Clustering with Diversity

Jian Li

University of Maryland, College Park

Ke Yi and Qin Zhang

Hong Kong University of Science & Technology

ICALP 2010

July 2010

Clustering - overview

- A general clustering problem:

  Given a set V of points in a metric space, partition V into a set of disjoint clusters such that a certain objective function is minimized,

  subject to some

Clustering - overview

- A general clustering problem:

  Given a set V of points in a metric space, partition V into a set of disjoint clusters such that a certain objective function is minimized,

  subject to some

  - cluster-level constraints: impose restrictions, for example, on the number of clusters (exactly k clusters) or on the size of each cluster (at least l elements).

Clustering - overview

- A general clustering problem:

  Given a set V of points in a metric space, partition V into a set of disjoint clusters such that a certain objective function is minimized,

  subject to some

  - cluster-level constraints: impose restrictions, for example, on the number of clusters (exactly k clusters) or on the size of each cluster (at least l elements).

  - instance-level constraints: specify whether particular pair of items can be clustered together based on some background knowledge.

Clustering with diversity

- ℓ-diversity: Given a set V of points in a metric space, each of which has a color, subject to:
  - each cluster has size $\geq l$.
  - all points partitioned into one cluster must have distinct colors;

Goal (this paper): minimize $\max\{$diameter of clusters$\}$.

Clustering with diversity

- ℓ-diversity: Given a set V of points in a metric space, each of which has a color, subject to:
  - each cluster has size $\geq l$.
  - ~~all points partitioned into one cluster must have distinct colors;~~ $\Rightarrow$ ℓ-anonymity

Goal (this paper): minimize $\max\{$diameter of clusters$\}$.

The main motivation

ℓ-diversity is motivated from privacy preservation for data publication (Machanavajjhala et. al. 2006), follows ℓ-anonymity (a.k.a. k-anonymity, Samarati 2001).

quasi-identifier (QI)

sensitive attribute

T-1D(*Name*)	Age	Gender	Degree	Disease
1 (*Adam*)	29	M	M.Sc.	HIV
2 (*Bob*)	25	M	M.Sc.	HIV
3 (*Calvin*)	25	M	B.Sc.	pneumonia
4 (*Daisy*)	29	F	B.Sc.	bronchitis
5 (*Elam*)	40	M	B.Sc.	bronchitis
6 (*Frank*)	45	M	B.Sc.	bronchitis
7 (*George*)	35	M	B.Sc.	pneumonia
8 (*Henry*)	37	M	B.Sc.	pneumonia
9 (*Ivy*)	50	F	Ph.D.	dyspepsia
10 (*Jane*)	60	F	Ph.D.	pneumonia

(a) The microdata

The main motivation

ℓ-diversity is motivated from privacy preservation for data publication (Machanavajjhala et. al. 2006), follows ℓ-anonymity (a.k.a. k-anonymity, Samarati 2001).

quasi-identifier (QI)

sensitive attribute

T-1D(*Name*)	Age	Gender	Degree	Disease
1 (*Adam*)	29	M	M.Sc.	HIV
2 (*Bob*)	25	M	M.Sc.	HIV
3 (*Calvin*)	25	M	B.Sc.	pneumonia
4 (*Daisy*)	29	F	B.Sc.	bronchitis
5 (*Elam*)	40	M	B.Sc.	bronchitis
6 (*Frank*)	45	M	B.Sc.	bronchitis
7 (*George*)	35	M	B.Sc.	pneumonia
8 (*Henry*)	37	M	B.Sc.	pneumonia
9 (*Ivy*)	50	F	Ph.D.	dyspepsia
10 (*Jane*)	60	F	Ph.D.	pneumonia

(a) The microdata

QIs	Disease
(25-29, M, MSc)	HIV
	HIV
(25-29, *, BSc)	pneumonia
	bronchitis
(40-45, M, BSc)	bronchitis
	bronchitis
(35-37, M, BSc)	pneumonia
	pneumonia
(50-60, F, PhD)	dyspepsia
	pneumonia

(b) A 2-anonymous table

The main motivation

ℓ-diversity is motivated from privacy preservation for data publication (Machanavajjhala et. al. 2006), follows ℓ-anonymity (a.k.a. k-anonymity, Samarati 2001).

quasi-identifier (QI)

sensitive attribute

T-1D(*Name*)	Age	Gender	Degree	Disease
1 (*Adam*)	29	M	M.Sc.	HIV
2 (*Bob*)	25	M	M.Sc.	HIV
3 (*Calvin*)	25	M	B.Sc.	pneumonia
4 (*Daisy*)	29	F	B.Sc.	bronchitis
5 (*Elam*)	40	M	B.Sc.	bronchitis
6 (*Frank*)	45	M	B.Sc.	bronchitis
7 (*George*)	35	M	B.Sc.	pneumonia
8 (*Henry*)	37	M	B.Sc.	pneumonia
9 (*Ivy*)	50	F	Ph.D.	dyspepsia
10 (*Jane*)	60	F	Ph.D.	pneumonia

(a) The microdata

QIs	Disease
(25-29, M, MSc)	HIV
	HIV
(25-29, *, BSc)	pneumonia
	bronchitis
(40-45, M, BSc)	bronchitis
	bronchitis
(35-37, M, BSc)	pneumonia
	pneumonia
(50-60, F, PhD)	dyspepsia
	pneumonia

(b) A 2-anonymous table

QIs	Disease
(25-29, M, *)	HIV
	pneumonia
(25-29, *, *)	HIV
	bronchitis
(35-40, M, BSc)	bronchitis
	pneumonia
(37-45, M, BSc)	bronchitis
	pneumonia
(50-60, F, PhD)	dyspepsia
	pneumonia

(c) An 2-diverse table

Previous work

- ℓ-anonymity
  - A 2-approximation in the metric space is known (Aggarwal el. al. 2006).

Previous work

- ℓ-anonymity
  - A 2-approximation in the metric space is known (Aggarwal el. al. 2006).
- ℓ-diversity
  - Many heuristic solutions have been proposed in the DB community (i.e. LeFevre et. al. 2006, Ghinita et. al. 2007). But no theoretical result.

Previous work

- ℓ-anonymity
  - A 2-approximation in the metric space is known (Aggarwal el. al. 2006).
- ℓ-diversity
  - Many heuristic solutions have been proposed in the DB community (i.e. LeFevre et. al. 2006, Ghinita et. al. 2007). But no theoretical result.
- Outliers (remove some points to get better clusters)
  - First considered by Charikar et. al. 2001 for facility location and k-median.
  - A 4-approximation is known for ℓ-anonymity (Aggarwal el. al. 2006).

Related work on instance-level constraints

- ML constraints and CL constraints (Wagstaff and Cardie 2000)
  - *must-link* (ML): two points must be clustered together.
    cannot-link (CL): two points must be separated.
  - ℓ-diverse clustering can be seen as a special case where nodes with the same color must satisfy CL constraints.
  - No approximation algorithm is studied.

Related work on instance-level constraints

- ML constraints and CL constraints (Wagstaff and Cardie 2000)
  - *must-link* (ML): two points must be clustered together.
    cannot-link (CL): two points must be separated.
  - ℓ-diverse clustering can be seen as a special case where nodes with the same color must satisfy CL constraints.
  - No approximation algorithm is studied.
- Correlation clustering (Bansal et. al. 2004)
  - Minimize the violation of the given constraints.
  - Best approximation algorithms are due to Ailon et. al. 2008

Our results for l-diversity

- A 2-approximation algorithm

  (if the problem has feasible solutions).

- A matching lower bound assuming $P \neq NP$

  (even there are only 3 colors)

- An $O(1)$-approximation algorithm for the infeasible case.

  (if the problem does not have a feasible solution, we remove the least possible number of points to get a feasible solution)

2-approximation algorithm

Problems and definitions

- Recall ℓ-diversity: Given a set V of points in a metric space, each of which has a color, subject to:
  1. each cluster has size $\geq l$;
  2. all points partitioned into one cluster must have distinct colors.

  Goal: try to minimize the largest diameter of clusters.

Problems and definitions

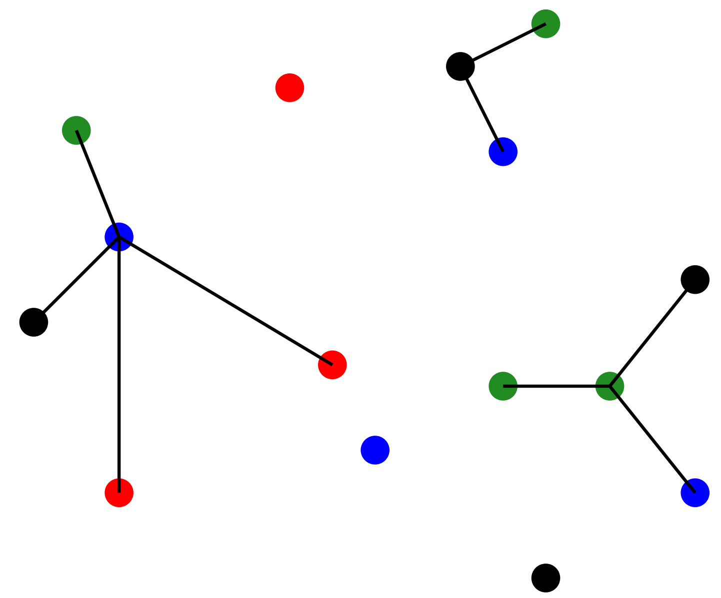
- Recall ℓ-diversity: Given a set V of points in a metric space, each of which has a color, subject to:
  1. each cluster has size $\geq l$;
  2. all points partitioned into one cluster must have distinct colors.

  Goal: try to minimize the largest diameter of clusters.

- Given V, construct $G(V, E)$ with
  - each $v \in V$ has a color $c(v)$;
  - for each pair of points u,v with distinct colors, creat an edge $e = (u, v)$ with weight $w(e)$, which is the distance of u, v in the metric space.
  - diameter of $C \subseteq V$: $\max_{u,v \in C}(w(e(u, v)))$.

Definitions (cont.)

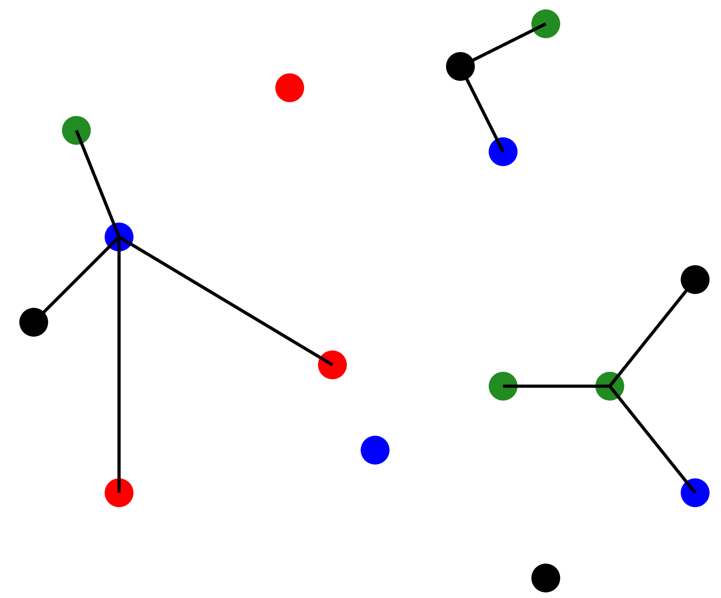
- *star forest*: a forest where each connected component is a star.



Definitions (cont.)

- *star forest*: a forest where each connected component is a star.

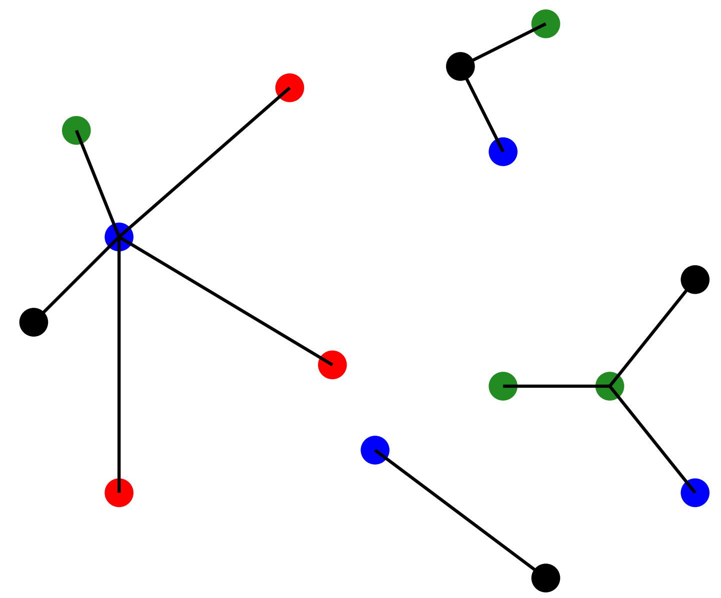
- *spanning star forest*: a star forest with no isolated point.



Definitions (cont.)

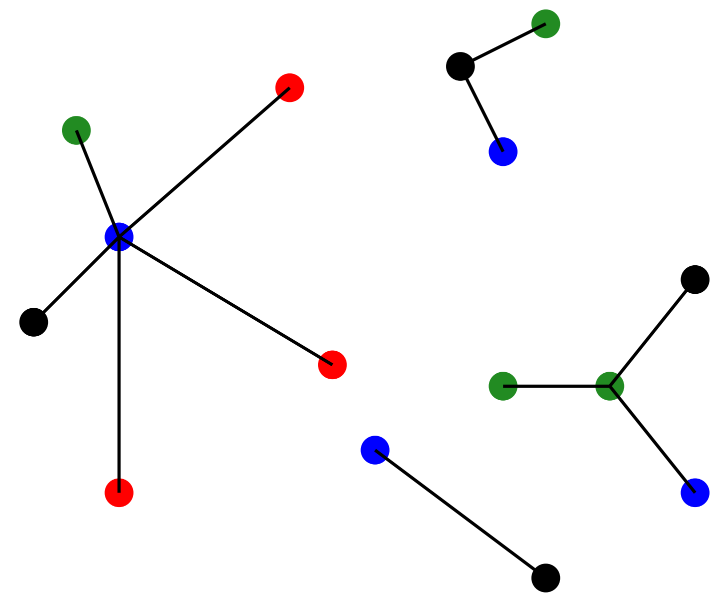
- *star forest*: a forest where each connected component is a star.

- *spanning star forest*: a star forest with no isolated point.



Definitions (cont.)

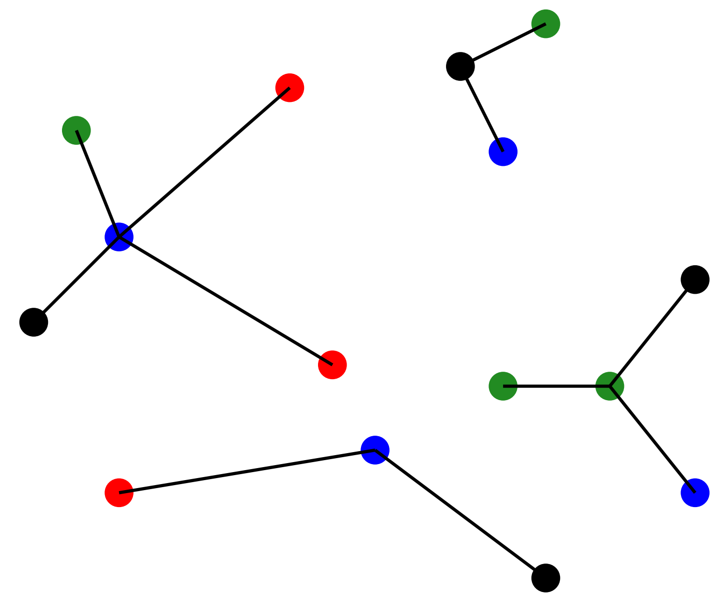
- *star forest*: a forest where each connected component is a star.
- *spanning star forest*: a star forest with no isolated point.



$\ell = 3$

- *semi-valid spanning star forest*: a spanning star forest with each component containing at least ℓ colors.

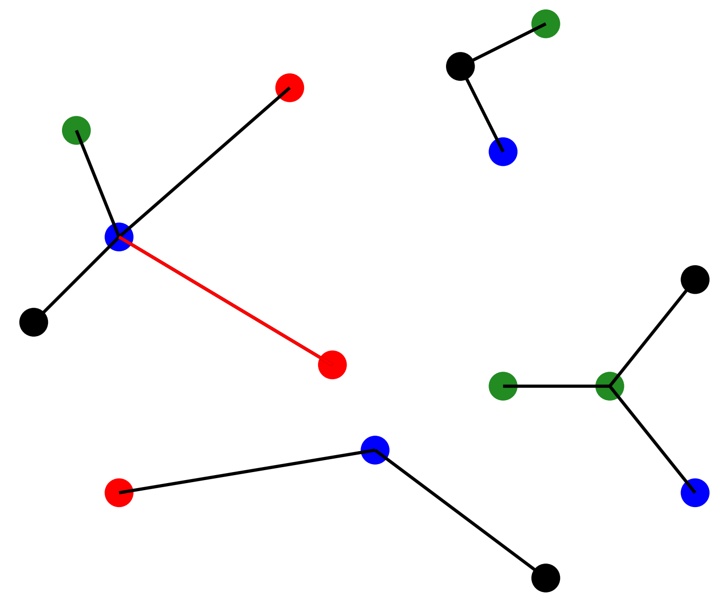
Definitions (cont.)



- *star forest*: a forest where each connected component is a star.
- *spanning star forest*: a star forest with no isolated point.
- *semi-valid spanning star forest*: a spanning star forest with each component containing at least ℓ colors.

Definitions (cont.)

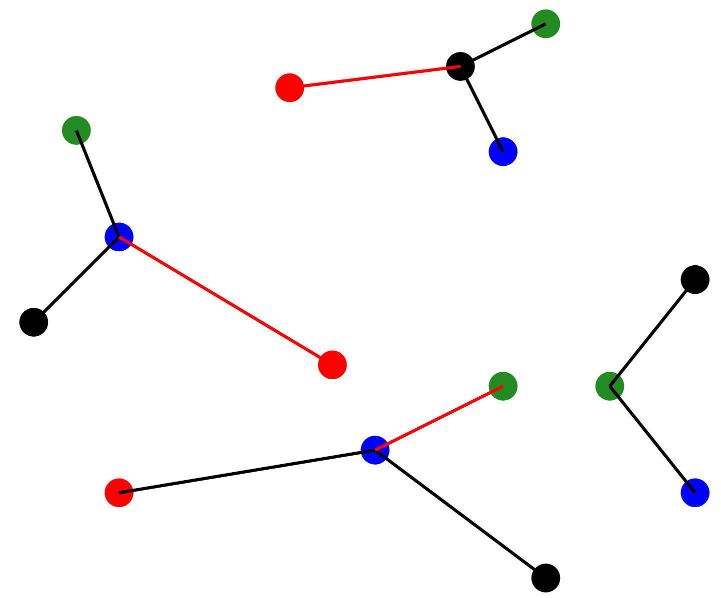
- *star forest*: a forest where each connected component is a star.
- *spanning star forest*: a star forest with no isolated point.



$\ell = 3$

- *semi-valid spanning star forest*: a spanning star forest with each component containing at least ℓ colors.
- *valid spanning star forest*: a semi-valid spanning star forest with each component containing points with distinct colors.

Definitions (cont.)



$\ell = 3$

- *star forest*: a forest where each connected component is a star.
- *spanning star forest*: a star forest with no isolated point.
- *semi-valid spanning star forest*: a spanning star forest with each component containing at least ℓ colors.
- *valid spanning star forest*: a semi-valid spanning star forest with each component containing points with distinct colors.

A review on 2-approximation for ℓ-anonymity

- General idea of the algorithm in Aggarwal el. al. 2006.
  1. Let $e_1, e_2, \ldots$ be the edge of G in a non-decreasing order of weights.
  2. Consider each graph G_i formed by the first i edges $E_i = \{e_1, e_2, \ldots, e_i\}$.
  3. For each $G_i = (V, E_i)$, try to

A review on 2-approximation for ℓ-anonymity

- General idea of the algorithm in Aggarwal el. al. 2006.
  1. Let $e_1, e_2, \ldots$ be the edge of G in a non-decreasing order of weights.
  2. Consider each graph G_i formed by the first i edges $E_i = \{e_1, e_2, \ldots, e_i\}$.
  3. For each $G_i = (V, E_i)$, try to

     find a maximal independent set (IS) I s.t.

     (1) there is a spanning star forest in G_i with the nodes in I being the star centers,

     (2) each star has at least ℓ nodes.

A review on 2-approximation for ℓ-anonymity

- General idea of the algorithm in Aggarwal el. al. 2006.

  1. Let $e_1, e_2, \ldots$ be the edge of G in a non-decreasing order of weights.
  2. Consider each graph G_i formed by the first i edges $E_i = \{e_1, e_2, \ldots, e_i\}$.
  3. For each $G_i = (V, E_i)$, try to

     find a maximal independent set (IS) I s.t.

     (1) there is a spanning star forest in G_i with the nodes in I being the star centers,

     (2) each star has at least ℓ nodes.

- Let the diameter of the OPT clustering be d^* with $w(e_{i^*}) = d^*$. Aggarwal el. al. shows that the trial on G_{i^*} must succeed.

A review on 2-approximation for ℓ-anonymity

- General idea of the algorithm in Aggarwal el. al. 2006.
  1. Let $e_1, e_2, \ldots$ be the edge of G in a non-decreasing order of weights.
  2. Consider each graph G_i formed by the first i edges $E_i = \{e_1, e_2, \ldots, e_i\}$.
  3. For each $G_i = (V, E_i)$, try to

     find a maximal independent set (IS) I s.t.

     (1) there is a spanning star forest in G_i with the nodes in I being the star centers,

     (2) each star has at least ℓ nodes.
- Let the diameter of the OPT clustering be d^* with $w(e_{i^*}) = d^*$. Aggarwal el. al. shows that the trial on G_{i^*} must succeed.

$\Rightarrow$ 2-approximation (SOL $\leq 2d^*$)

Our algorithm for ℓ-diversity

- Our algorithm for ℓ-diversity follows the same framework but we try to

  find a maximal IS I s.t.

  there is a valid spanning star forest in G_i with the nodes in I being the star centers.

Our algorithm for ℓ-diversity

- Our algorithm for ℓ-diversity follows the same framework but we try to

  find a maximal IS I s.t.

  there is a valid spanning star forest in G_i with the nodes in I being the star centers.

  We can also prove that the algorithm will succeed on G_{i^*}, $\Rightarrow$ 2-approximation.

Our algorithm for ℓ-diversity

- Our algorithm for ℓ-diversity follows the same framework but we try to

  find a maximal IS I s.t.

  there is a valid spanning star forest in G_i with the nodes in I being the star centers.

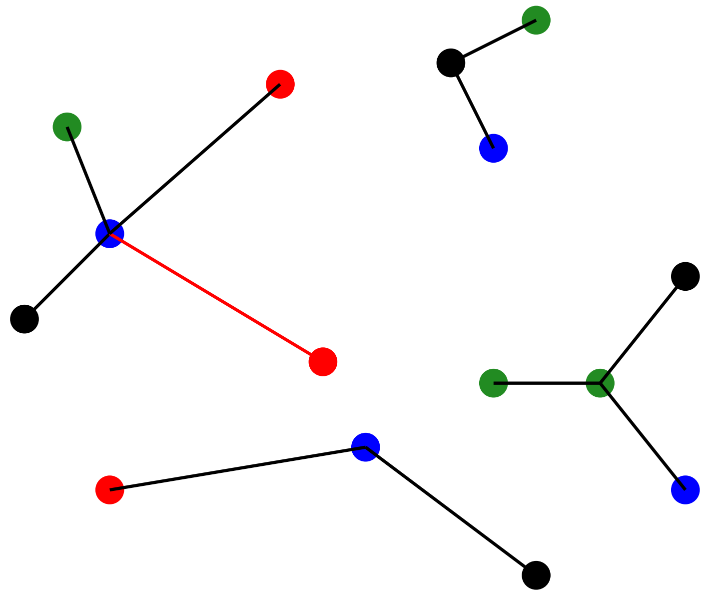
  We can also prove that the algorithm will succeed on G_{i^*}, $\Rightarrow$ 2-approximation.

- The additional challenge: nodes in a star have distinct colors.

Two tests

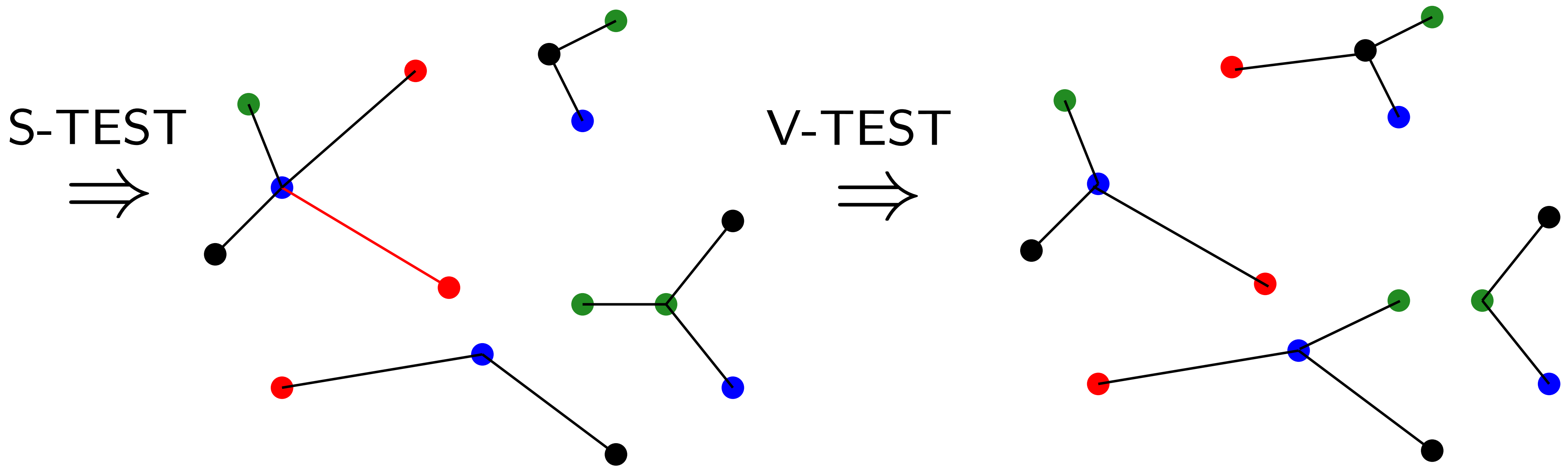
- S-TEST(G_i, I): check if there exists a semi-valid spanning star forest in G_i with nodes in I being star centers.

S-TEST
$\Rightarrow$



Two tests

- S-TEST(G_i, I): check if there exists a semi-valid spanning star forest in G_i with nodes in I being star centers.



- V-TEST(G_i, I): check if there exists a valid spanning star forest in G_i with nodes in I being star centers.

The 2-approximate algorithm

- We just perform the trial on graphs $G_1, G_2, \ldots$ one by one, until we find on some G_i a valid spanning forest with a maximal independent set as centers.

1. Let I be an arbitrary maximal IS in G_i
2. **While** S-TEST(G_i, I) is passed
   (a) $(S, S') \leftarrow$ V-TEST(G_i, I) /* $S \subseteq V - I, S' \subseteq I$ */
   (b) **If** $S = \emptyset$ **then** *Succeed*; **else**
      i. $I \leftarrow I - S' + S$
      /* $|S'| < |S|$, $|I|$ increase and I is still an IS */
      ii. Add nodes to I until it is a maximal IS
3. *Fail*;

The 2-approximate algorithm

- We just perform the trial on graphs $G_1, G_2, \ldots$ one by one, until we find on some G_i a valid spanning forest with a maximal independent set as centers.

1. Let I be an arbitrary maximal IS in G_i
2. **While** S-TEST(G_i, I) is passed
   (a) $(S, S') \leftarrow$ V-TEST(G_i, I) /* $S \subseteq V - I, S' \subseteq I$ */
   (b) **If** $S = \emptyset$ **then** *Succeed*; **else**
      i. $I \leftarrow I - S' + S$
      /* $|S'| < |S|$, $|I|$ increase and I is still an IS */
      ii. Add nodes to I until it is a maximal IS
3. *Fail*;

redistribute points
and
pick new centers

The 2-approximate algorithm

- We just perform the trial on graphs $G_1, G_2, \ldots$ one by one, until we find on some G_i a valid spanning forest with a maximal independent set as centers.

1. Let I be an arbitrary maximal IS in G_i
2. **While** S-TEST(G_i, I) is passed
   (a) $(S, S') \leftarrow$ V-TEST(G_i, I) /* $S \subseteq V - I, S' \subseteq I$ */
   (b) **If** $S = \emptyset$ **then** *Succeed*; **else**
      i. $I \leftarrow I - S' + S$
      /* $|S'| < |S|$, $|I|$ increase and I is still an IS */
      ii. Add nodes to I until it is a maximal IS
3. *Fail*;

- **Claim**: Both tests succeed on G_{i^*}.
$\Rightarrow \exists$ a valid spanning star forest on G_{i^*}

Lowerbound

- **Theorem**: There is no polynomial-time approximation algorithm for ℓ-diversity that achieves an approximation factor less than 2 unless $\mathsf{P} = \mathsf{NP}$.
  - By reduction to 3D-matching.

Lowerbound

- **Theorem**: There is no polynomial-time approximation algorithm for ℓ-diversity that achieves an approximation factor less than 2 unless P $=$ NP.
  - By reduction to 3D-matching.
  - Holds even there are only 3 colors.
  - If there are 2 colors, the problem can be solved in polynomial time by finding perfect matchings in the $G_1, G_2, \ldots$.

The infeasible case (high-level sketch)

- If some color has more than $\lfloor n/l \rfloor$ points, there is no feasible solution, thus we must remove some points.

  Least number of points that should be removed to get a feasible solution can be computed, say, k points.

  Goal: Compute an optimal clustering by deleting k points.

The infeasible case (high-level sketch)

- If some color has more than $\lfloor n/l \rfloor$ points, there is no feasible solution, thus we must remove some points.

  Least number of points that should be removed to get a feasible solution can be computed, say, k points.

  Goal: Compute an optimal clustering by deleting k points.

- Additional challenge: have to decide (even know k):

  which points should be deleted?

The infeasible case (high-level sketch)

- If some color has more than $\lfloor n/l \rfloor$ points, there is no feasible solution, thus we must remove some points.

  Least number of points that should be removed to get a feasible solution can be computed, say, k points.

  Goal: Compute an optimal clustering by deleting k points.

- Additional challenge: have to decide (even know k):

  which points should be deleted?

- Our strategy: to find

  a valid star forest spanning $n - k$ nodes in $G_{i^*}^{28}$.

  $\Rightarrow$ 56-approximation

Futher work

- Try to minimize the sum of the diameters of the clusters.

Futher work

- Try to minimize the sum of the diameters of the clusters.

- Design approximation algorithms for the problem by removing any fixed number of points.

Futher work

- Try to minimize the sum of the diameters of the clusters.

- Design approximation algorithms for the problem by removing any fixed number of points.

- Consider other instance-level constraints like the general Must-Link and Cannot-Link constraints,

The End

THANK YOU

Q and A