An Efficient Local Search Algorithm for Correlation Clustering on Large Graphs

Nathan Cordner and George Kollios

COCOA 2023

Talk Outline

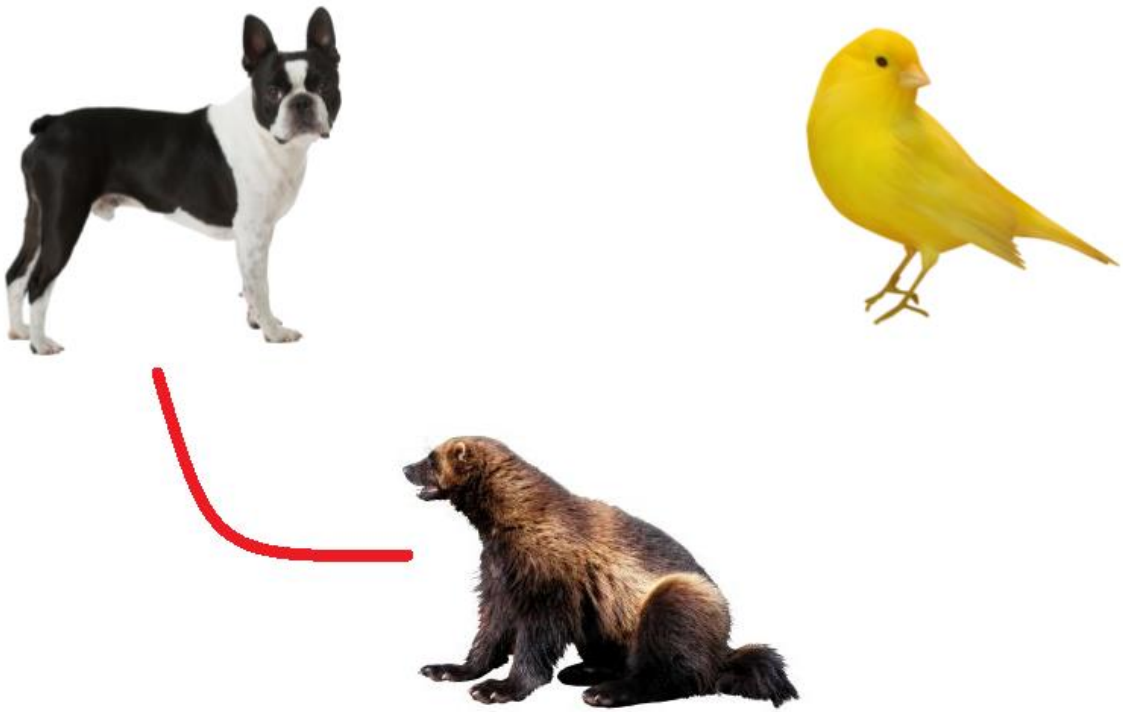
- Introduce the Correlation Clustering problem
- Overview the Pivot algorithm for Correlation Clustering
- Introduce Local Search methods, including **Inner Local Search** (ILS)
- Show effectiveness of Pivot with ILS against other methods

Correlation Clustering

Which animals belong together?

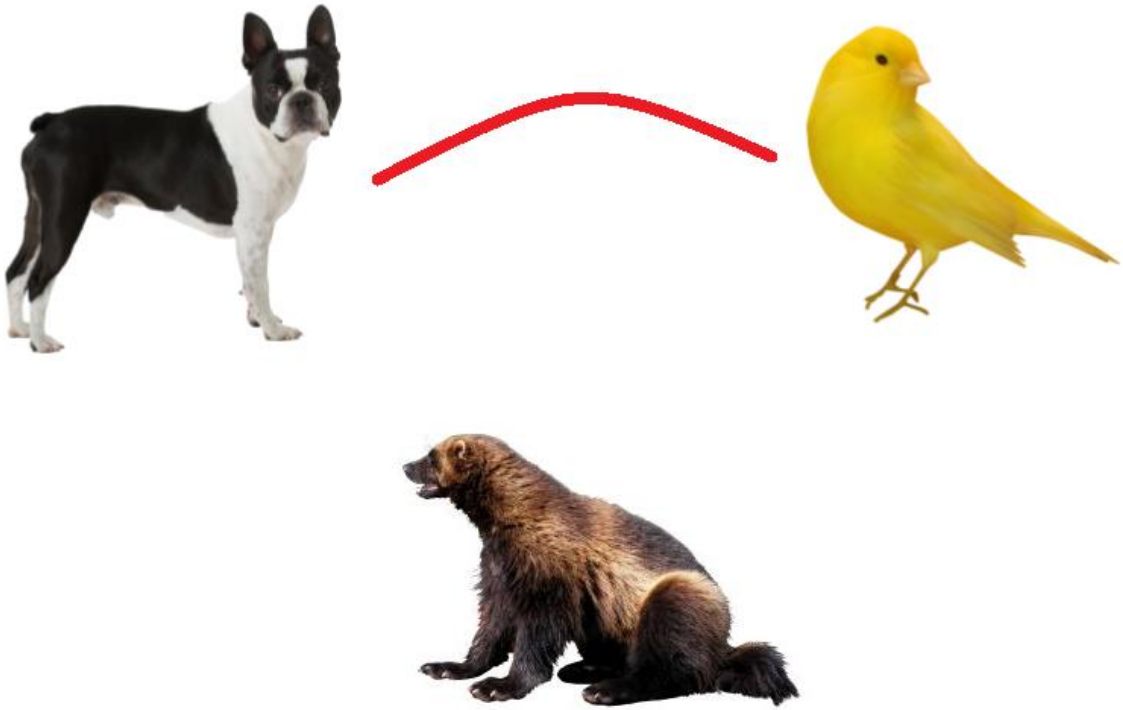
Correlation Clustering

Which animals belong together?



Correlation Clustering

Which animals belong together?



Correlation Clustering

Given a simple undirected graph G = (V, E)

- Let edges represent node similarity
- Clustering cost
  - +1 for every pair of nodes inside a cluster not connected by an edge
  - +1 for every pair of nodes in different clusters that are connected by an edge
- **Goal**: minimize the overall clustering cost
  - Note that a particular number of clusters is not specified!

Correlation Clustering

Applications

- Entity resolution
  - Determine which objects belong to the same real-world entity
- Database deduplication
  - Combine and reduce database entries that represent the same object
- Social network analysis
  - Find tight-knit friend groups in social networks
- Bioinformatics
  - Group proteins together based on similarity

Correlation Clustering Algorithms

Approximation Algorithms

- LP Rounding Methods
  - 2.5-approximation (Ailon et al., STOC 2005)
  - 2.06-approximation (Chawla et al., STOC 2015)
  - 1.994-approximation (Cohen-Addad et al., FOCS 2022)
- **Problem**
  - LP size is large: $O(|V|^2)$ variables and $O(|V|^3)$ constraint equations
  - LP solvers are slow
  - Graph sizes are often large: millions of nodes and billions of edges

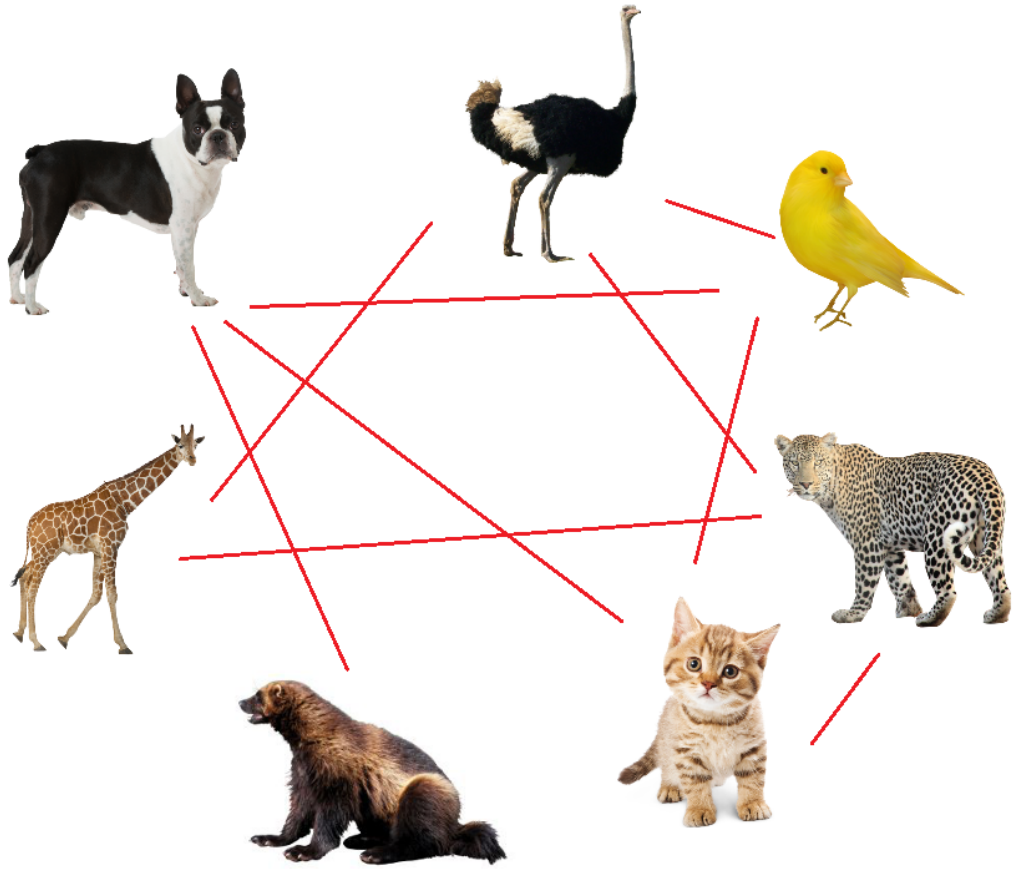
Correlation Clustering Algorithms

The Pivot Algorithm (Ailon et al., STOC 2005)

- Given G = (V, E)
  - Pick random pivot node u in V
  - Set C = {u}
  - For all neighbors v of u in V:
    - Add v to C
  - Add C to the clustering, and repeat on V = V \ C until V is empty
  - Return completed clustering
- Runs in O(|V| + |E|) time
- Randomized expected 3-approximation

Correlation Clustering Algorithms

Example



Correlation Clustering Algorithms

- The Pivot algorithm does not give the best approximation guarantees
- However, it is one of the most popular CC algorithms

Applications

- Clustering social network graphs (Kollios et al., IEEETKDE 2011)
- Clustering protein-protein interaction graphs (Halim et al., IS 2015)
- Clustering event graphs generated from news stories (Christiansen et al., ACM-HT 2017)

Adaptations

- Clustering probabilistic graphs (Kollios et al., IEEETKDE 2011; Mandaglio et al., SIGKDD 2020)
- Chromatic Correlation Clustering (Klodt et al., SIGKDD 2021)
- Fair Correlation Clustering (Ahmadian et al., AISTATS 2020)
- Online Correlation Clustering (Lattanzi et al., NIPS 2021)
- Size-constrained Correlation Clustering (Puleo and Milenkovic, SIAM-JO 2015)
- Parallel Pivot (Chierichetti et al., SIGKDD 2014; Pan et al., NIPS 2015)

Correlation Clustering Improvement

Local Search

- Start with a clustering
  - Repeatedly loop through the node set V in a random order
  - Move current node v to an existing cluster or to a new cluster if it reduces the overall clustering cost
  - Quit if no cost improvement is made after a full loop through V

Running time: $\Theta(|V| + |E|)$ per pass through V

- Worst case: $O(|V|^2)$ passes

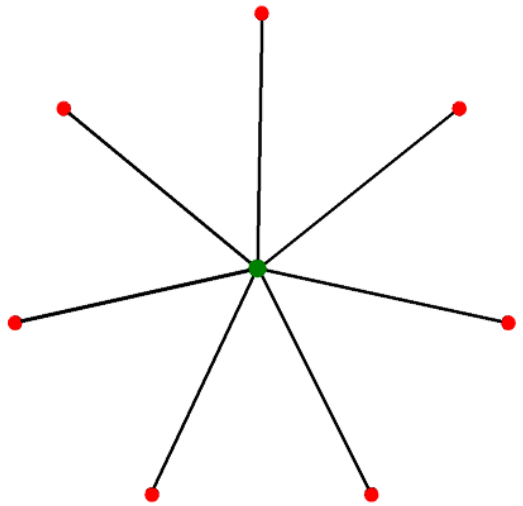
Approximation guarantees based on initial clustering

Correlation Clustering Improvement

- Pivot with Local Search is a common solution
  - Keep the speed of Pivot, but also try to improve the results
- **Problem**: Local Search iterations become expensive on larger graphs
- Example: Orkut social media network
  - 3 million nodes, 120 million edges
  - Pivot runs in about 2 seconds
  - Local Search takes over an hour!

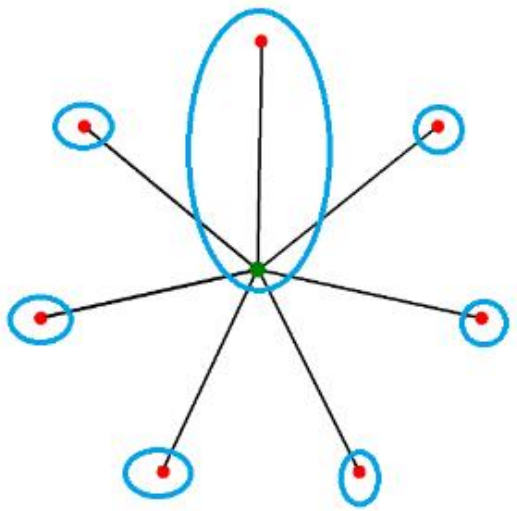
Correlation Clustering Improvement

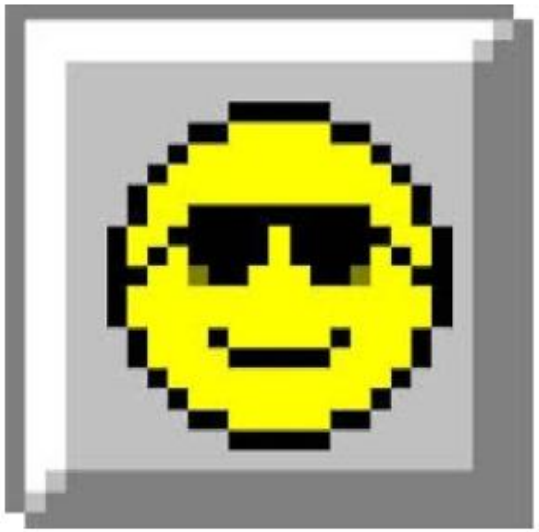
An observation: Pivot forms sparse, star-shaped clusters



Correlation Clustering Improvement

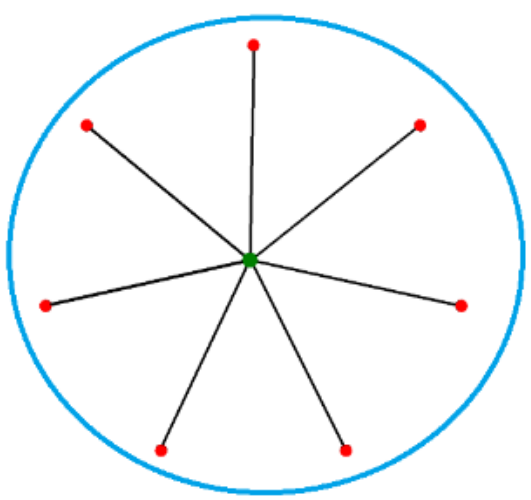
An observation: Pivot forms sparse, star-shaped clusters





Correlation Clustering Improvement

An observation: Pivot forms sparse, star-shaped clusters



Correlation Clustering Improvement

Inner Local Search (**Cordner** and Kollios, COCOA 2023)

- Given clusters C_1, …, C_k (from the Pivot algorithm)
  - Run Local Search to completion **inside** each cluster
  - Return the updated clustering

Running time: $\Theta(|\mathrm{V}| + |\mathrm{E}| + \sum_{i=1}^{k}(|C_i| + |E_i|)I_i)$

- E_i = edges between nodes in cluster C_i
- I_i = number of Local Search iterations used by cluster C_i
- **Big advantage when cluster sizes are much smaller than |V|**

Correlation Clustering Improvement

Inner Local Search (**Cordner** and Kollios, COCOA 2023)

- Given clusters C_1, …, C_k (from the Pivot algorithm)
  - Run Local Search to completion **inside** each cluster
  - Return the updated clustering

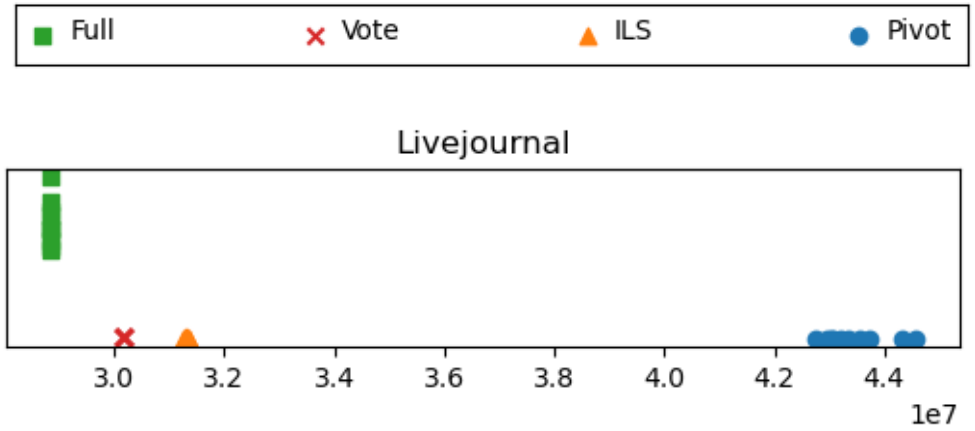
Other properties

- Can be run in parallel (unlike full Local Search)
- Reduction in clustering cost is nearly like full Local Search

Correlation Clustering Improvement

Data Set: Livejournal (a social media network)

- 4 million nodes, 35 million edges, largest Pivot cluster size = 926
- 10 runs per algorithm



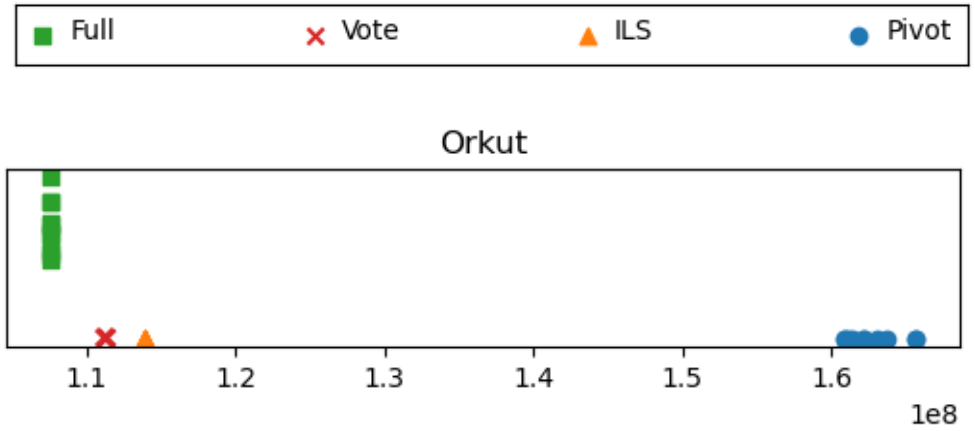
Algorithm	Running Time (s)	Relative Cost
Pivot	2.01	1.0
ILS	3.5	0.721
Full LS	1076	0.664
Match LS	47.8	0.721
Vote	20.5	0.694

https://snap.stanford.edu/data/com-LiveJournal.html

Correlation Clustering Improvement

Data Set: Orkut (a social media network)

- 3 million nodes, 120 million edges, largest Pivot cluster size = 1762
- 10 runs per algorithm



Algorithm	Running Time (s)	Relative Cost
Pivot	1.73	1.0
ILS	7.93	0.7
Full LS	4380	0.661
Match LS	134	0.7
Vote	61.5	0.684

https://snap.stanford.edu/data/com-Orkut.html

Talk Outline

- Introduce the Correlation Clustering problem
- Overview the Pivot algorithm for Correlation Clustering
- Introduce Local Search methods, including **Inner Local Search** (ILS)
- Show effectiveness of Pivot with ILS against other methods

Questions?