Chromatic Correlation Clustering and the Pivot

Bruno Ordozgoiti[1]

[1]Aalto University

Helsinki 2021

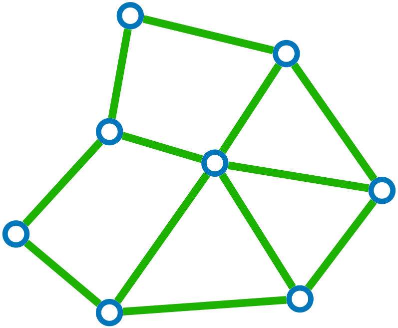
Correlation clustering

Definition

CorrelationClustering

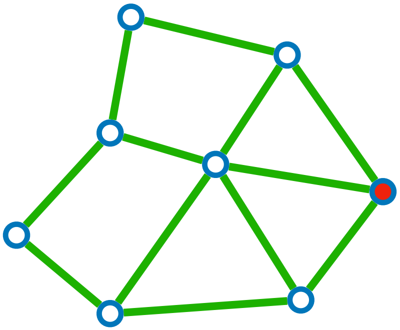
- Input: graph $G = (V, E)$.
- Solution: clustering $c : V \to \mathbb{N}$.
- Objective: minimize the number of disagreements:
  - $uv \notin E \wedge c(u) = c(v)$,
  - $uv \in E \wedge c(u) \neq c(v)$.

The PIVOT algorithm — example



Credit: Aris

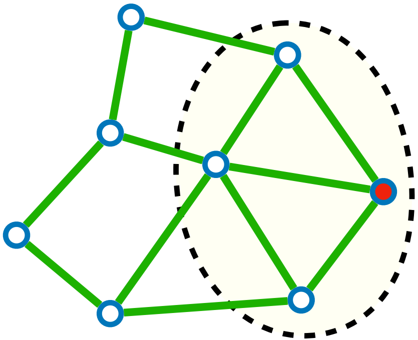
The PIVOT algorithm — example



a pivot is selected uniformly at random

Credit: Aris

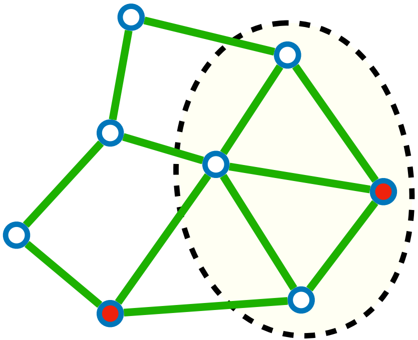
The PIVOT algorithm — example



a cluster is formed with the pivot and all its neighbors

Credit: Aris

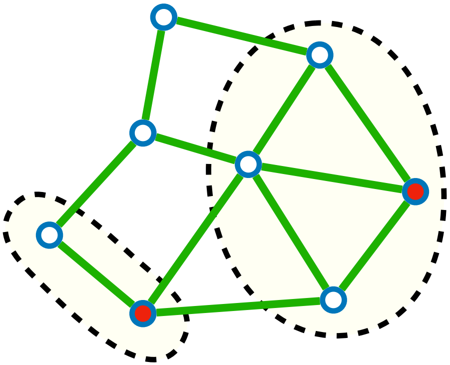
The PIVOT algorithm — example



a new pivot is selected from the remaining of the graph vertices

Credit: Aris

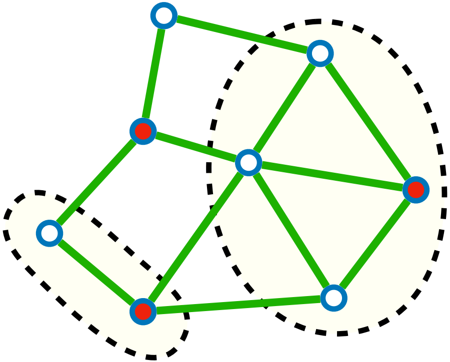
The PIVOT algorithm — example



a second cluster is formed with the pivot and all its neighbors

Credit: Aris

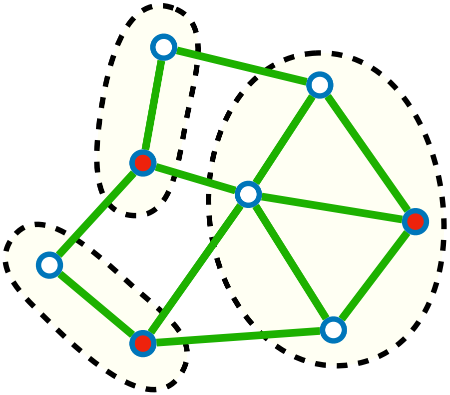
The PIVOT algorithm — example



and the process continues ...

Credit: Aris

The Pivot algorithm — example



... until the whole graph is consumed.

Credit: Aris

Correlation clustering — the KwikCluster (or Pivot) algorithm

```
KwikCluster(G = (V, E⁺, E⁻))

  If V = ∅ then return ∅
  Pick random pivot i ∈ V.
  Set C = {i}, V′ = ∅.

  For all j ∈ V, j ≠ i:
      If (i, j) ∈ E⁺ then
          Add j to C
      Else (If (i, j) ∈ E⁻)
          Add j to V′

  Let G′ be the subgraph induced by V′.

  Return C ∪ KwikCluster(G′) .
```

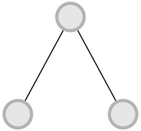
- The Pivot algorithm

  (Ailon et al., 2005)

+ An elegant randomized algorithm
+ Approximation ratio 3
+ Running time $\mathcal{O}(m)$
– It assumes a complete graph
– It assumes an unweighted graph

Analysis of Pivot for CorrelationClustering

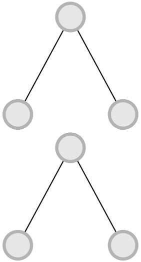
Mistakes come from wedges.
Let T be the set of all wedges.



Analysis of PIVOT for CORRELATIONCLUSTERING

Mistakes come from wedges.
Let T be the set of all wedges.

If all wedges are disjoint, then $OPT \geq |T|$.

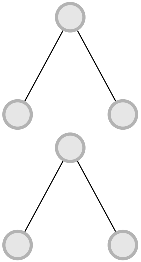


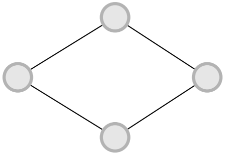
Analysis of Pivot for CorrelationClustering

Mistakes come from wedges.
Let T be the set of all wedges.

If all wedges are disjoint, then $OPT \geq |T|$.

Otherwise...



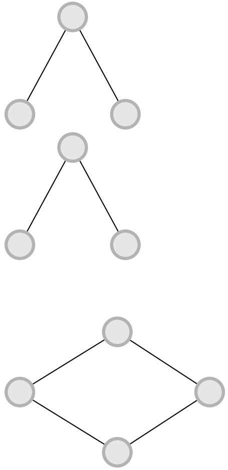


Analysis of Pivot for CorrelationClustering

Mistakes come from wedges.
Let T be the set of all wedges.

If all wedges are disjoint, then $OPT \geq |T|$.

Otherwise...



Lemma

Suppose that for each $t \in T$ we define $\beta_t \geq 0$ s.t. for every $e \in E$, $\sum_{t: e \in t} \beta_t \leq 1$. Then

$$OPT \geq \sum_{t \in T} \beta_t.$$

Analysis of Pivot for CorrelationClustering

If $\forall e \in E$, $\sum_{t:e\in t} \beta_t \le 1$, then $OPT \ge \sum_{t\in T} \beta_t$.

- For $t \in T$, define A_t: event that a vertex of t is the pivot and t is in the recursive call.
- $\mathbb{E}\left[cost_{\textsc{Pivot}}\right] = \sum_{t\in T} \mathbb{P}\left[A_t\right]$.
- B_e: event that e is a mistake.
- $\mathbb{P}\left[B_e \cap A_t\right] = \mathbb{P}\left[B_e | A_t\right] \mathbb{P}\left[A_t\right] = \frac{1}{3}\mathbb{P}\left[A_t\right]$.
- For t, t' s.t. $e \in t \cap t'$, $\mathbb{P}\left[(B_e \cap A_t) \cap (B_e \cap A_{t'})\right] = 0$. Therefore, $\sum_{t:e\in t} \frac{1}{3}\mathbb{P}\left[A_t\right] \le 1$.

So

$$OPT \ge \sum_{t\in T} \frac{1}{3}\mathbb{P}\left[A_t\right] = \frac{\mathbb{E}\left[cost_{\textsc{Pivot}}\right]}{3}.$$

Chromatic Correlation Clustering

Definition

CHROMATICCORRELATIONCLUSTERING

- Instance: edge-colored graph $G = (V, E, \ell)$, $\ell : E \to L \subset \mathbb{N}$.
- Solution:
  - $c : V \to \mathbb{N}$,
  - $\lambda : im(c) \to L$.
- Objective: minimize the number of disagreements:
  - $uv \notin E \wedge c(u) = c(v)$,
  - $uv \in E \wedge c(u) \neq c(v)$,
  - $uv \in E \wedge c(u) = c(v) \wedge \ell(uv) \neq \lambda(c(u))$.

ChromaticCorrelationClustering- The literature

Approximation algorithms for ChromaticCorrelationClustering.

Method	Factor	Work
ChromaticBalls	6Δ	Bonchi et al. (2015)
Reduce-And-Cluster	11	Anava et al. (2015)
LP	4	Anava et al. (2015)
Pivot	3	Klodt et al. (2021)

Chromatic Correlation Clustering

"Judge a vertex not for the color of its edge, but for the content of its entries in the adjacency matrix".

Chromatic Correlation Clustering

"Judge a vertex not for the color of its edge, but for the content of its entries in the adjacency matrix".

Algorithm 1: `Pivot`

Data: An undirected, edge-colored Graph $G = (V, E, col)$

Result: A clustering $C = \{(C_1, c_1), \ldots (C_m, c_m)\}$ with $C_i \subseteq V$ and $c_i \in \mathbb{N}$.

Pick a random pivot $v \in V$ as cluster-center;

$C \leftarrow \{v\}$;

for $u \in N(v)$ **do**

    $C \leftarrow C \cup \{u\}$;

$c \leftarrow argmax_{c \in Colors} |\{ab \in E \cap C^2 \mid col(ab) = c\}|$;

return $\{(C, c)\} \cup Pivot(G[V \setminus C])$;

Analysis of Pivot for ChromaticCorrelationClustering

Analysis by Klodt et al. (2021).

Consider three solutions:

- $opt = (C^*, \lambda^*)$,
- $S = (C, \lambda)$ (output of Pivot),
- $S' = (C, \lambda')$ (setting $\lambda'(C_i) = \lambda^*(p_i)$).

Note that $d(S) \leq d(S')$.

Definition

Critical iteration of $ab \in \binom{V}{2}$: iteration in which at least the first of a, b becomes clustered.

Analysis of PIVOT for CHROMATICCORRELATIONCLUSTERING

Let ab be a disagreement. Let c be the pivot in the critical iteration of a (first of a, b w.l.o.g).
We will charge $B_{ab} \in \{ab, ac, bc\}$ such that $B_{ab} \in D(opt)$.
Three cases:

1. ab is a non-edge within a cluster of S';
2. ab is an edge between clusters of S';
3. If ab is an edge in a cluster of S' but does not have the color of its cluster.

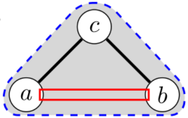
Analysis of Pivot for ChromaticCorrelationClustering

Case 1: ab is a non-edge within a cluster of S';
Note that ac and bc are both edges.

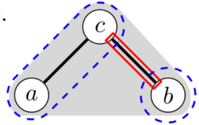
1.1 If a and b are in the same cluster in opt then $ab \in D(opt)$. We charge $B_{ab} := ab$.

1.2 If a and b are in separate clusters in opt: then either ac or bc is an edge between clusters in opt and so in $D(opt)$. We charge $B_{ab} := ac$ if $ac \in D(opt)$, and $B_{ab} := bc$ otherwise.

1.1.



1.2.



Analysis of Pivot for ChromaticCorrelationClustering

Three cases:

1. ab is a non-edge within a cluster of S';
2. ab is an edge between clusters of S';
3. If ab is an edge in a cluster of S' but does not have the color of its cluster.

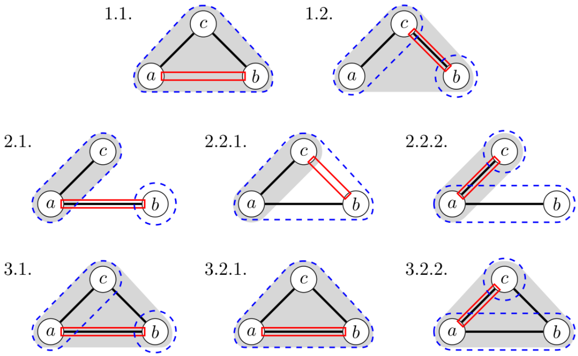


Figure 1: Charging cases in the proof of Theorem 2.1. Gray: Cluster of Sol' during critical iteration of ab. Blue: Cluster of Opt. Red: Charged edge. Note that slight variants of these examples are possible.

Analysis of Pivot for ChromaticCorrelationClustering

For each $e \in d(opt)$, $M_e \subset d(S')$ is the set of disagreements charged to e.

Let $uv \in d(opt)$, $S = N(u) \cup N(v) \cup \{u, v\}$. The pivot at the c.i. of uv is a vertex of S, each with prob. $1/|S|$.
Three cases:

1. uv is a non-edge in a cluster of opt.
2. uv is an edge between clusters of opt.
3. uv is an edge in a cluster of opt with the wrong color.

Analysis of PIVOT for CHROMATICCORRELATIONCLUSTERING

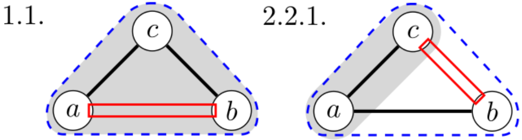
Case 1: uv is a non-edge in a cluster of *opt*.

1.1 Charged only for itself when $p \in N(u) \cap N(v)$

2.1.1 p is u or v.

Thus,

$$\mathbb{E}[M_{uv}] = \frac{|N(u) \cap N(v)|}{|S|} + \frac{2|N(u) \cap N(v)|}{|S|} < 3.$$



Similarly, all three cases yield an expected error of at most 3.

Theorem

The color-blind PIVOT *yields*

$$3OPT \geq \mathbb{E}\left[cost_{\text{PIVOT}}\right].$$

Thanks!

References I

Ailon, N., Charikar, M., and Newman, A. (2005). Aggregating inconsistent information: ranking and clustering. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 684–693.

Anava, Y., Avigdor-Elgrabli, N., and Gamzu, I. (2015). Improved theoretical and practical guarantees for chromatic correlation clustering. In *Proceedings of the 24th International Conference on World Wide Web*, pages 55–65.

Bonchi, F., Gionis, A., Gullo, F., Tsourakakis, C. E., and Ukkonen, A. (2015). Chromatic correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(4):1–24.

Klodt, N., Seifert, L., Zahn, A., Casel, K., Issac, D., and Friedrich, T. (2021). A color-blind 3-approximation for chromatic correlation clustering and improved heuristics. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 882–891.