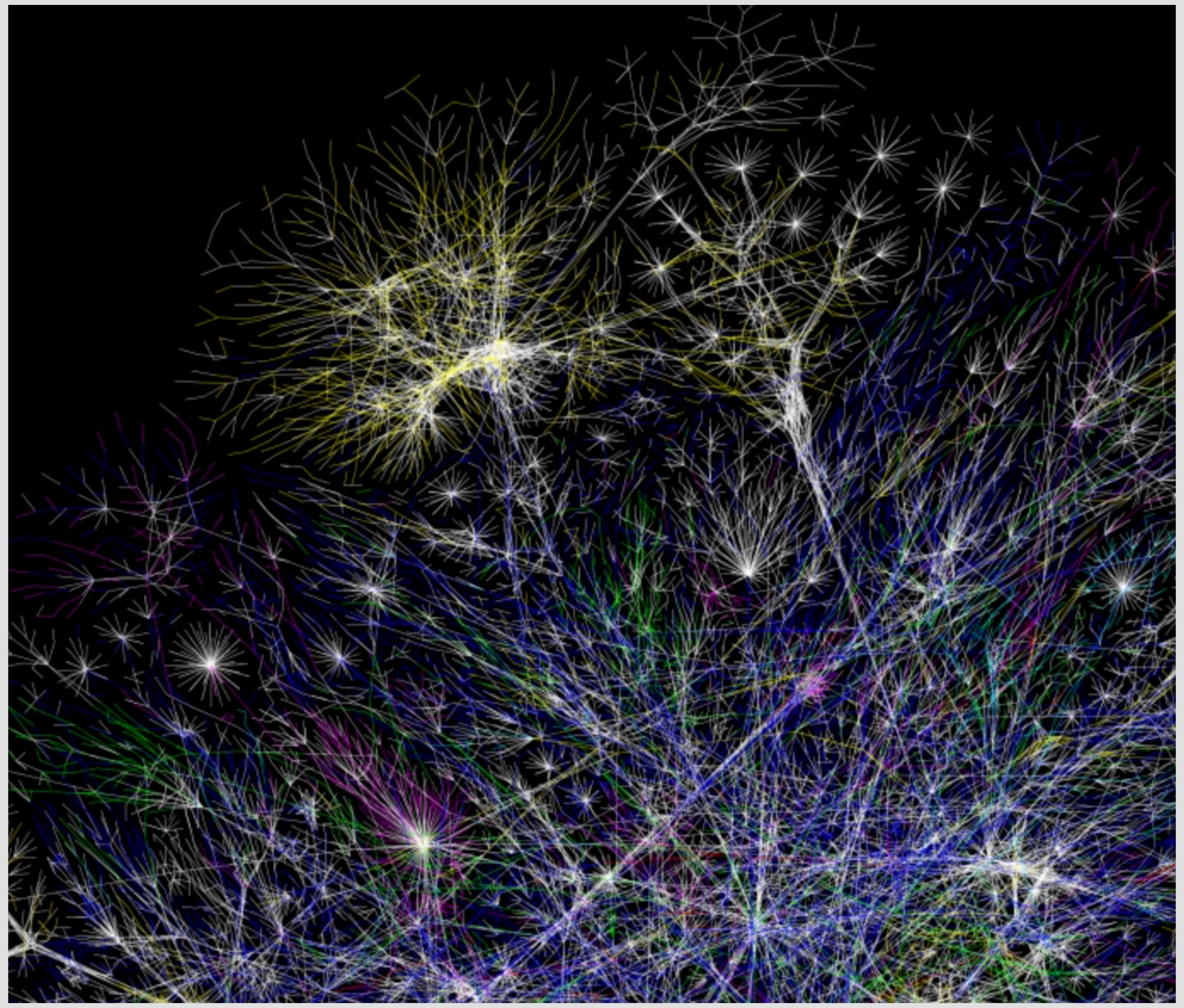
Combinatorial ℓ_p-norm Correlation Clustering

Sami Davies (UC Berkeley/ Simons), Benjamin Moseley (CMU), Heather Newman (CMU)

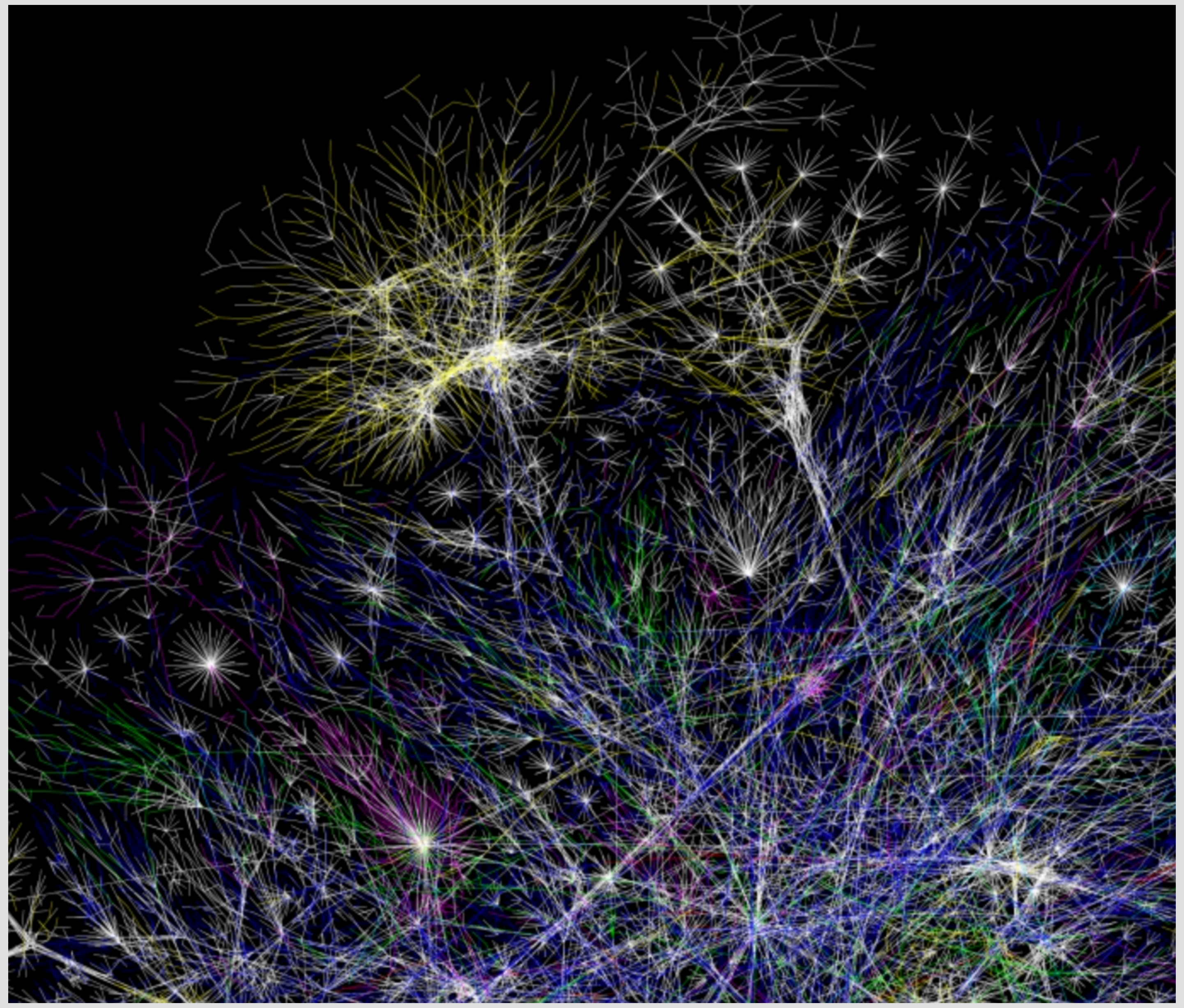
Community detection



- ✦ Creating large-scale maps with meta nodes
- ✦ Understanding *community* vs *aggregate* features
- ✦ Identifying topological/ spectral properties

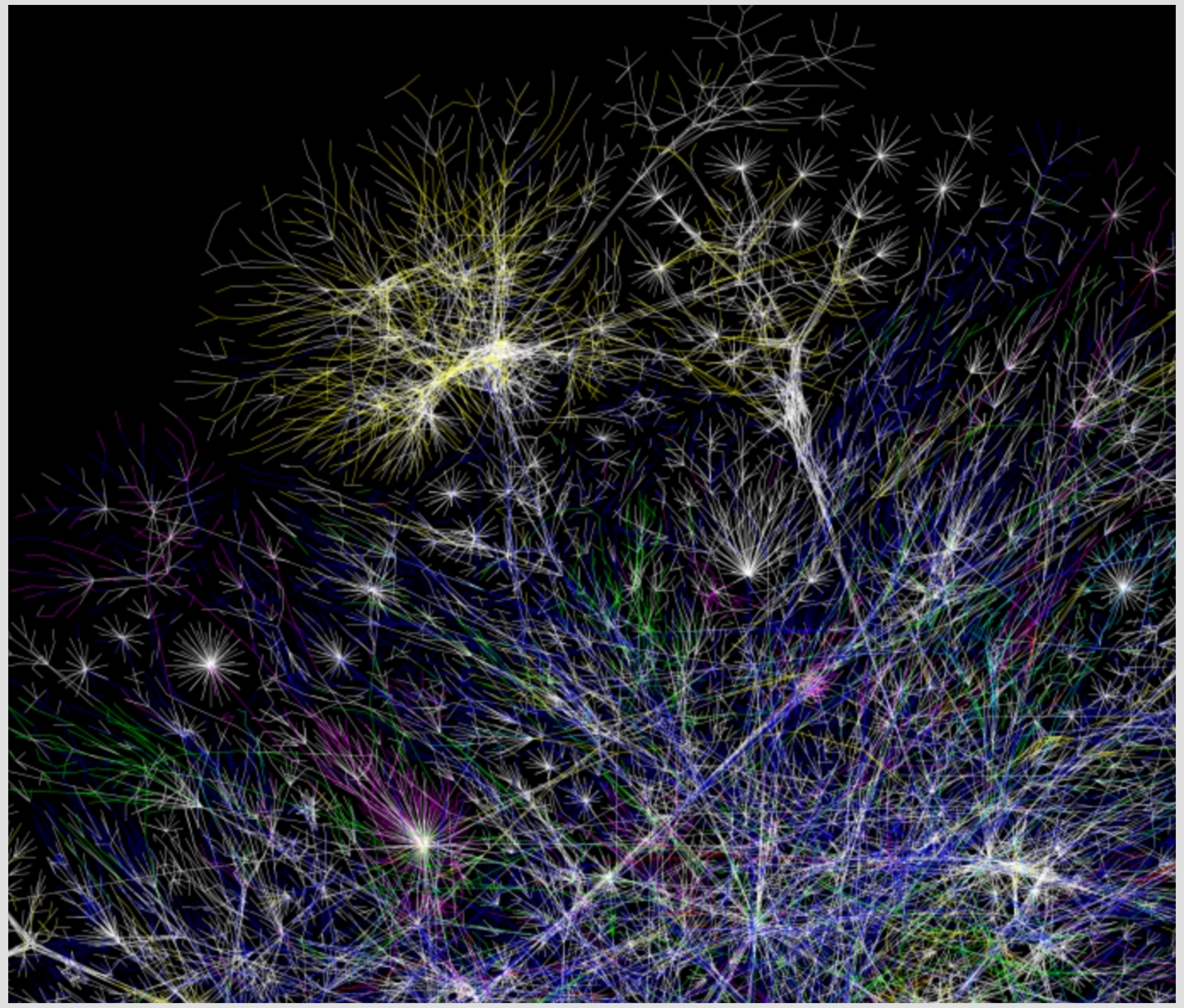
Community detection

Many different models



- Hierarchical clustering — Best for data with underlying heirarchy
- Minimum cut clustering — Fixed # of clusters
- Girvan Newman algorithm — Runtime $O(m^2n)$
- Modularity maximization — Doesn't find small clusters
-

Community detection



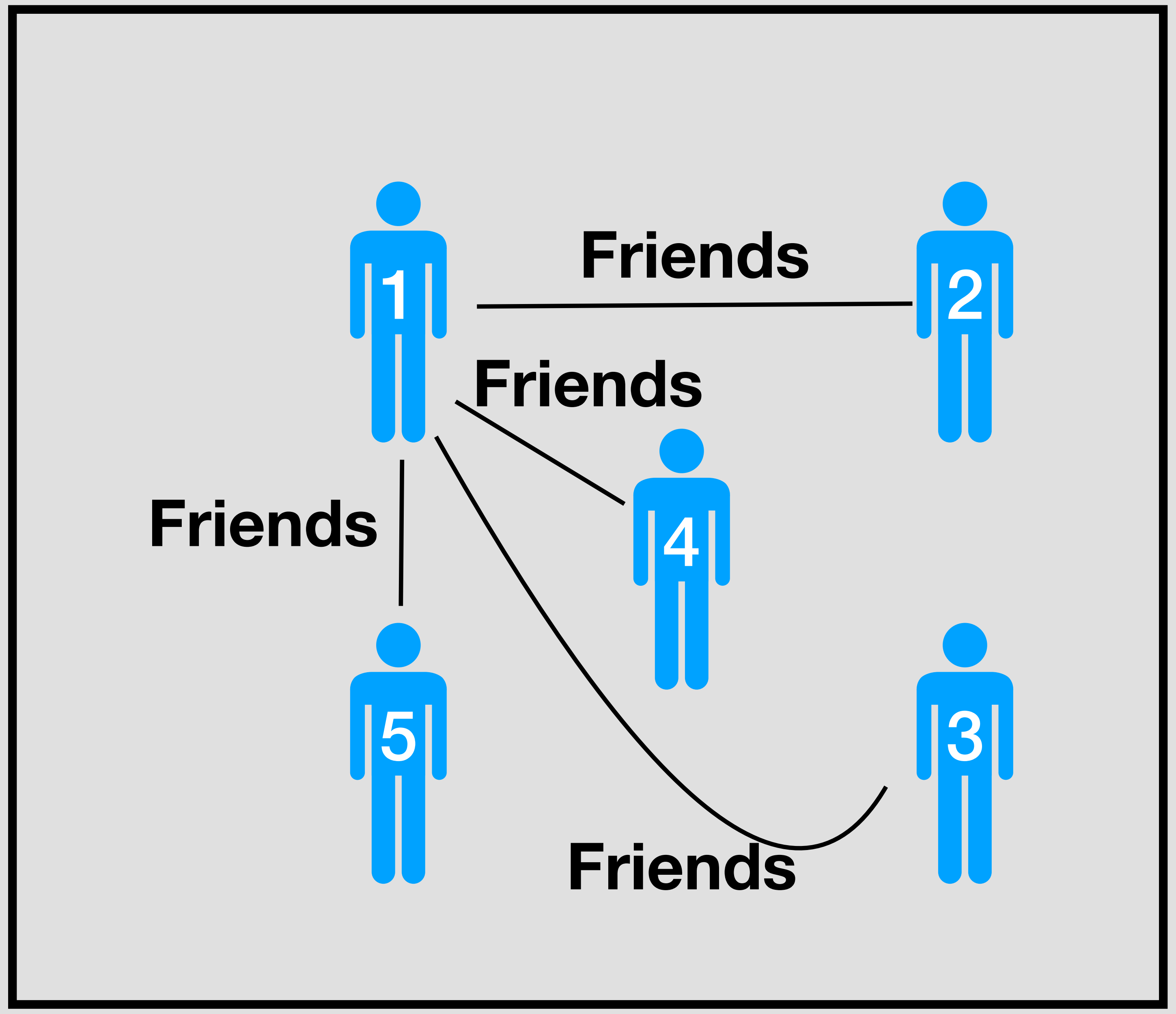
Many different models

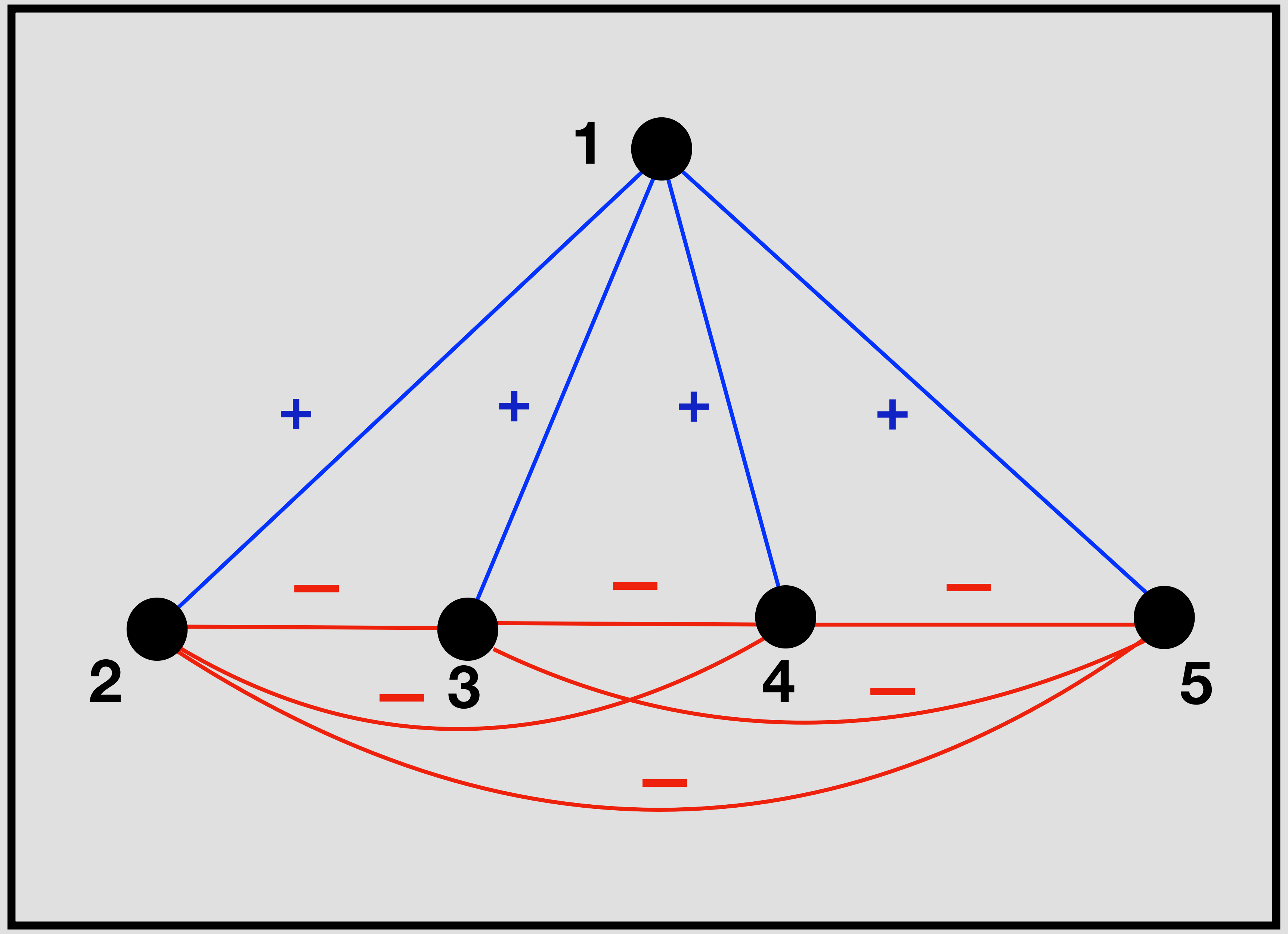
- ✦ Hierarchical clustering — Best for data with underlying heirarchy
- ✦ Minimum cut clustering — Fixed # of clusters
- ✦ Girvan Newman algorithm — Runtime $O(m^2 n)$
- ✦ Modularity maximization — Doesn't find small clusters
- ✦ **Correlation clustering**

Correlation clustering

Model:

- Cluster ***similar*** nodes together, separate ***dissimilar*** nodes
- No pre-fixed # of clusters, complete unweighted graph



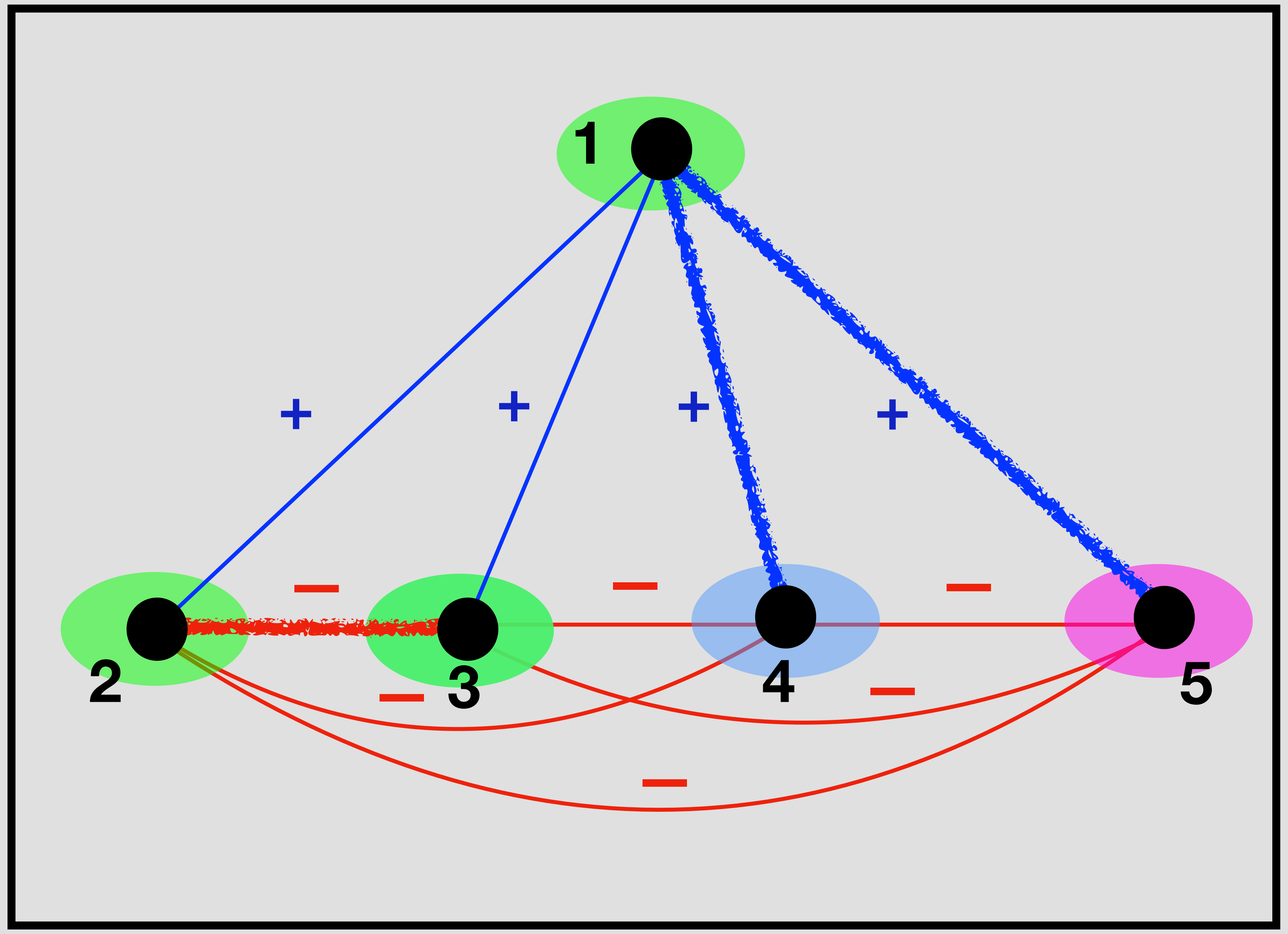


Correlation clustering

Model:

- Cluster ***similar*** nodes together, separate ***dissimilar*** nodes
- No pre-fixed # of clusters, complete unweighted graph
- Edge (u,v) in ***disagreement*** w.r.t C if
  - ✦ (+) with u, v different clusters or
  - ✦ (–) with u, v same cluster

Original objective for Correlation Clustering = minimize # of edges in disagreement



ℓ_p Correlation clustering

Model:

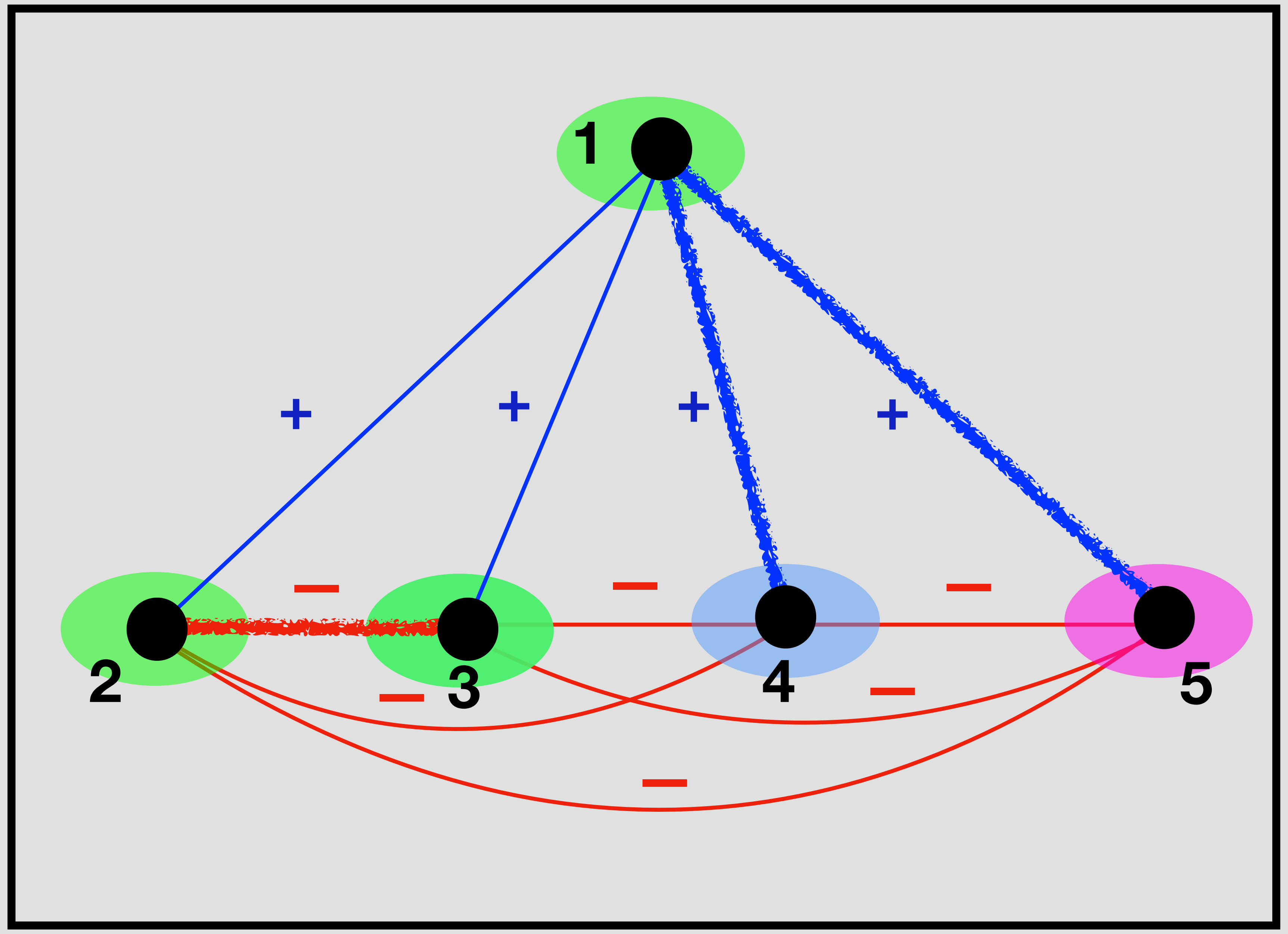
- Cluster ***similar*** nodes together, separate ***dissimilar*** nodes
- No pre-fixed # of clusters, complete unw
- Edge (u,v) in ***disagreement*** w.r.t C if
  - (+) with u, v different clusters or
  - (–) with u, v same cluster

p small = global obj ⟷ p large = local/fair obj

- y_i^C = # disagreements w.r.t. C incident to v_i
- Goal: find $\mathbf{argmin}_C ||y^C||_p$

$p \geq 1$

ℓ_1 = original cc
ℓ_∞ = min max norm



Previous work

For ℓ_1-norm (original) objective:

- Introduced by [Bansal, Blum, Chawla '04]
- Linear time Pivot algorithm gives 3-apx

  [Ailon, Charikar, Newman JACM08] [Chierichetti, Dalvi, Kumar KDD14]
- APX-hard

  [Charikar, Guruswami, Wirth JCSS05]
- Many other active threads of research!

  [Ahmadi, Khuller, Saha IPCO19] [Veldt ICML22] [Cohen-Addad, Lee, Li, Newman FOCS23]

Previous work

For ℓ_1-norm (original) objective:

- Introduced by [Bansal, Blum, Chawla '04]
- Linear time Pivot algorithm gives 3-apx

  [Ailon, Charikar, Newman JACM08] [Chierichetti, Dalvi, Kumar KDD14]
- APX-hard

  [Charikar, Guruswami, Wirth JCSS05]
- Many other active threads of research!

  [Ahmadi, Khuller, Saha IPCO19] [Veldt ICML22] [Cohen-Addad, Lee, Li, Newman FOCS23]

For general ℓ_p-norm objectives:

- 5-approximation algorithm; NP-hard

  [Puleo, Milenkovic ICML16], [Charikar, Gupta, Schwartz IPCO17], [Kalhan, Makarychev, Zhou ICML19]
- Techniques round solution to a convex program

Previous work

For ℓ_1-norm (original) objective:

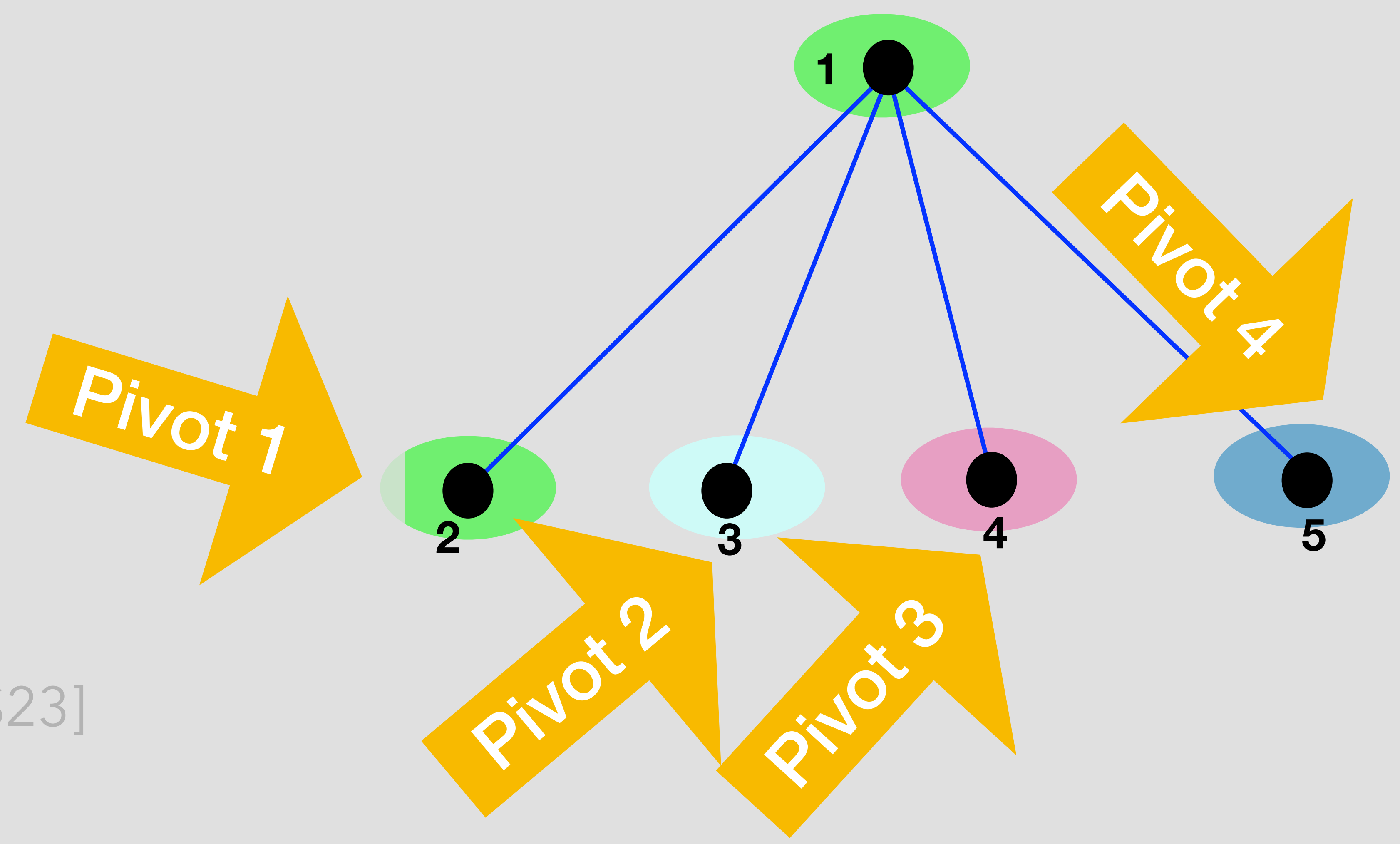
- Introduced by [Bansal, Blum, Chawla '04]
- Linear time Pivot algorithm gives 3-apx
  [Ailon, Charikar, Newman JACM08] [Chierichetti, Dalvi, Kumar KDD14]
- APX-hard
  [Charikar, Guruswami, Wirth JCSS05]
- Many other active threads of research!
  [Ahmadi, Khuller, Saha IPCO19] [Veldt ICML22] [Cohen-Addad, Lee, Li, Newman FOCS23]

For general ℓ_p-norm objectives:

- 5-approximation algorithm; NP-hard
  [Puleo, Milenkovic ICML16], [Charikar, Gupta, Schwartz IPCO17], [Kalhan, Makarychev, Zhou ICML19]
- Techniques round solution to a convex program

Pivot algorithm

- Randomly choose a *pivot* (unclustered vertex)
- Make new cluster with pivot and all its unclustered positive neighbors



Trouble with the convex program

ℓ_p-norm correlation clustering algs solve a convex program

Solving ***metric constrained*** LPs on large networks is slow!

Not very amenable to online settings

Solution specific to ***one fixed*** ℓ_p-norm

Work on solving CC LPs fast only scales to graphs with few thousand vertices!

[Ruggles et al. '20], [Sonthalia & Gilbert '20], [Veldt '22]

All-norms objective = simultaneously optimize all ℓ_p-norms

Introduced by [Azar, Epstein, Richter, Woeginger '04]

Universal algorithms produce a solution good for many objs

In, e.g., Steiner tree, TSP, clustering

Trouble with the convex program

ℓ_p-norm correlation clustering algs solve a convex program

Solving ***metric constrained*** LPs on large networks is slow!

Work on solving CC LPs fast only scales to graphs with few thousand vertices!

[Ruggles et al. '20], [Sonthalia & Gilbert '20], [Veldt '22]

Not very amenable to online settings

Solution specific to ***one fixed*** ℓ_p-norm

All-norms objective = simultaneously optimize all ℓ_p-norms

Introduced by [Azar, Epstein, Richter, Woeginger '04]

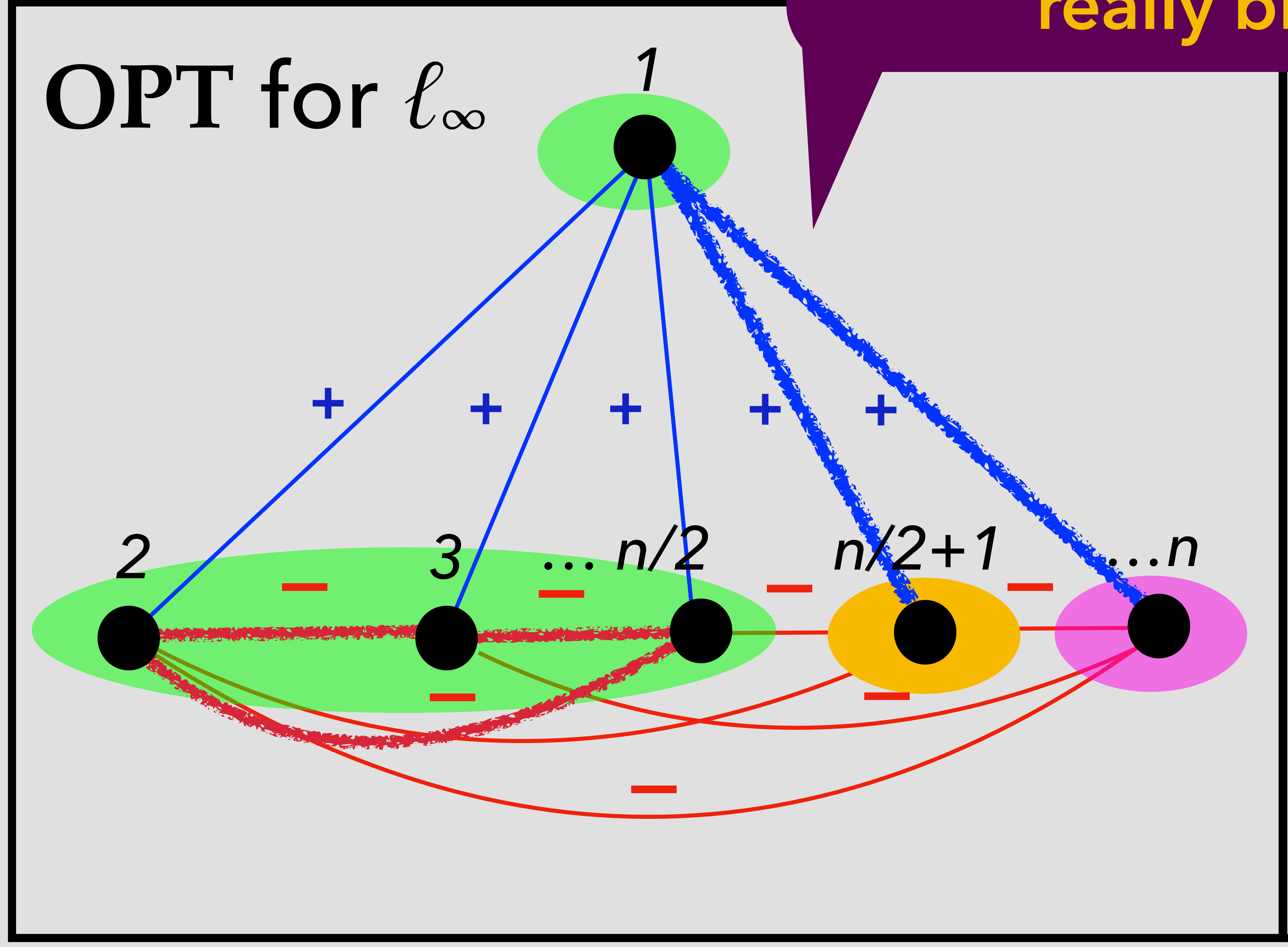
Universal algorithms produce a solution good for many objs

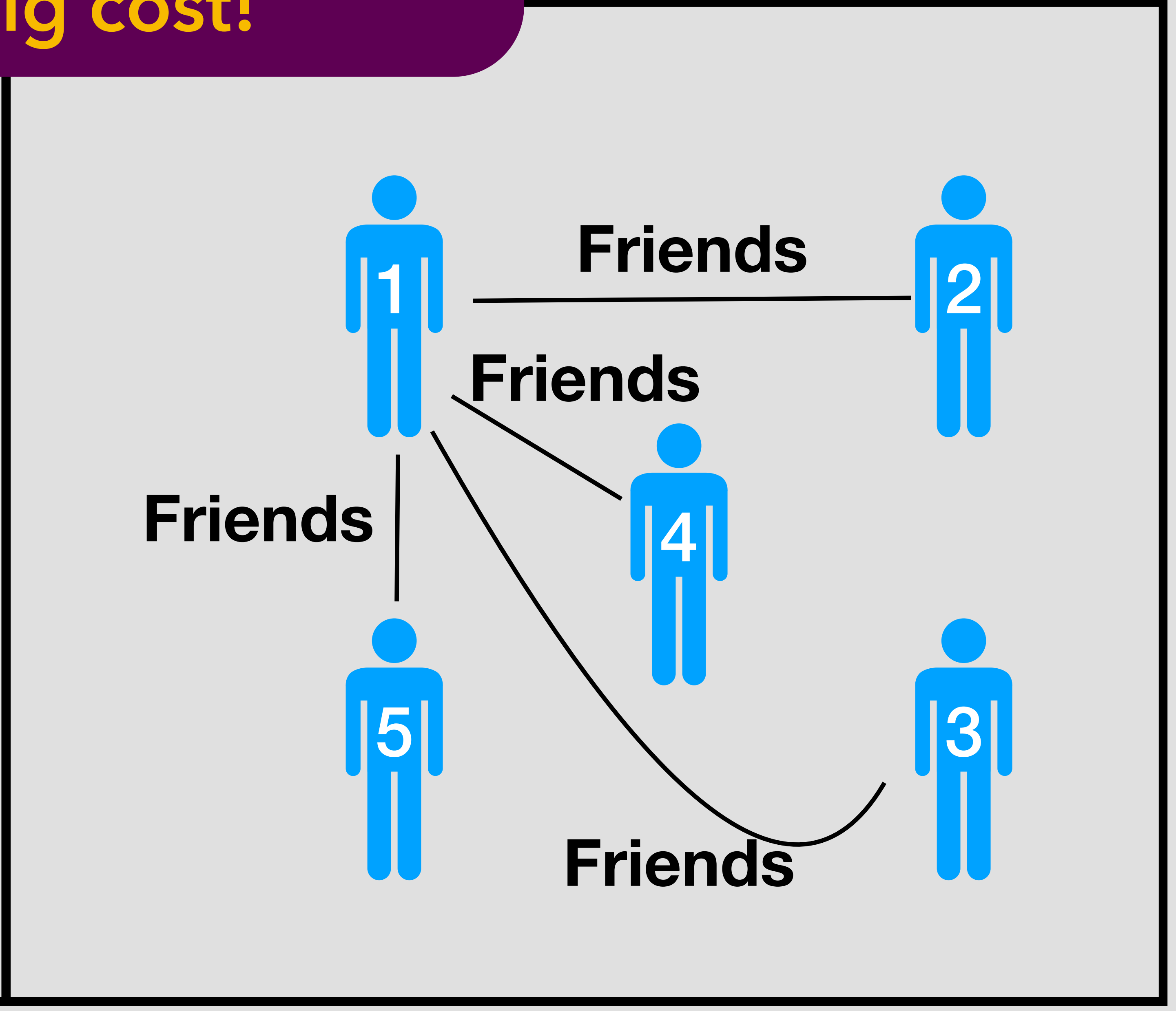
In, e.g., Steiner tree, TSP, clustering

Trouble with the convex program

OPT for one ℓ_p-norm can be really bad for others!

Cost for ℓ_1 norm is $\theta(n^2)$, **really big cost!**





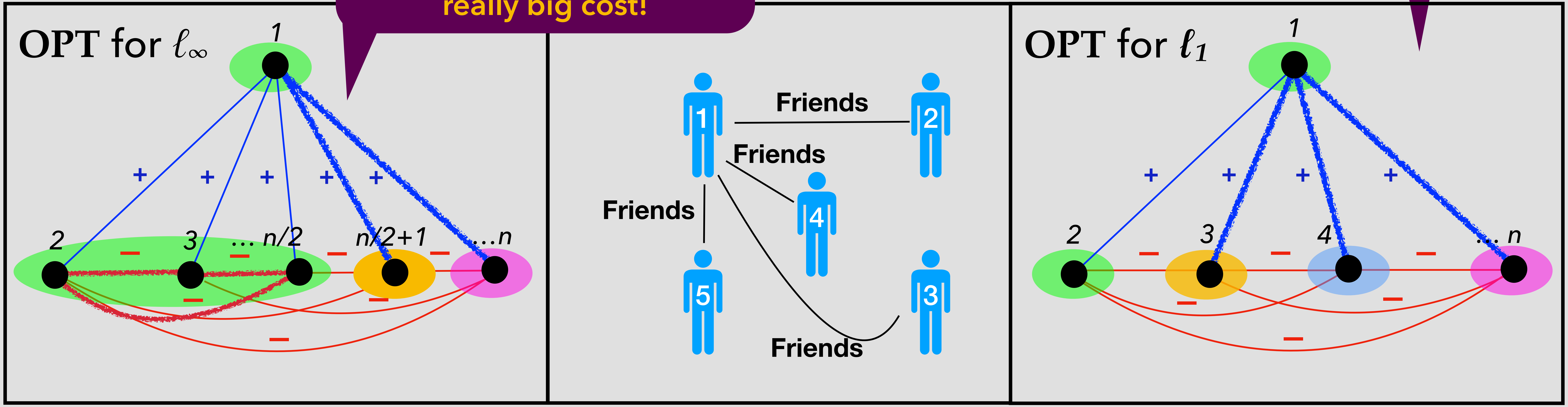
Trouble with the convex program

OPT for one ℓ_p-norm can be really bad for others!

Does there exist a "pretty good" solution for all ℓ_p-norms?

Cost for ℓ_1 norm is $\theta(n^2)$, **really big cost!**

Cost for ℓ_1 norm is $\theta(n)$



Trouble with the convex program

ℓ_p-norm correlation clustering algs solve a convex program

Solving *metric constrained* LPs on large networks is slow!

Not very amenable to online/ streaming settings

Solution specific to *one fixed* ℓ_p-norm

Our combinatorial approach

Initial constant was 40
Heidrich, Irmai, Andres built off us, improve to 4!

"Fast Combinatorial Algorithms for Min Max Correlation Clustering"
ICML23

(1) Develop faster $O(1)$-apx alg for min max objective;
↪ near-linear time on networks with small positive degree

"One Partition Approximating All ℓ_p-norm Objectives in Correlation Clustering"
In sub

(2) Find ***simultaneously*** $O(1)$-apx clustering for all ℓ_p-norm objs

Not possible for *k*-center & *k*-median [Alamdari & Shmoys WAOA17]

(3) Algorithms in the online setting

In progress

Today

- ✦ Introduction (the model, prior work, our results) 🔥
- ✦ The correlation metric (constructing a "guess" for the fraction solution, an inherent asymmetry) 🔥 🔥
- ✦ Proof sketch for the ℓ_∞-norm 🔥 🔥 🔥
- ✦ Adjusting the correlation metric (regular graphs are easy, dealing with negative edges) 🔥🔥
- ✦ Conclusions (mainly vibes) 🔥

Today

- ~~Introduction (the model, prior work, our results)~~ 🔥
- The correlation metric (constructing a "guess" for the fraction solution, an inherent asymmetry) 🔥 🔥
- Proof sketch for the ℓ_∞-norm 🔥 🔥 🔥
- Adjusting the correlation metric (regular graphs are easy, dealing with negative edges) 🔥🔥
- Conclusions (mainly vibes) 🔥

Previous techniques

Integer convex program

Not (generally) practical to solve

$x_{uv} = 0$ then u, v same cluster
$x_{uv} = 1$ then u, v different clusters

$$\min ||y||_p$$

$$y(u) = \sum_{v \in N_u^+} x_{uv} + \sum_{v \in N_u^-} (1 - x_{uv}) \qquad \forall u \in V$$

$$x_{uv} \leq x_{vw} + x_{uw} \qquad \forall u, v, w \in V$$

$$x_{uv} \in \mathbb{Z}_{\geq 0} \qquad \forall u, v \in V$$

Convex program relaxation

Can be solved efficiently

$$\min ||y||_p$$

$$y(u) = \sum_{v \in N_u^+} x_{uv} + \sum_{v \in N_u^-} (1 - x_{uv}) \qquad \forall u \in V$$

$$x_{uv} \leq x_{vw} + x_{uw} \qquad \forall u, v, w \in V$$

$$0 \leq x_{uv} \leq 1 \qquad \forall u, v \in V$$

Previous techniques

Integer convex program

Not (generally) practical to solve

$x_{uv} = 0$ then u, v same cluster

$x_{uv} = 1$ then u, v c

m

$y(u) = \sum_{v \in N_u^+} x_{uv} \ldots \sum_{v \in N_u^-} \ldots$

$x_{uv} \le x_{vw} + x_{uw} \qquad \forall u, v, w \in V$

$x_{uv} \in \mathbb{Z}_{\ge 0} \qquad \forall u, v \in V$

Step 1: Solve convex program

Step 2: "Round" fractional solution to integral one

Convex program relaxation

Can be solved efficiently

$$\min ||y||_p$$

$$y(u) = \sum_{v \in N_u^+} x_{uv} + \sum_{v \in N_u^-} (1 - x_{uv}) \qquad \forall u \in V$$

$$x_{uv} \le x_{vw} + x_{uw} \qquad \forall u, v, w \in V$$

$$0 \le x_{uv} \le 1 \qquad \forall u, v \in V$$

Previous techniques

Input: semi-metric x on V

Let $r = 1/5$

While there is some unclustered vertex

  Find "densest" cluster with center c^* and radius r

  Create cluster C around c^* with radius $2r$

Return clusters

Convex program for ℓ_p correlation clustering

$$\min ||y||_p$$

$$y(u) = \sum_{v \in N_u^+} x_{uv} + \sum_{v \in N_u^-} (1 - x_{uv}) \qquad \forall u \in V$$

$$x_{uv} \leq x_{vw} + x_{uw} \qquad \forall u, v, w \in V$$

$$0 \leq x_{uv} \leq 1 \qquad \forall u, v \in V$$

Constraints induce a semi-metric space

LP solution → Rounding algorithm by Kalhan, Makarychev, Zhou → **Clustering**

Correlation metric

Correlation metric for ℓ_∞:

(1) satisfies triangle inequality

(2) has $\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \le O(1) \cdot \max_{w \in V} y(w)$

$$\sum_{u \in V} \left(\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \right)^p \le O(1) \cdot \sum_{w \in V} y(w)^p$$

$$\min ||y||_p$$

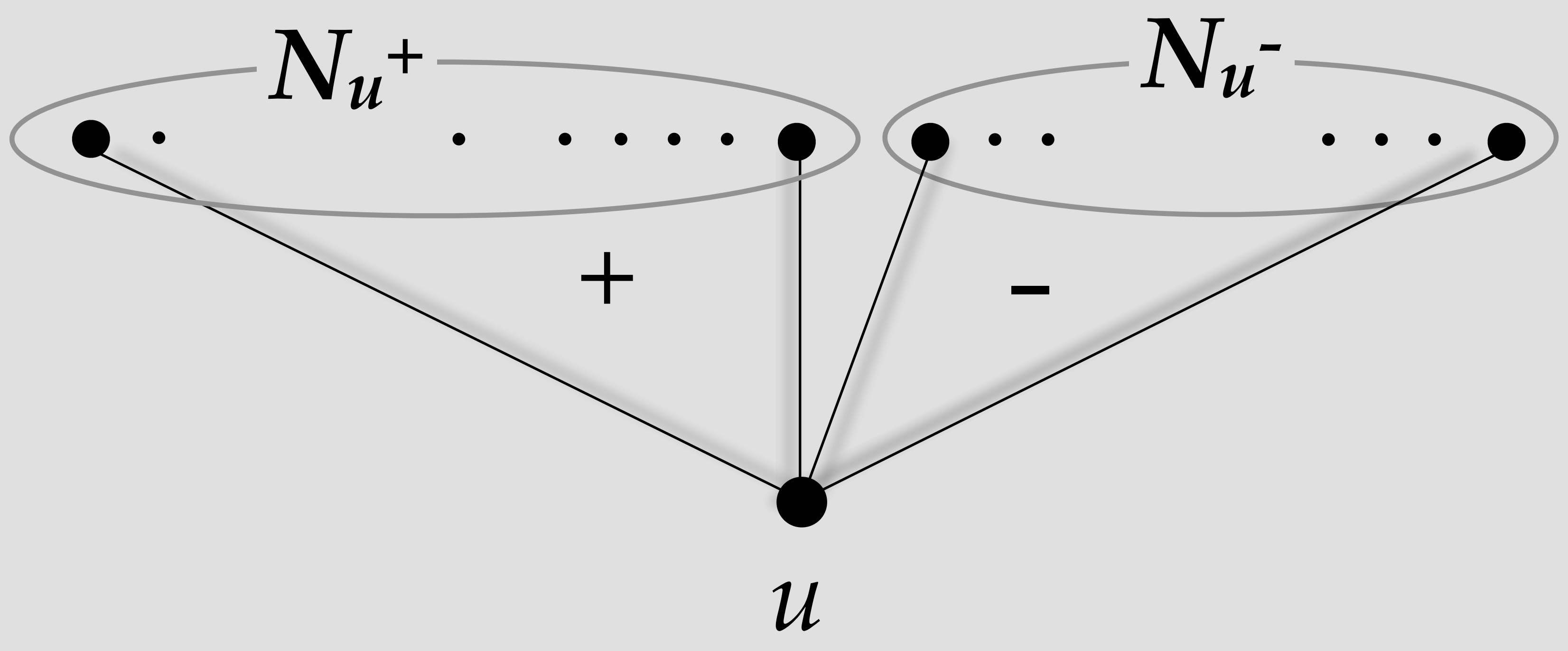
$$y(u) = \sum_{v \in N_u^+} x_{uv} + \sum_{v \in N_u^-} (1 - x_{uv}) \qquad \forall u \in V$$

$$x_{uv} \le x_{vw} + x_{uw} \qquad \forall u, v, w \in V$$

$$0 \le x_{uv} \le 1 \qquad \forall u, v \in V$$

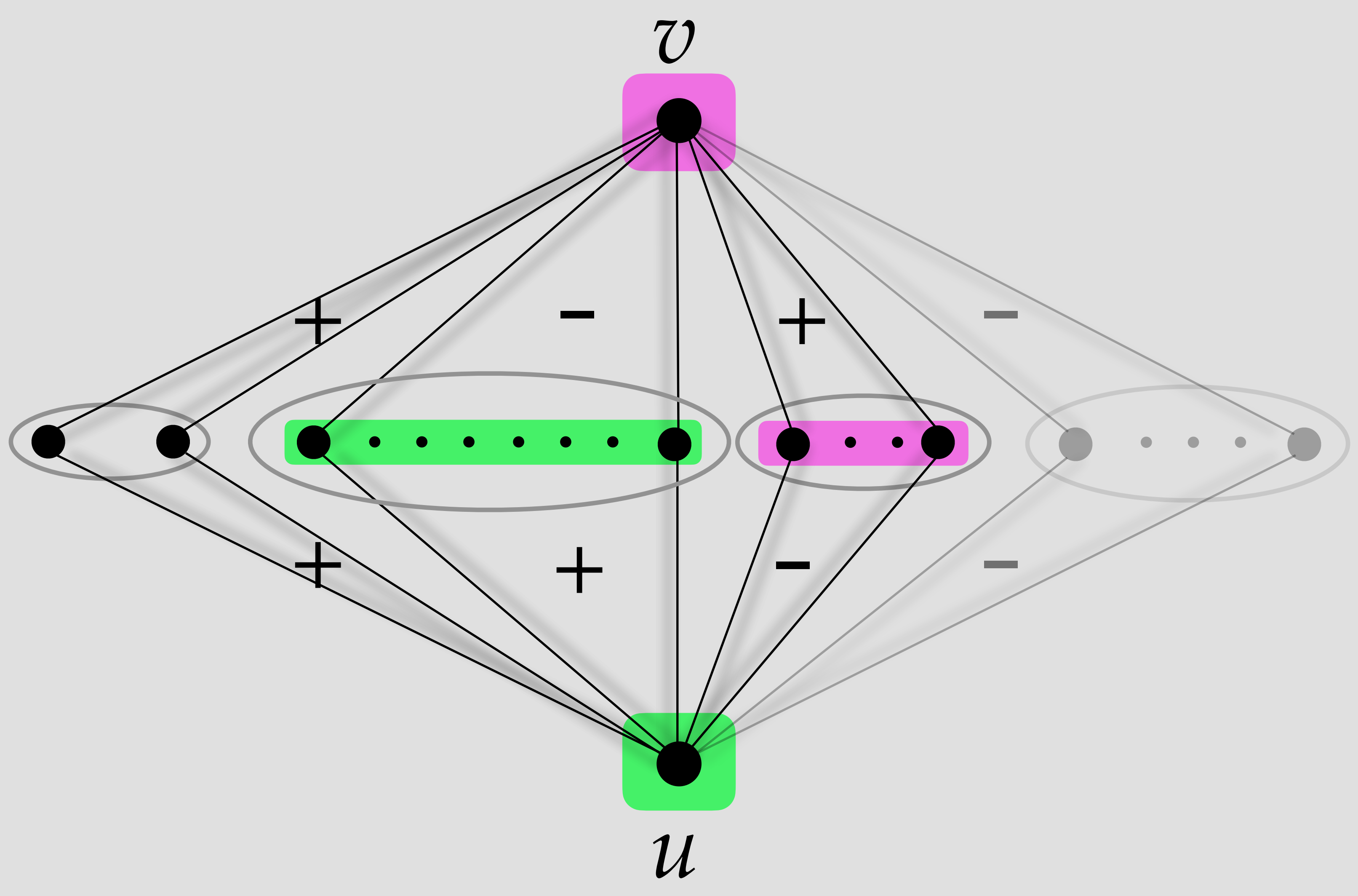
Correlation metric

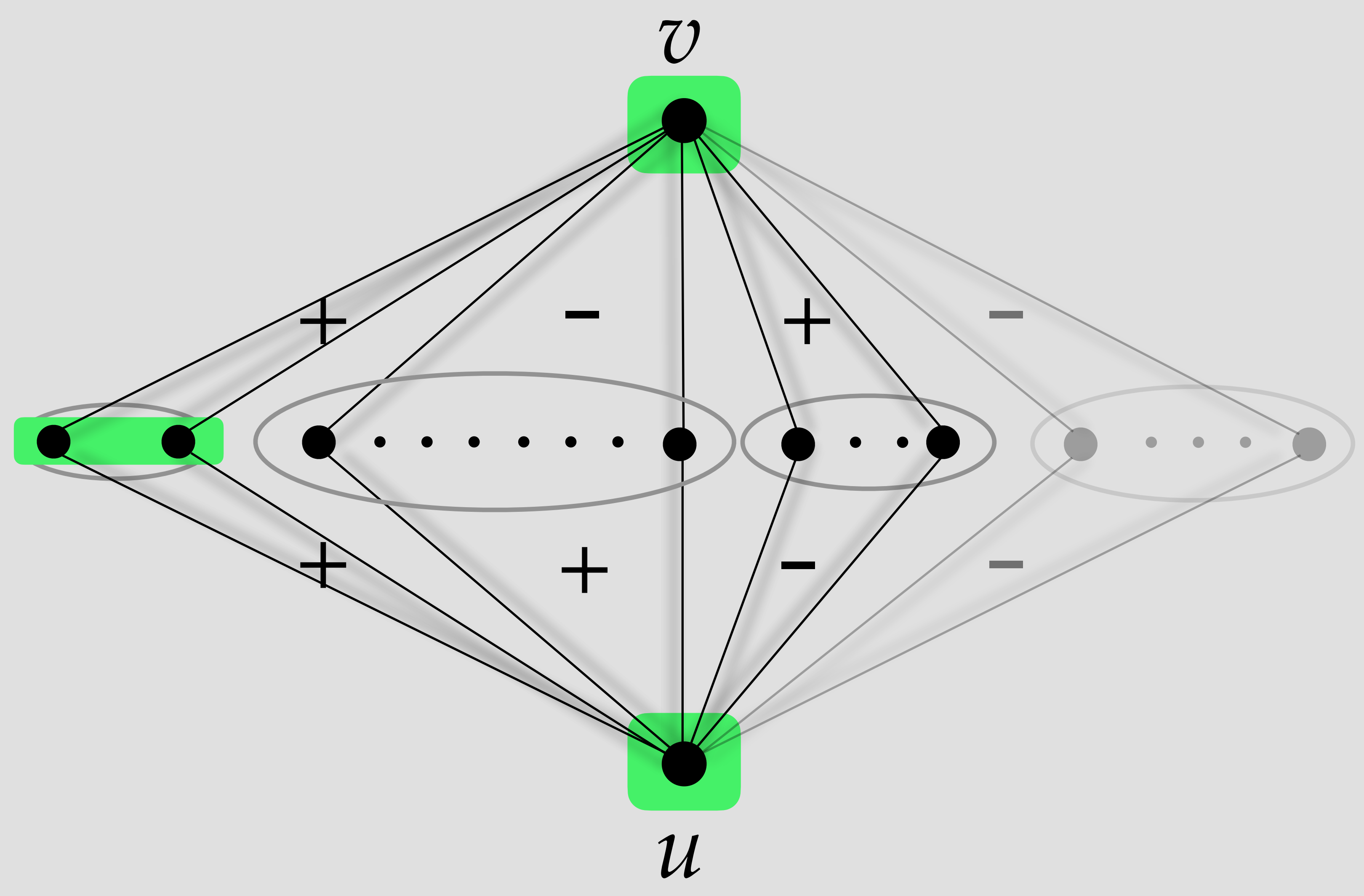
- N_u^+ = (+) neighbors of u, N_u^- = (–) neighbors of u



Correlation metric

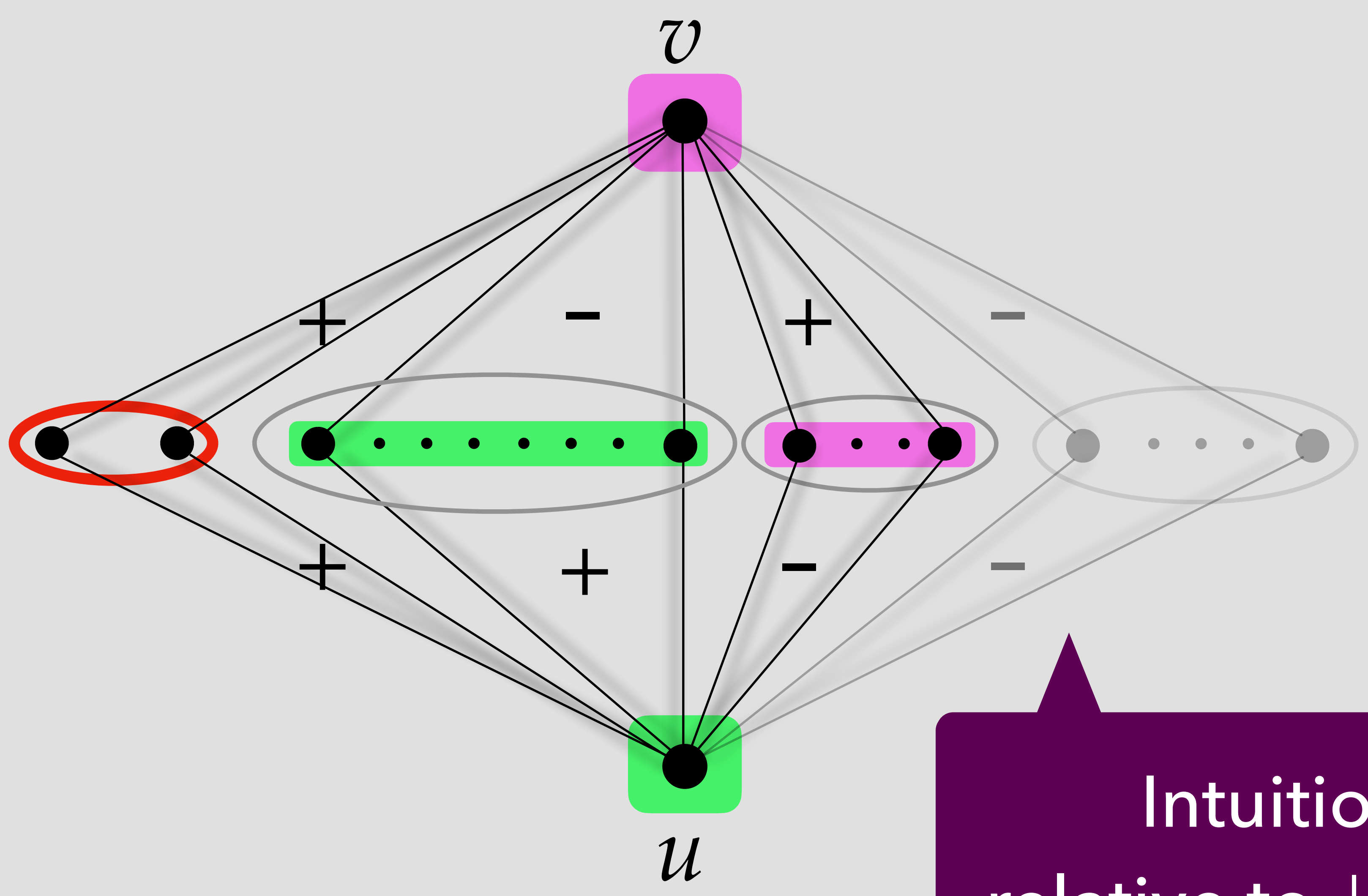
- N_u^+ = (+) neighbors of u, N_u^- = (–) neighbors of u

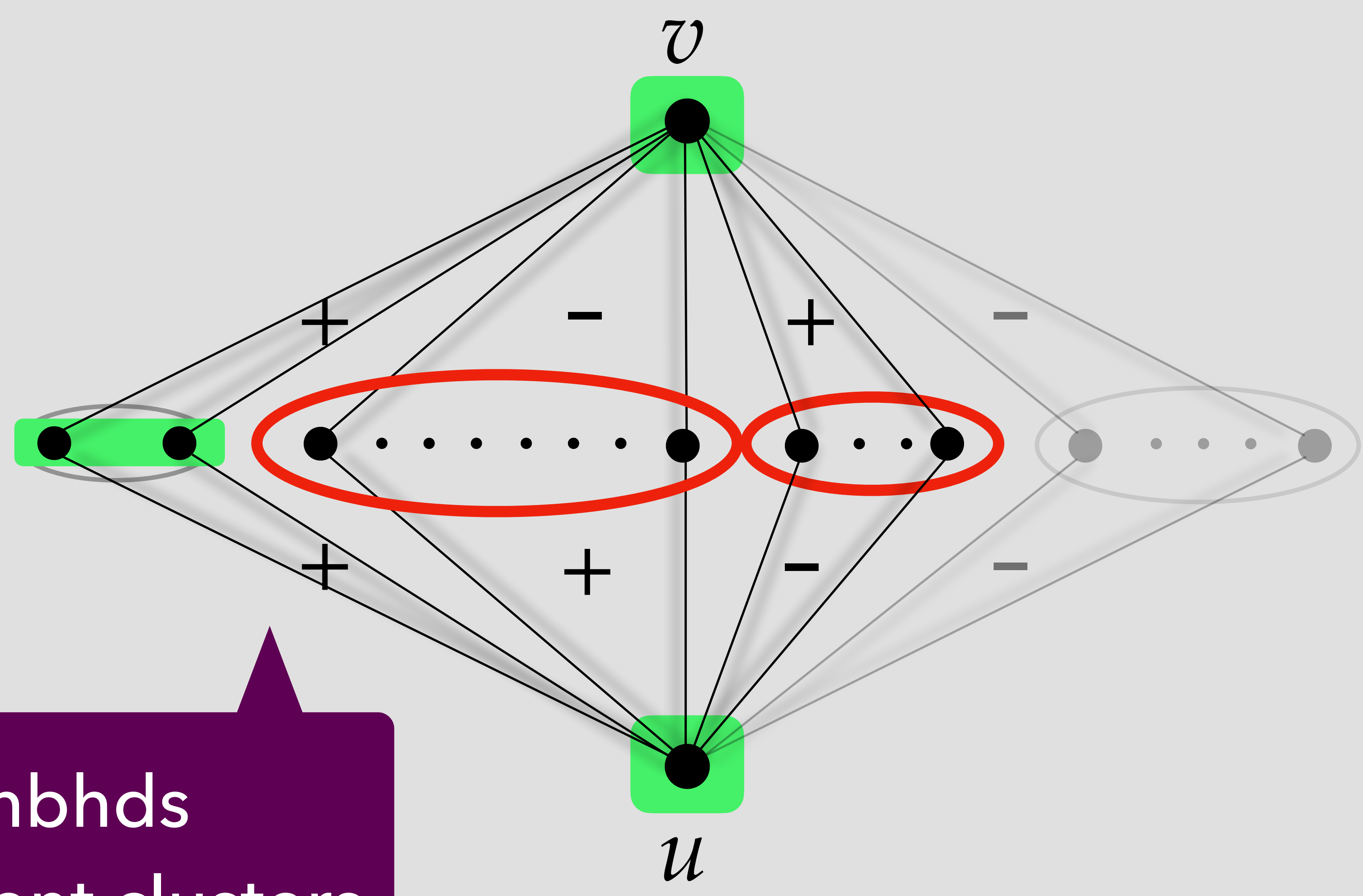




Correlation metric

- N_u^+ = (+) neighbors of u, $N_u^- =$ (−) neighbors of u



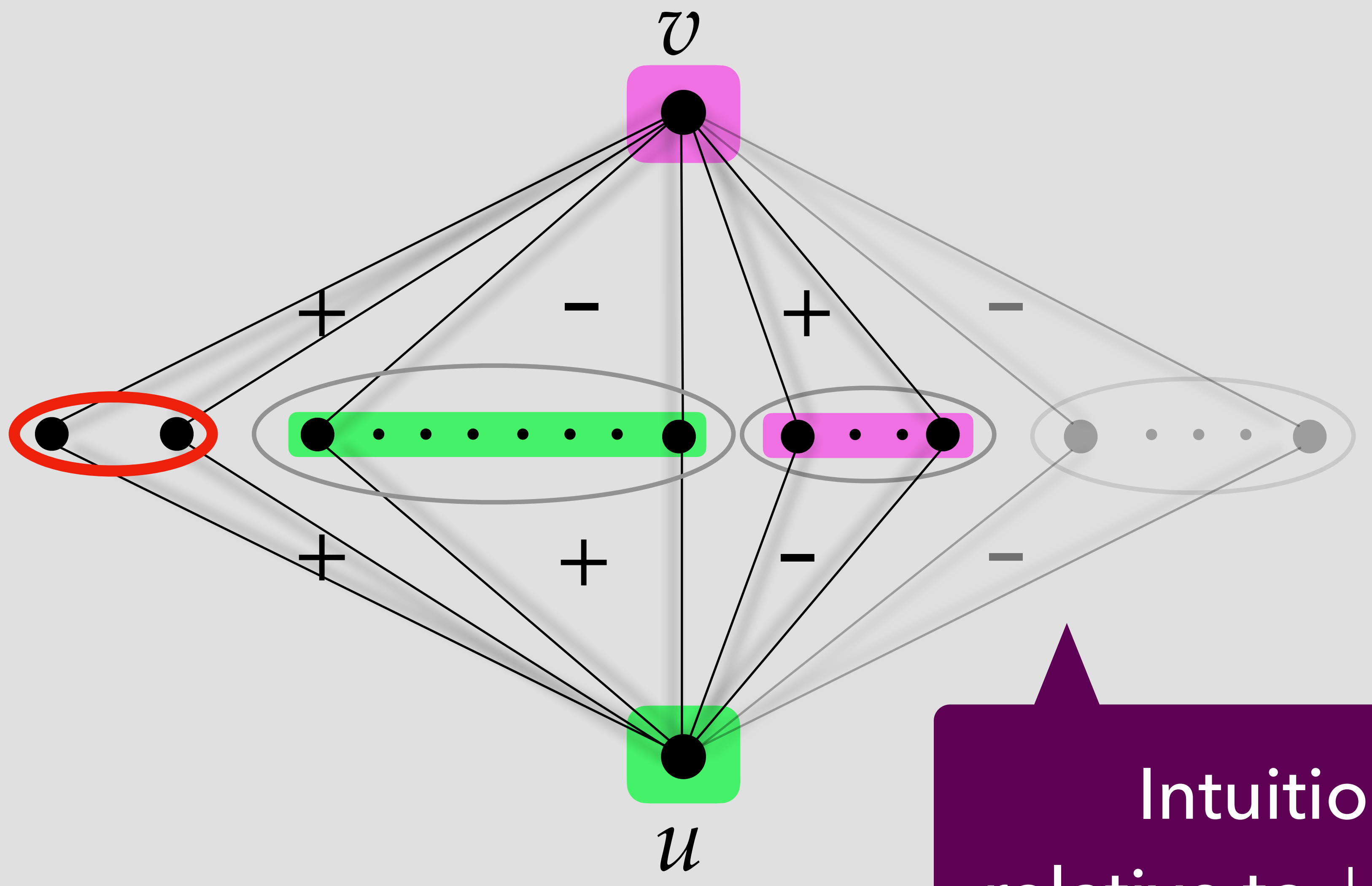


Intuition: if u and v have large mixed nbhds relative to $|N_u^+ \cup N_v^+|$, want them in different clusters

Correlation metric

Very coarse approximation for probability
Pivot separates u,v

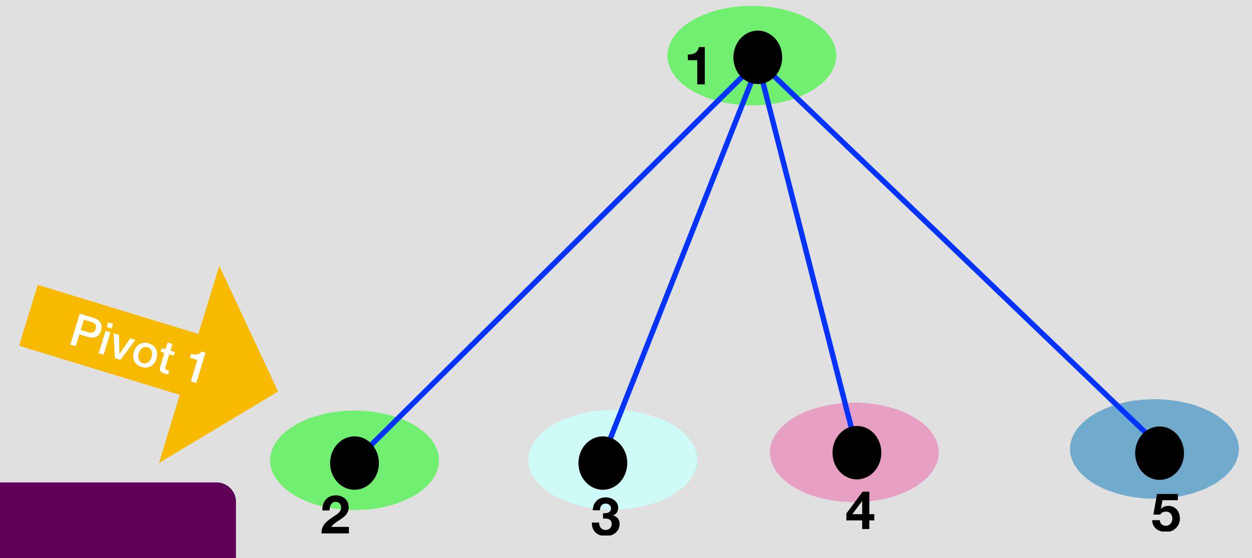
$$\text{Correlation metric} = d_{uv} = 1 - \frac{|N_u^+ \cap N_v^+|}{|N_u^+ \cup N_v^+|} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$



Intuition: if u and v have large mixed nbhds relative to $|N_u^+ \cup N_v^+|$, want them in different clusters

Pivot algorithm

- Randomly choose a *pivot* (unclustered vertex)
- New cluster with pivot + all its unclustered + neighbors



Correlation metric

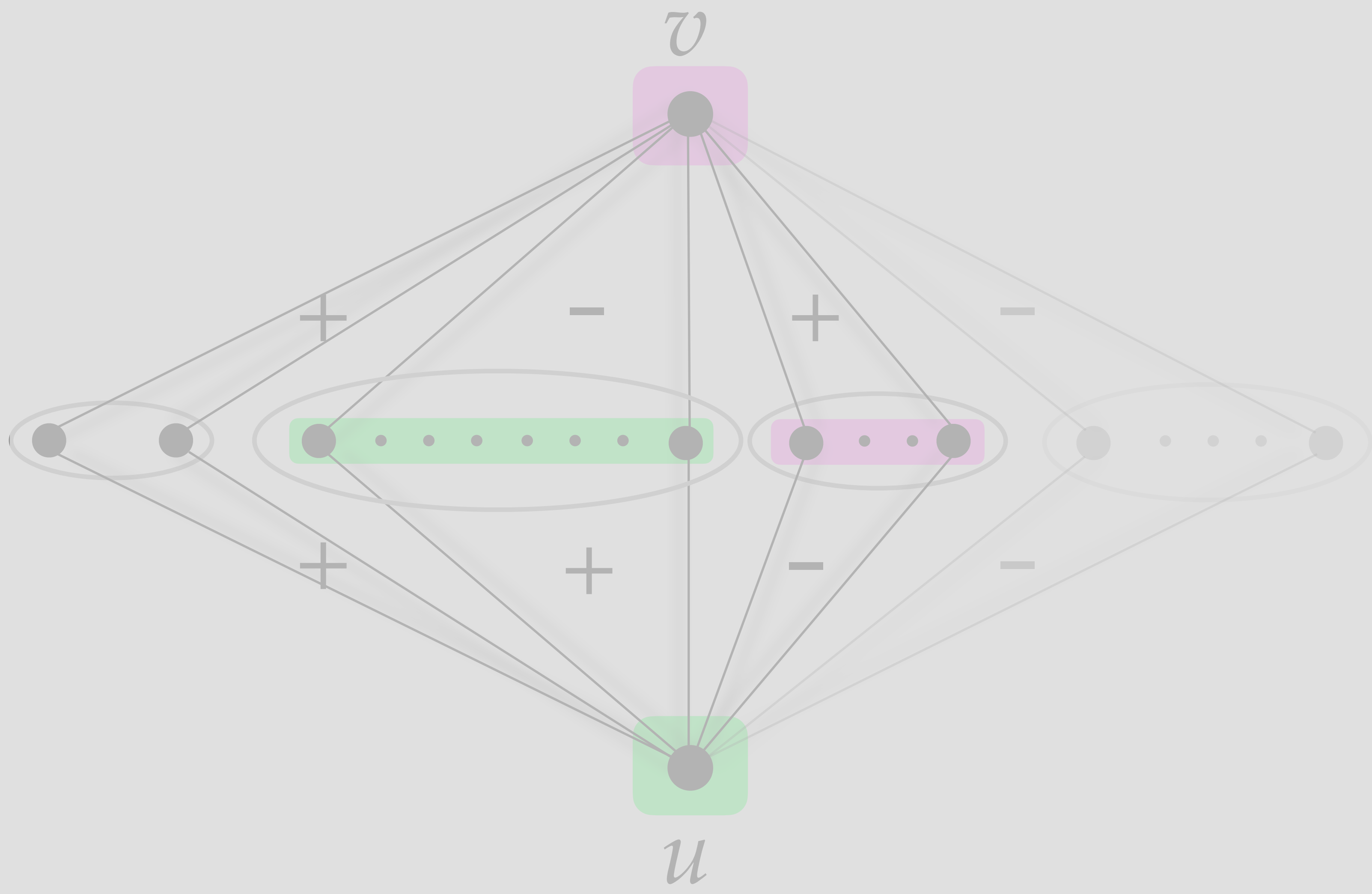
Correlation metric fast to compute, time $O(n^\omega)$.

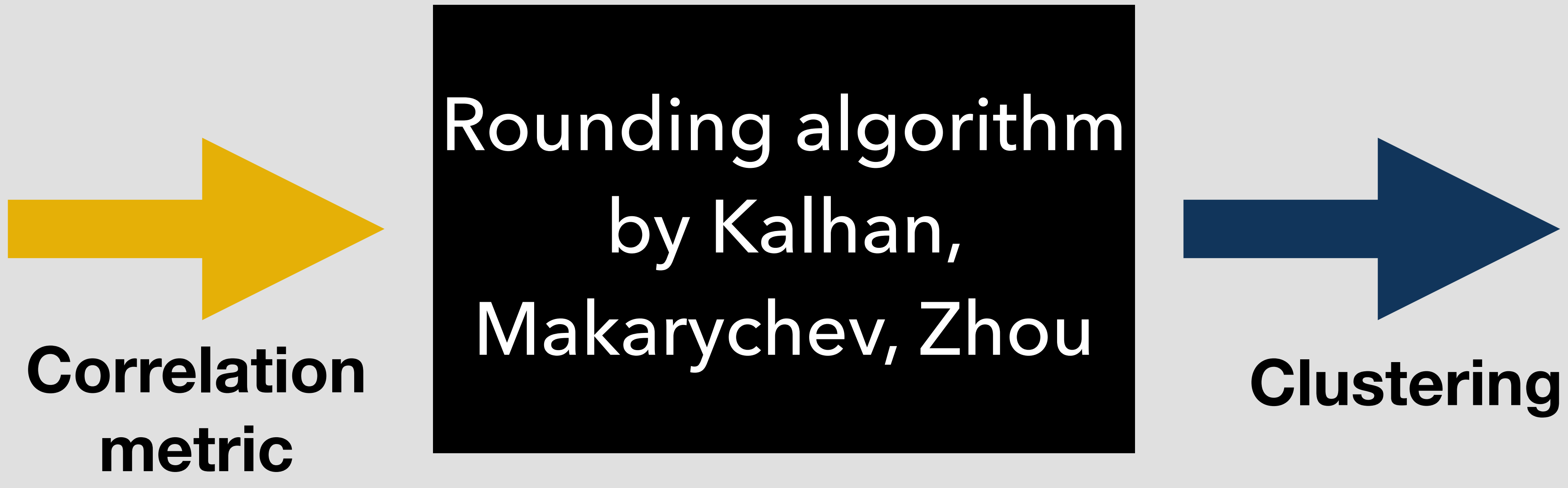
↪Even faster on sparse graphs $O(n \cdot \Delta^2 \cdot \log n)$

↪Further sped up on any graph with sampling procedure

Works as is for ℓ_∞ norm objective

Correlation metric = $d_{uv} = 1 - \frac{|N_u^+ \cap N_v^+|}{|N_u^+ \cup N_v^+|} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$





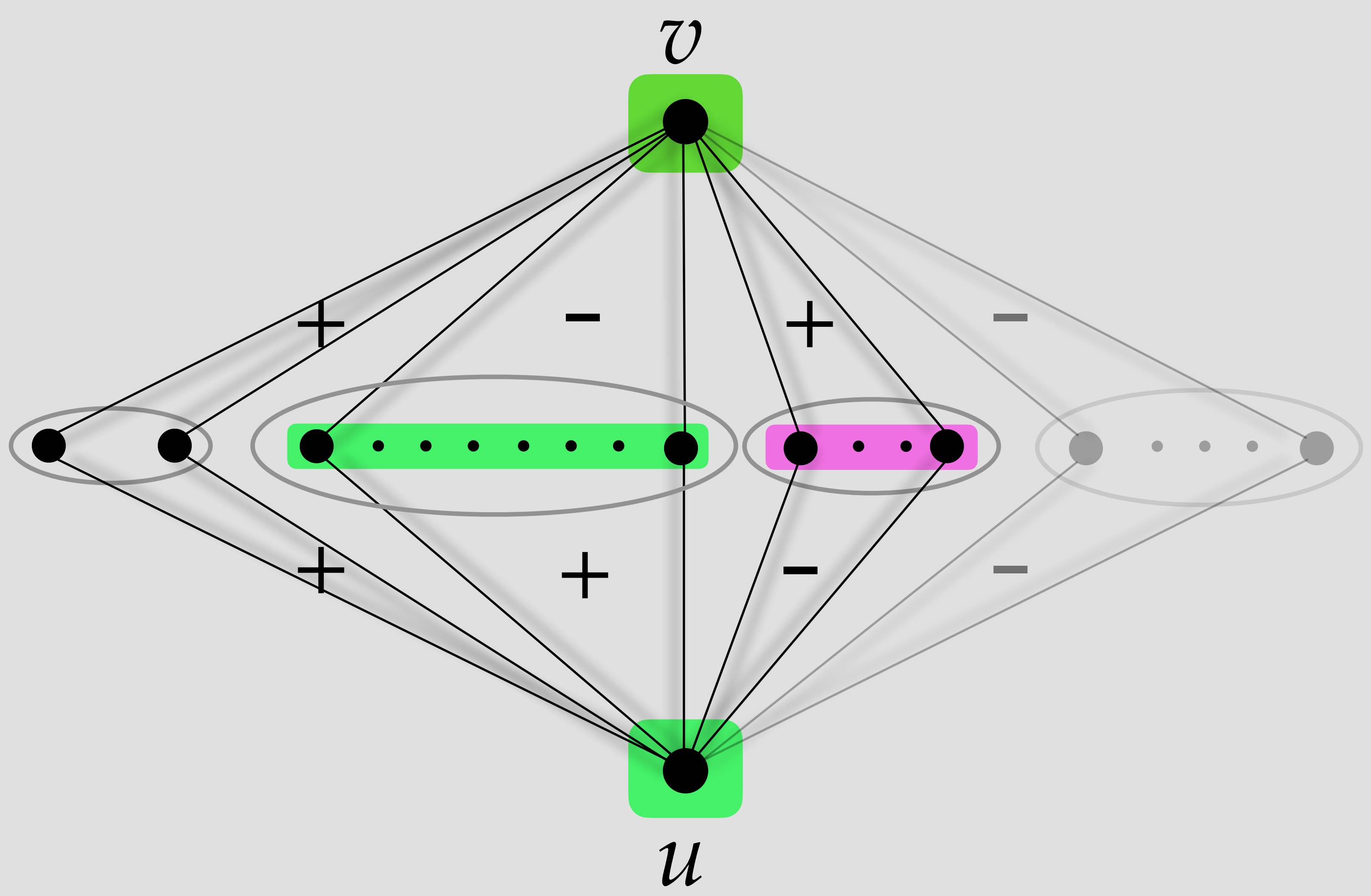
Today

- ~~Introduction (the model, prior work, our results)~~ 🔥
- ~~The correlation metric (constructing a "guess" for the fraction solution, an inherent asymmetry)~~ 🔥 🔥
- Proof sketch for the ℓ_∞-norm 🔥 🔥 🔥
- Adjusting the correlation metric (regular graphs are easy, dealing with negative edges) 🔥🔥
- Conclusions (mainly vibes) 🔥

Correlation metric for ℓ_∞

$$d_{uv} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$

Want to show for ℓ_∞: $\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \leq O(1) \cdot \max_{w \in V} y(w)$



Easy to bound positive edges for ℓ_∞ objective!

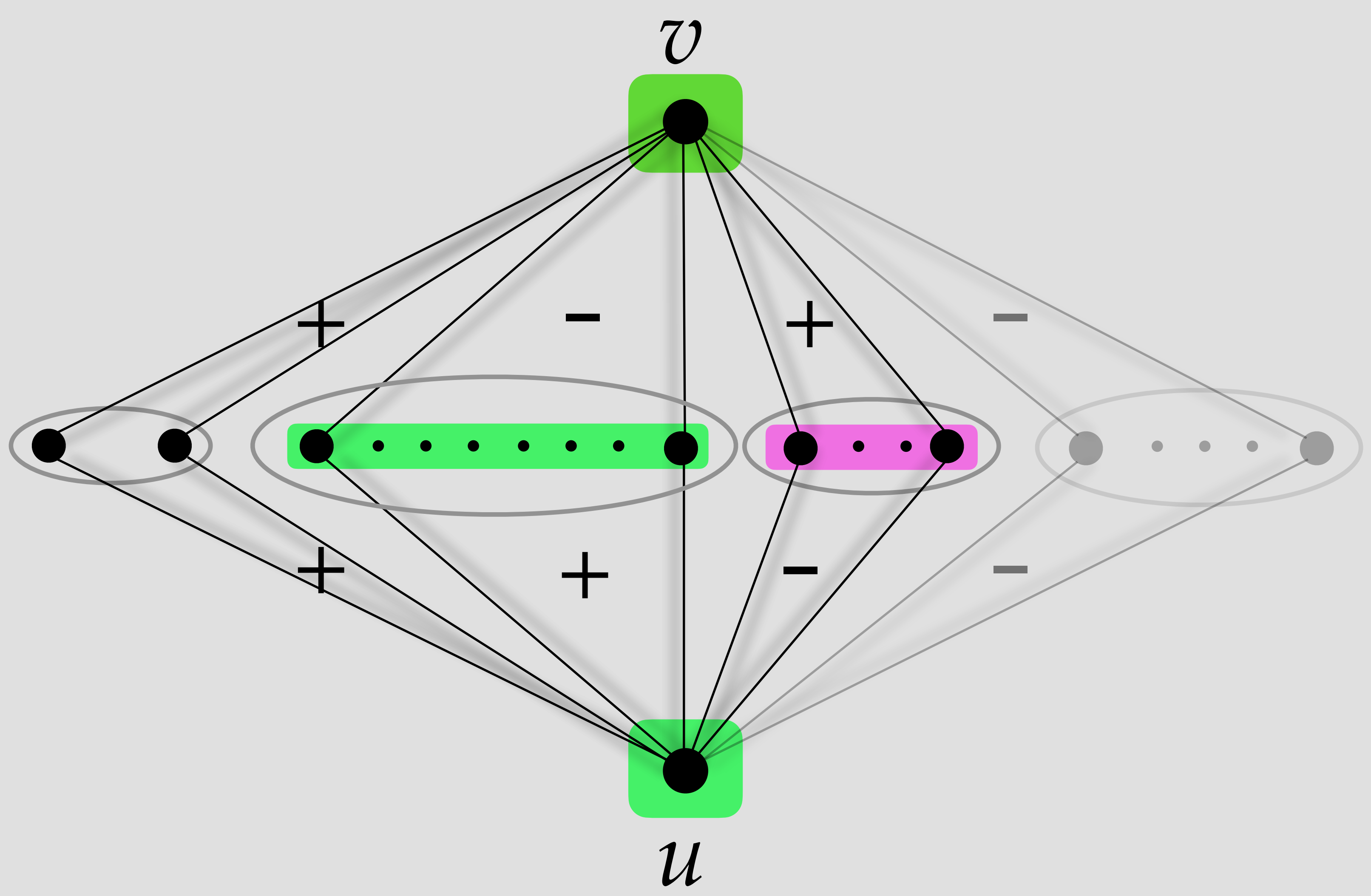
$$\sum_{v \in N_u^+} d_{uv} \leq \sum_{v \in N_u^+ \cap C(u)} \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|} + \sum_{v \in N_u^+ \cap \overline{C(u)}} 1$$

Correlation metric for ℓ_∞

$$d_{uv} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$

Want to show for ℓ_∞: $\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \leq O(1) \cdot \max_{w \in V} y(w)$

Recall $y(u)$ = # disagreements incident to u



Easy to bound positive edges for ℓ_∞ objective!

$$\sum_{v \in N_u^+} d_{uv} \leq \sum_{v \in N_u^+ \cap C(u)} \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|} + \sum_{v \in N_u^+ \cap \overline{C(u)}} 1$$

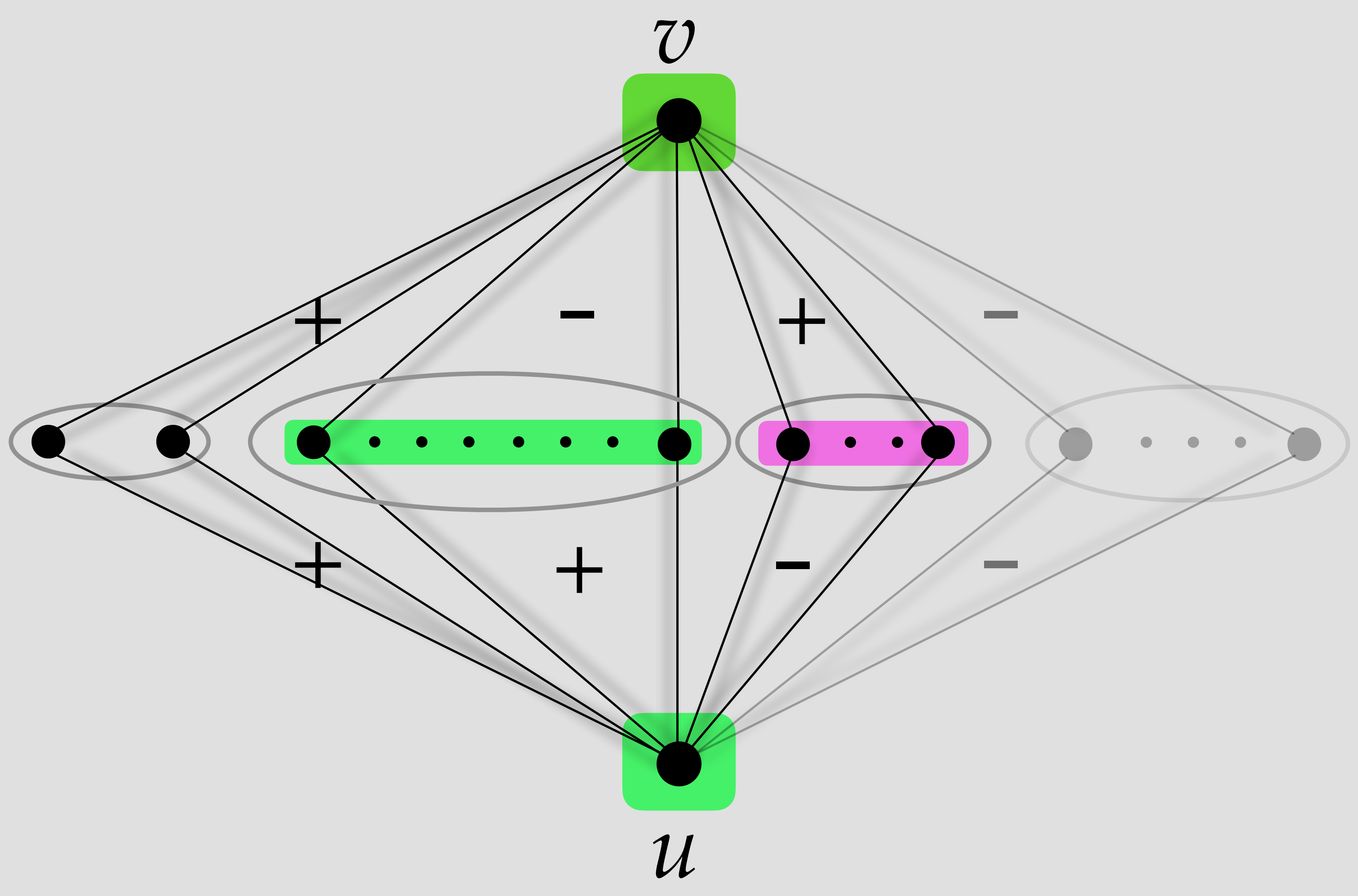
$$\leq \frac{1}{|N_u^+|} \sum_{v \in N_u^+ \cap C(u)} \left(|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|\right) + y(u)$$

Correlation metric for ℓ_∞

$$d_{uv} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$

Want to show for ℓ_∞: $\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \leq O(1) \cdot \max_{w \in V} y(w)$

Recall $y(u)$ = # disagreements incident to u



Easy to bound positive edges for ℓ_∞ objective!

$$\sum_{v \in N_u^+} d_{uv} \leq \sum_{v \in N_u^+ \cap C(u)} \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|} + \sum_{v \in N_u^+ \cap \overline{C(u)}} 1$$

$$\leq \frac{1}{|N_u^+|} \sum_{v \in N_u^+ \cap C(u)} (|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|) + y(u)$$

$$\leq \frac{1}{|N_u^+|} \sum_{v \in N_u^+ \cap C(u)} (y(u) + y(v)) + y(u)$$

$$\leq 2y(u) + \max_z y(z) \leq 3 \cdot \mathrm{OPT}\,.$$

Correlation metric for ℓ_∞

$$d_{uv} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$

Want to show for ℓ_∞: $\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \leq O(1) \cdot \max_{w \in V} y(w)$

Easy to bound positive edges for ℓ_∞ objective!

$$\sum_{v \in N_u^+} d_{uv} \leq 3 \cdot \mathrm{OPT}\,.$$

Correlation metric for ℓ_∞

$$d_{uv} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$

Want to show for ℓ_∞: $\sum_{v \in N_u^+} d_{uv} + \sum_{v \in N_u^-} (1 - d_{uv}) \leq O(1) \cdot \max_{w \in V} y(w)$

Bound on negative edges

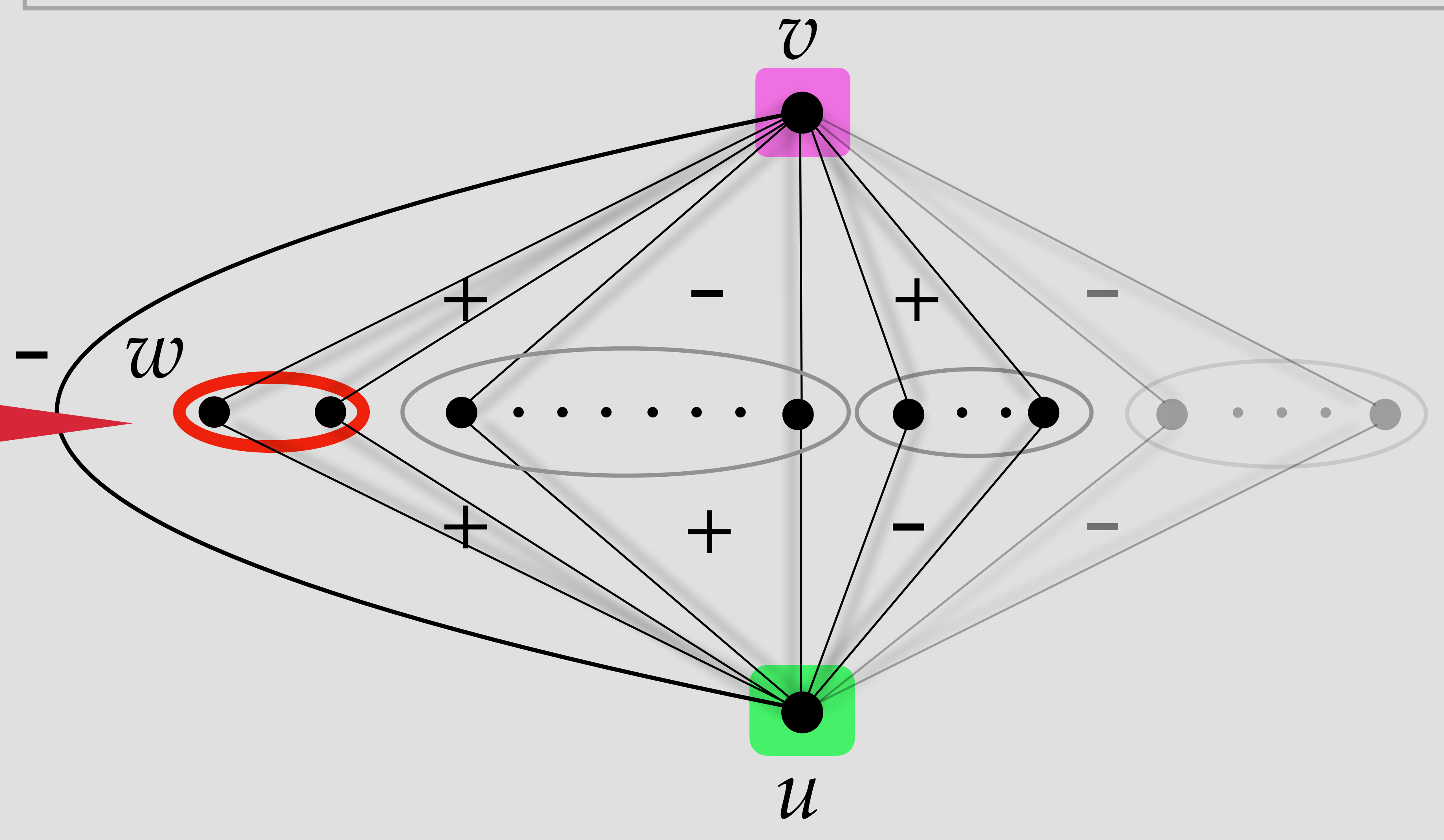
$$\sum_{v \in N_u^-} (1 - d_{uv}) = \sum_{v \in N_u^- \cap C(u)} (1 - d_{uv}) + \sum_{v \in N_u^- \cap \overline{C(u)}} (1 - d_{uv})$$

$$= y(u) + \sum_{v \in N_u^- \cap \overline{C(u)}} \frac{|N_u^+ \cap N_v^+|}{n - |N_u^- \cap N_v^-|}$$

Every w in $|N_u^+ \cap N_v^+|$ incident to an edge in disagreement, charge to carefully chosen $v^*(w)$ in $C(w)$

Easy to bound positive edges for ℓ_∞ objective!

$$\sum_{v \in N_u^+} d_{uv} \leq 3 \cdot \mathrm{OPT}\,.$$



Today

- ~~Introduction (the model, prior work, our results)~~ 🔥
- ~~The correlation metric (constructing a "guess" for the fraction solution, an inherent asymmetry)~~ 🔥 🔥
- ~~Proof sketch for the ℓ_∞-norm~~ 🔥 🔥 🔥
- Adjusting the correlation metric (regular graphs are easy, dealing with negative edges) 🔥🔥
- Conclusions (mainly vibes) 🔥

Adjusted correlation metric

$$d_{uv} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$$

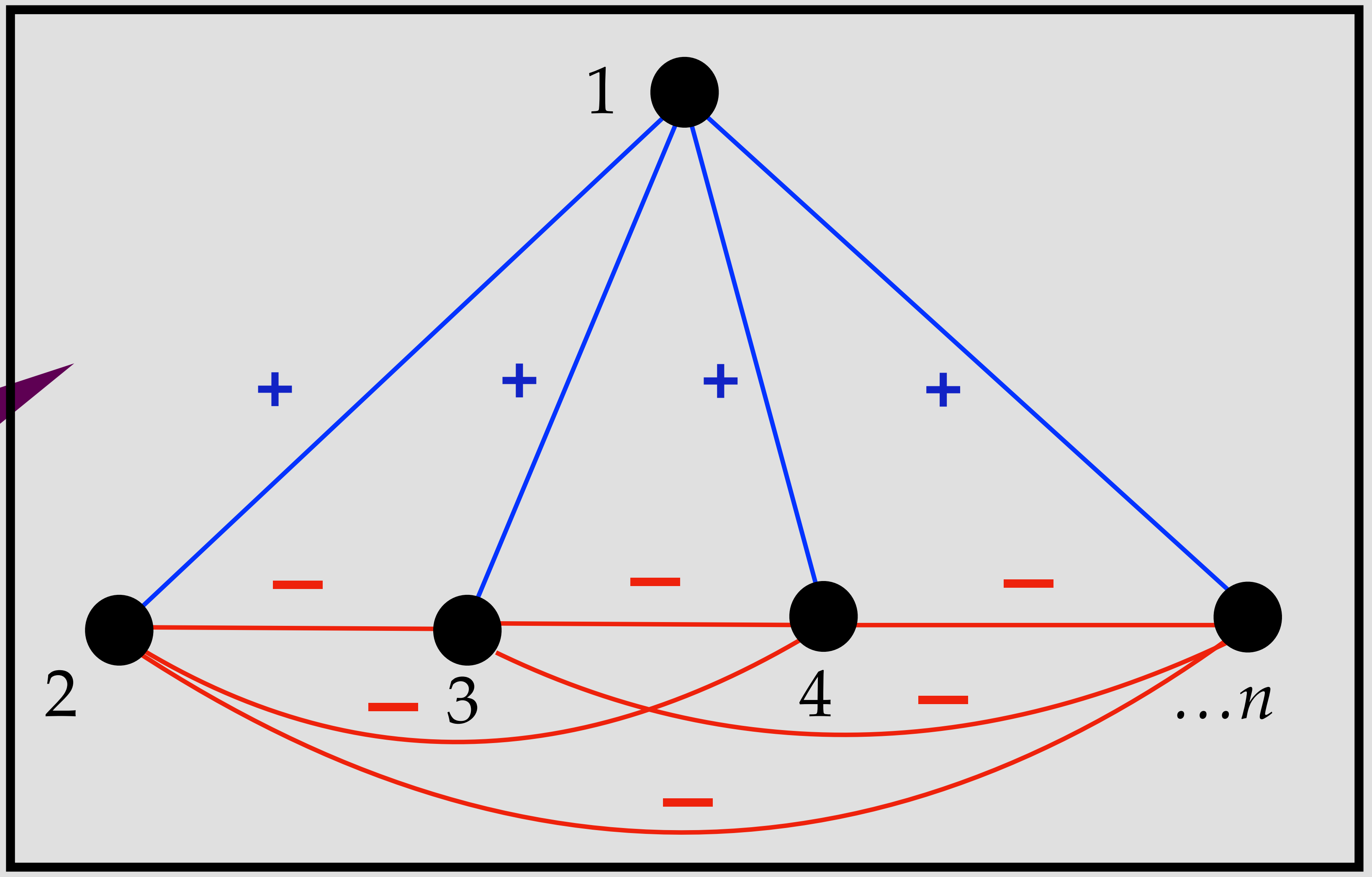
→ Simultaneous approximation for ℓ_1- and ℓ_∞-norm objectives

For ***regular graphs***, correlation metric $O(1)$–apxs ℓ_1-norm

- ✦ Proof via dual fitting!
- ✦ Problem is when graph is far from regular

Must ***adjust*** correlation metric for non-regular graphs for general ℓ_p-norms

$d_{uv} = 2/3$ for all u,v in $\{2,\ldots,n\}$, so fractional cost w.r.t d is $\Theta(n^2)$



Adjusted correlation metric

Correlation metric = $d_{uv} = 1 - \frac{|N_u^+ \cap N_v^+|}{|N_u^+ \cup N_v^+|} = \frac{|N_u^+ \cap N_v^-| + |N_u^- \cap N_v^+|}{|N_u^+ \cup N_v^+|}$

For *all* ℓ_p norms

- If negative edge (u,v) has $d_{uv} > 0.7$, update $d_{uv} \leftarrow 1$
- For u with $|N_u^- \cap \{v : d_{uv} \leq 0.7\}| \geq \frac{10}{3}\Delta_u$, update $d_{uv} \leftarrow 1$

Adjusted correlation metric

Rounding algorithm by Kalhan, Makarychev, Zhou

Clustering

Today

- ~~Introduction (the model, prior work, our results)~~ 🔥
- ~~The correlation metric (constructing a "guess" for the fraction solution, an inherent asymmetry)~~ 🔥 🔥
- ~~Proof sketch for the ℓ_∞-norm~~ 🔥 🔥 🔥
- ~~Adjusting the correlation metric (regular graphs are easy, dealing with negative edges)~~ 🔥🔥
- Conclusions (mainly vibes) 🔥

Summary

ℓ_p-norm correlation clustering algs solve a convex program

- Solving *metric constrained* LPs on large networks is slow!
- Not very amenable to online settings
- Solution specific to *one fixed* ℓ_p-norm

Summary

ℓ_p-norm correlation clustering algs solve a convex program

Solving *metric constrained* LPs on large networks is slow!

Not very amenable to online settings

Solution specific to *one fixed* ℓ_p-norm

Combinatorial techniques can resolve these issues

Summary

Δ = max (+) degree of any vertex
ω = matrix multiplication exponent

Sometimes called *universality* property

Result 1: $O(1)$-apx alg with run-time $O(\min\{ n\cdot\Delta^2\cdot\log n , n^\omega\})$. Near-linear for sparse graphs.

Result 2: ∃ an alg producing a clustering that is $O(1)$-apx for all ℓ_p-norms, simultaneously.

Result 3: (In progress, probably true) Given a random ε–fraction of the network, ∃ a ***semi-online algorithm*** that for any ℓ_p-norm objective produces a $O(\log n)$-competitive algorithm.

Correlation clustering has interesting combinatorial structure that can be exploited

What's next?

- 👩‍💻**In progress:** Extend to a semi-online setting

  ↪ Factor depends on p. For $p=\infty$, the algorithm is $\theta(\log n)$-competitive

- 🔥**Hot conjecture:** Exists a combinatorial alg simultaneously 4-approximating all ℓ_p-norms running in $O(n^\omega)$ time

- 😶‍🌫️ **Broader Qs:**

  1. Combinatorial algorithms by designing "approximate LP solution"
  2. Further study on the all-norms objective

Thank you!

samidavies@berkeley.edu