Intro to Graph Analytics for Data Scientists

Paul Nguyen, Associate Software Engineer at Anaconda

Apri 21, 2021

Objectives

What this talk is:

- Introduction to new tools for your tool box, e.g.
    - How to represent data via graphs.
    - How to use graph algorithms on graph data.
- Focused on applications of graph analytics.
- Tips on when to and not to use graph analytics.

Objectives

What this talk is NOT:

- A deep-dive into graph theory.
- A deep-dive into graph algorithms.
- A claim that graph analytics can solve all problems.
- A claim that graph analytics will replace any current data analytics methods.

Outline

1 | What are graphs?

2 | Graph data vs Tabular data

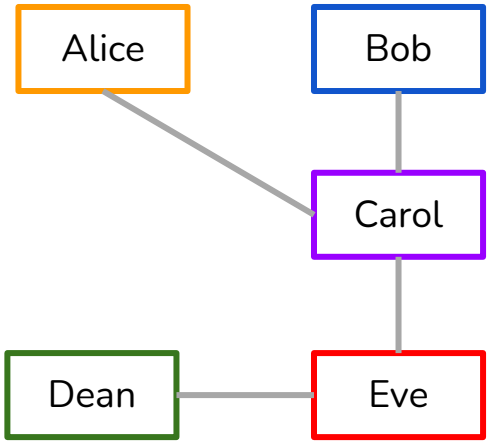
3 | Example use cases

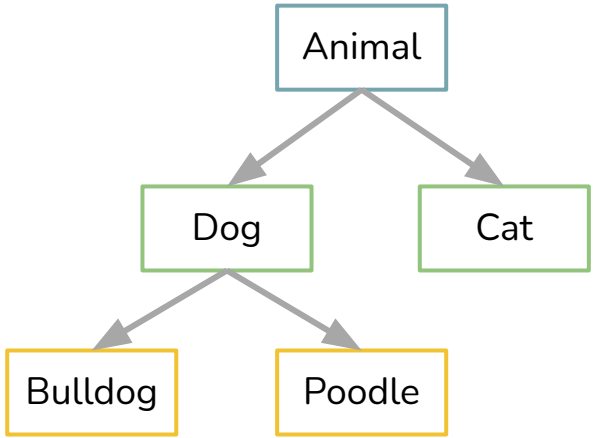
4 | Practical graph analytics tips

5 | Resources to dive deeper

What are graphs?

What are graphs?





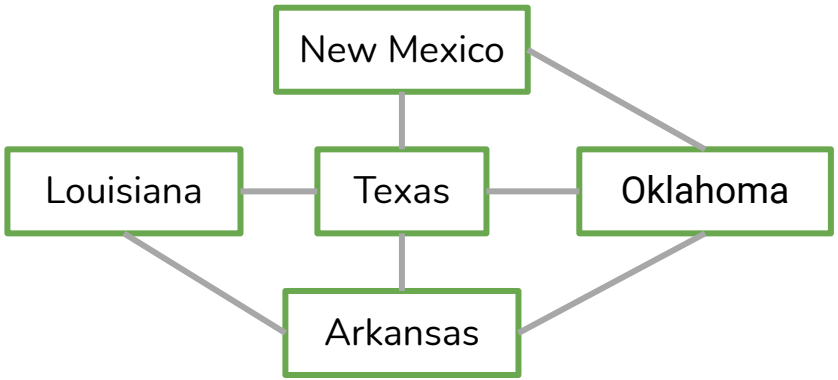
Different types of graphs

Common types of graphs:

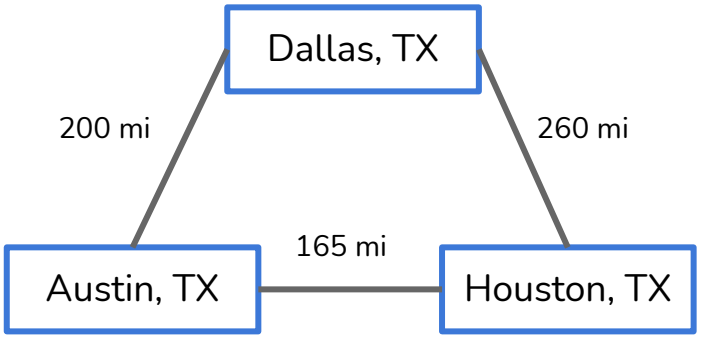
- Weighted/unweighted graphs
- Directed/undirected graphs
- Bipartite graphs
- Many more, e.g. multigraphs, hypergraphs

Weighted/unweighted graphs

Neighboring States

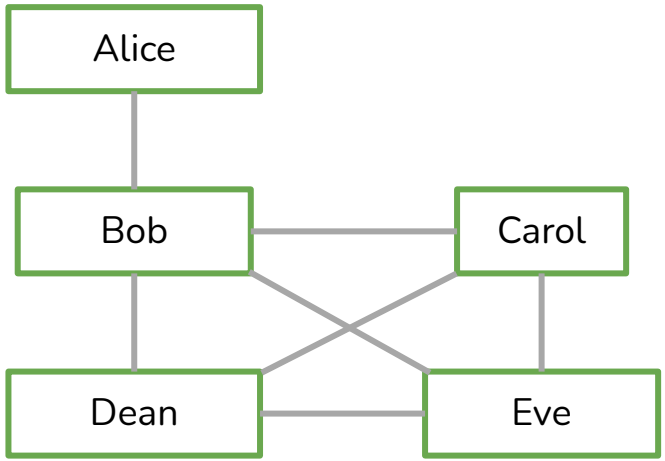


Distance Between Cities

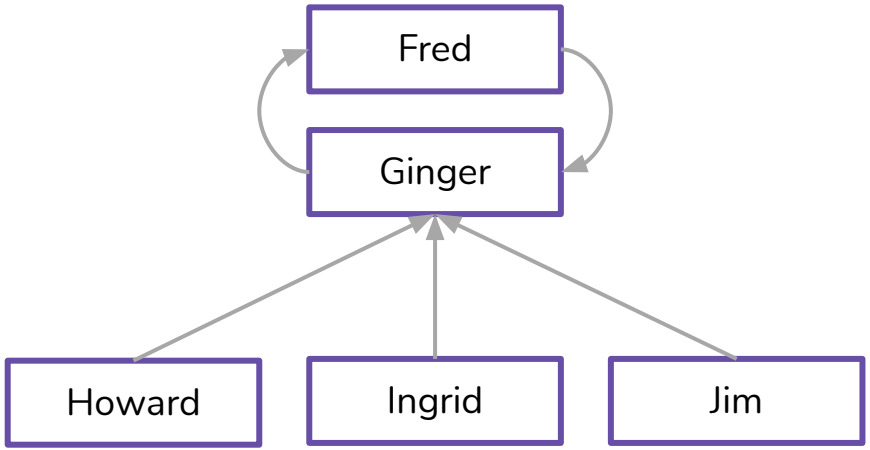


Directed/undirected graphs

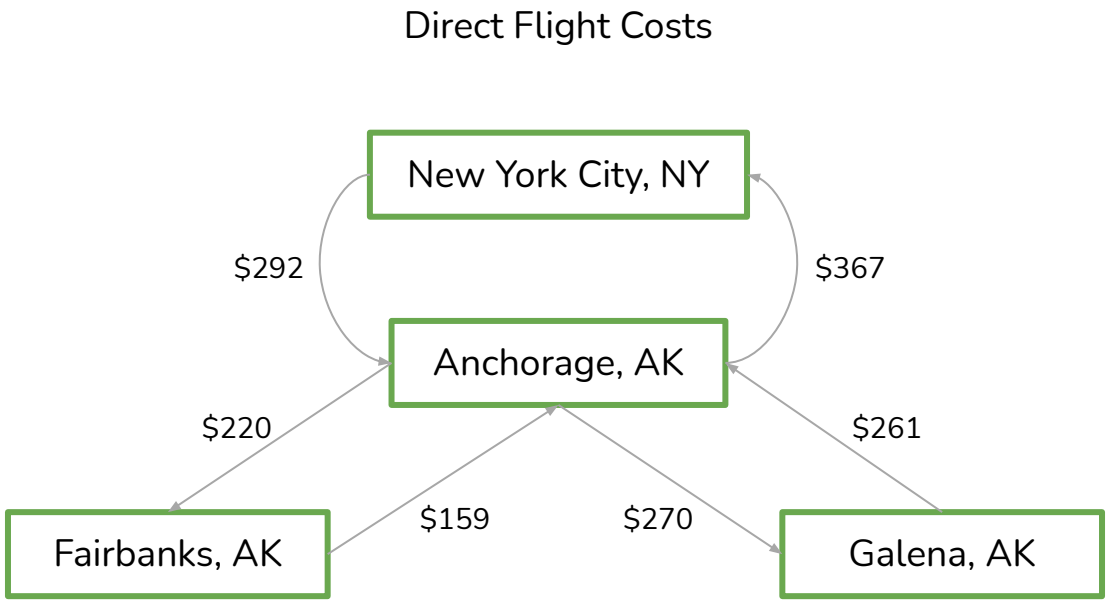
Facebook Friendship



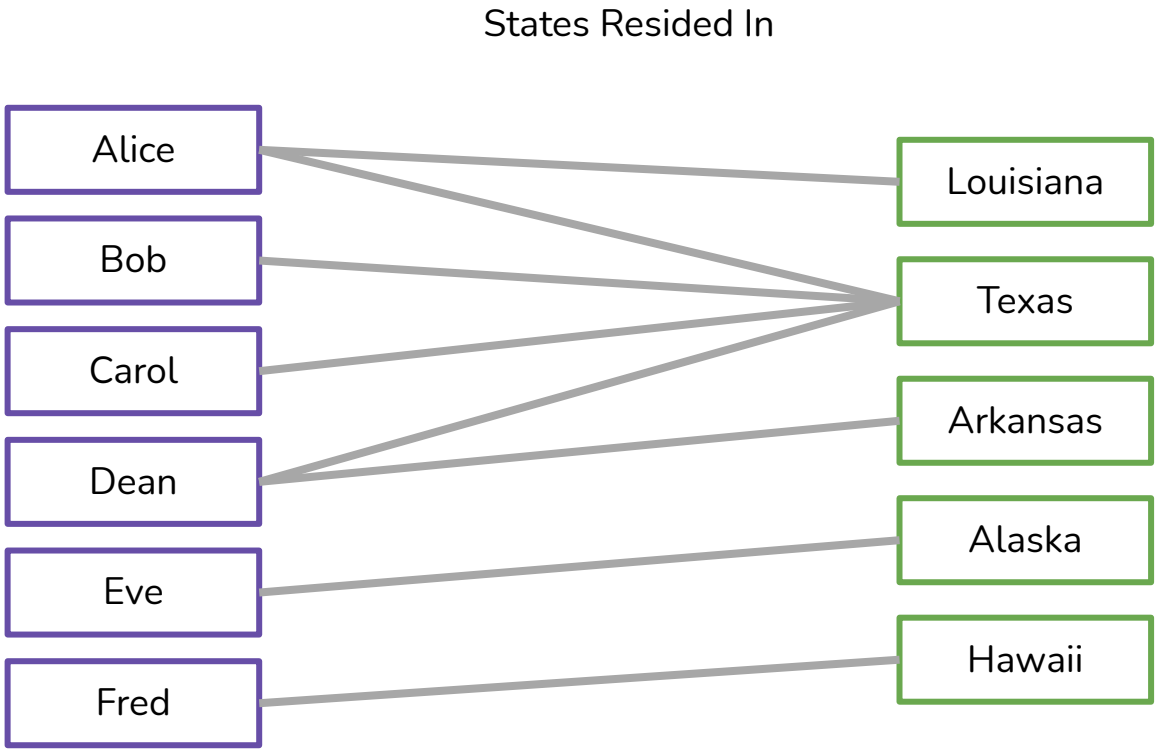
Twitter Followers



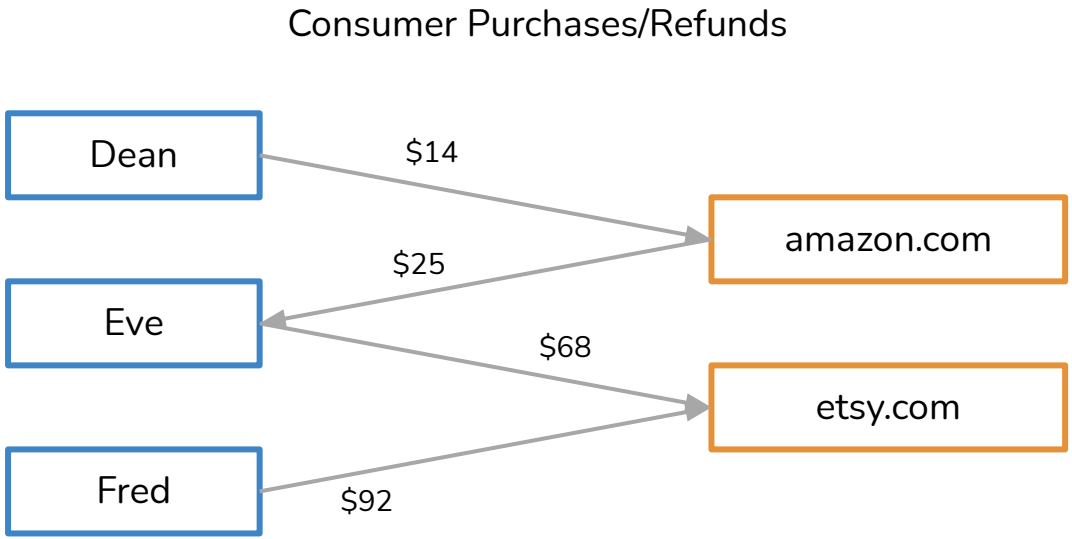
Graphs can be both (and more)



Bipartite graphs



Bipartite graphs (can be weighted/directed also)

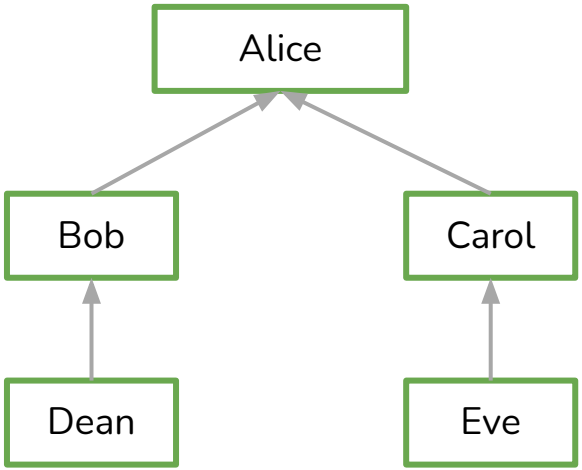


Graph data vs Tabular data

Graph data vs Tabular data

Graph Data

- Great for representing data about the relationship between entities.
- Great for representing transitivity.



Tabular Data

- Great for representing data about individual entities.

	Name	Age	Location	Manager
0	Alice	21	Texas	NULL
1	Bob	33	Louisiana	Alice
2	Carol	33	Texas	Alice
3	Dean	37	Florida	Bob
4	Eve	22	Texas	Carol

Example use cases

Examples

- Use Case 1: Identifying reposting bots
- Use Case 2: Customer clustering
- Use Case 3: Identifying transportation network bottlenecks

Use Case 1:
Identifying reposting bots

Use Case 1: Identifying reposting bots

- Bots posting messages can make analysing how human users view a product difficult.
- Differentiating bots from humans can be difficult as humans can influenced by bots and post similar messages.
- We'll show how to use graph analytics to identify bots.

Use Case 1: Identifying reposting bots

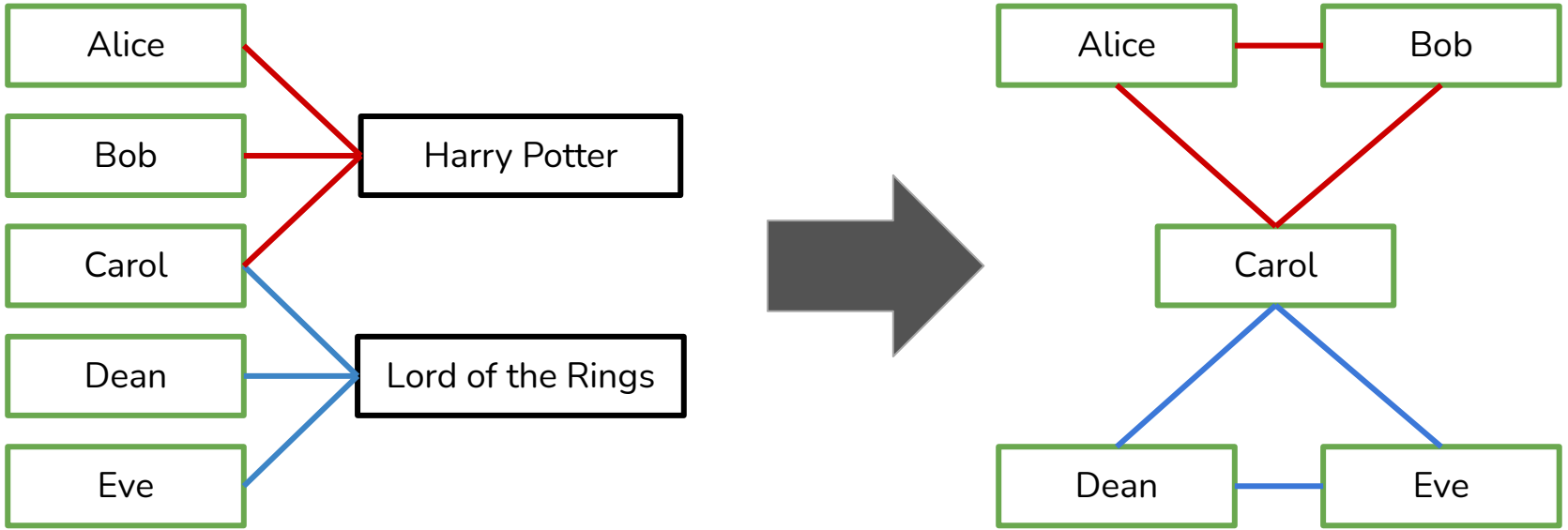
- We'll go over two graph algorithms necessary to understanding how our method will work:

    a. Bipartite graph projection

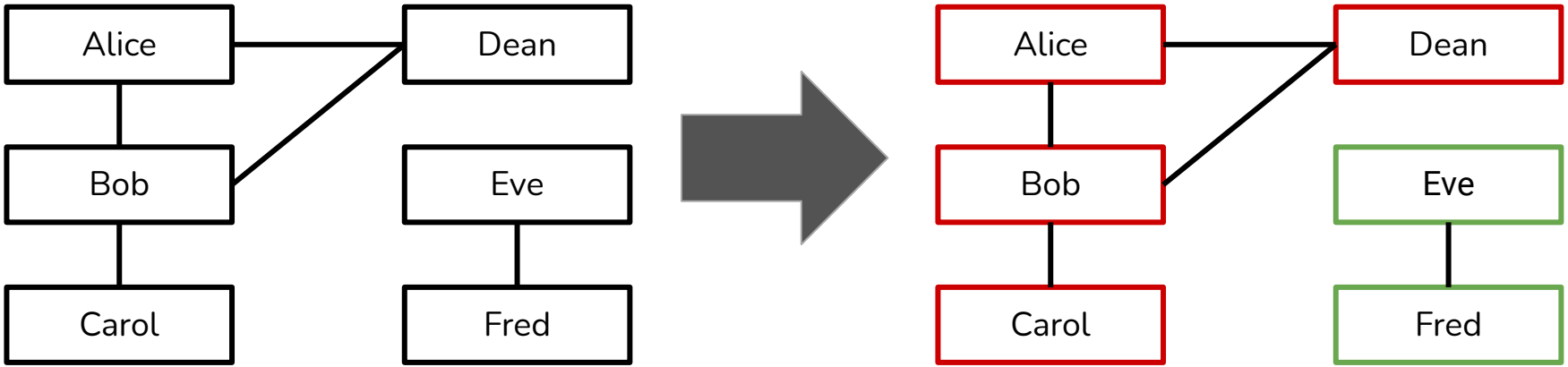
    b. Connected components

Bipartite Graph projection

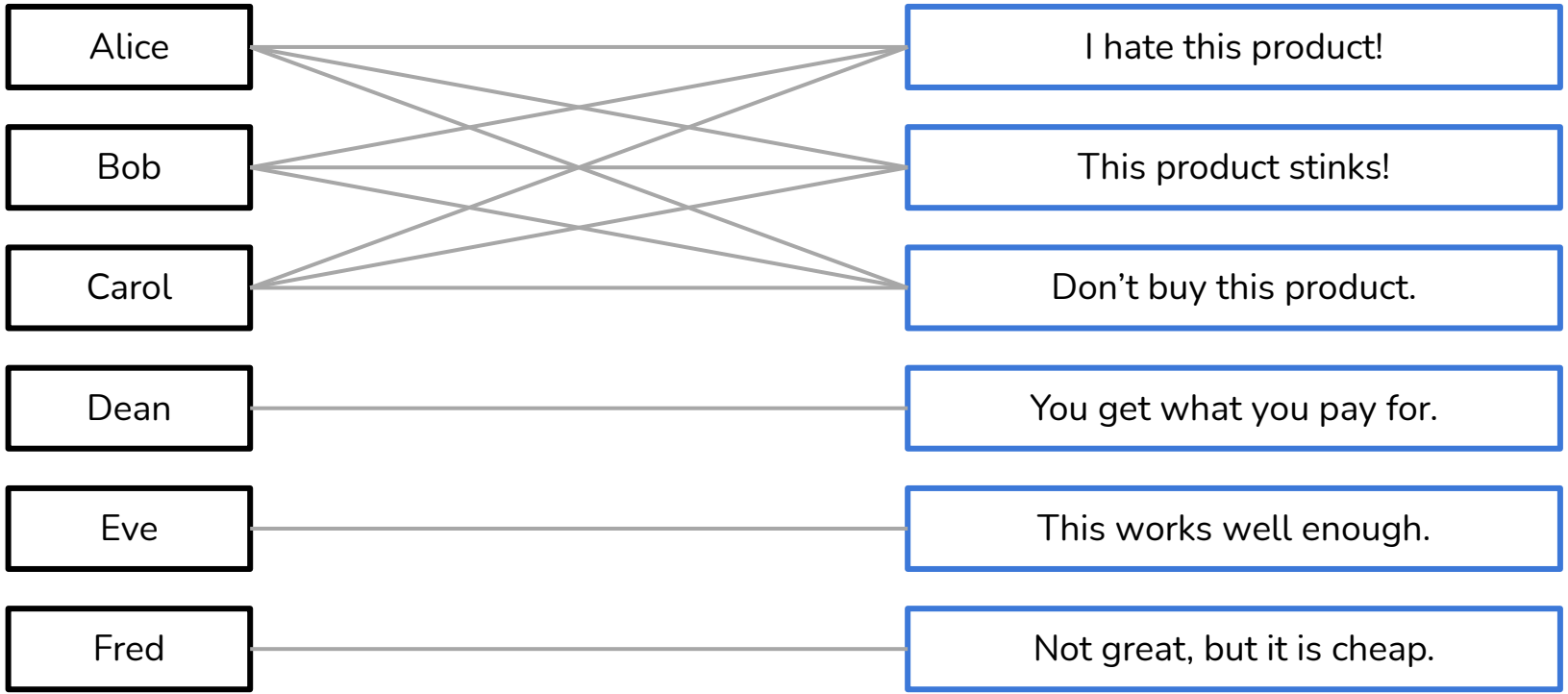
- Transforms a bipartite graph into a single-partition graph.



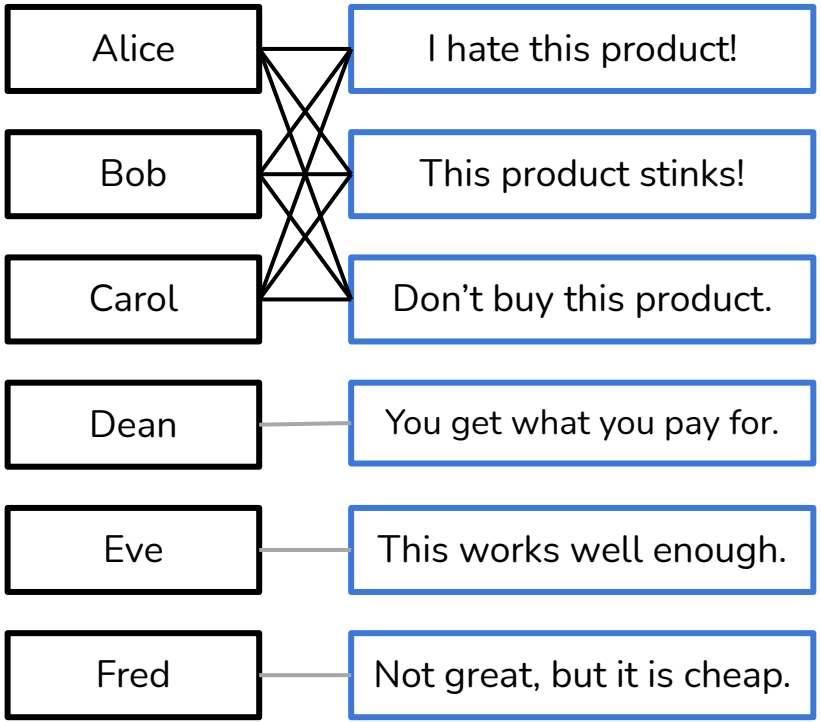
Connected components



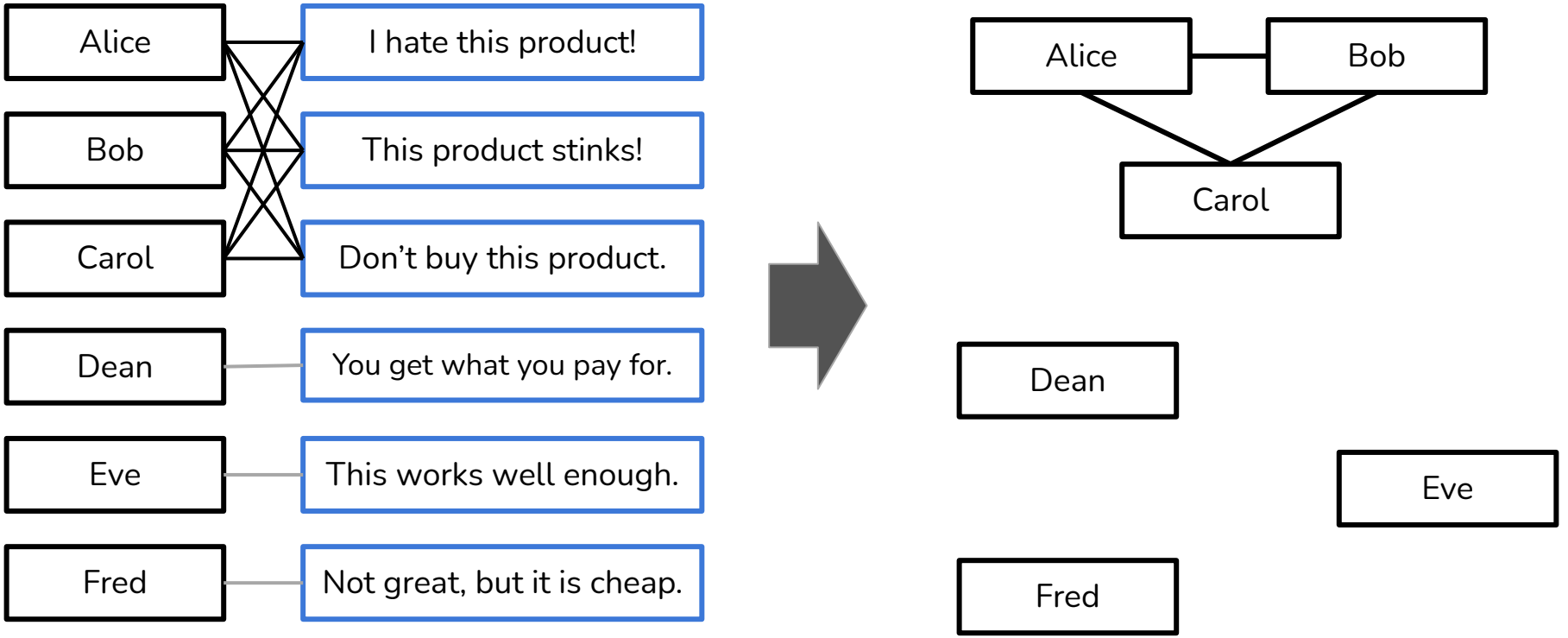
Use Case 1: Identifying reposting bots



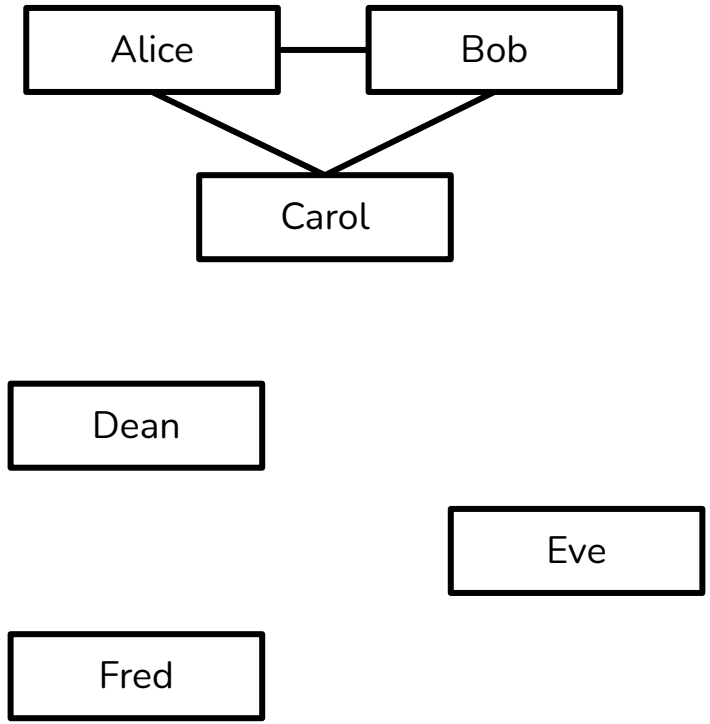
Step 1: Graph projection



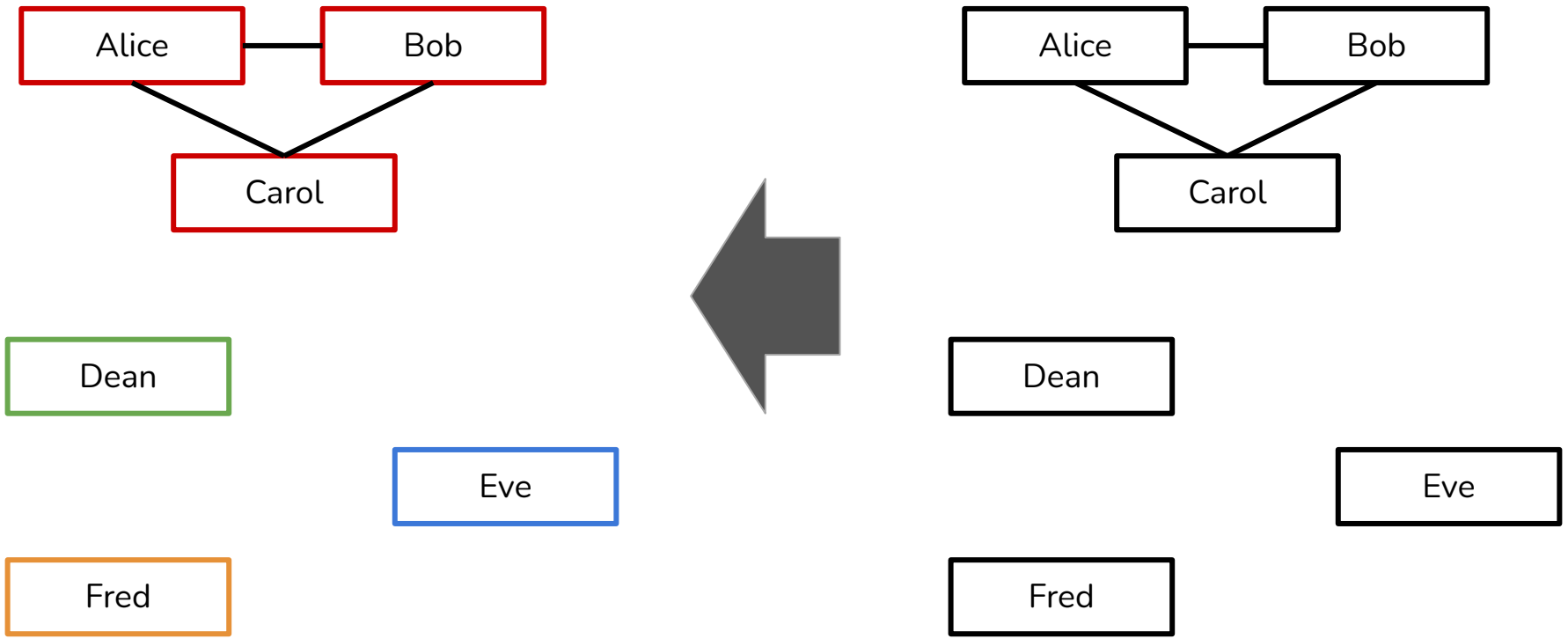
Step 1: Graph projection



Step 2: Connected components



Step 2: Connected components



Step 3: Find large connected components

Use Case 1: Identifying reposting bots

Practical Tips:

- Humans who share posts similar to those made by bots can be mistakenly labelled as bots.
- Different methods for weighing edges can lead to better results.
- Different clustering methods can lead to better results.

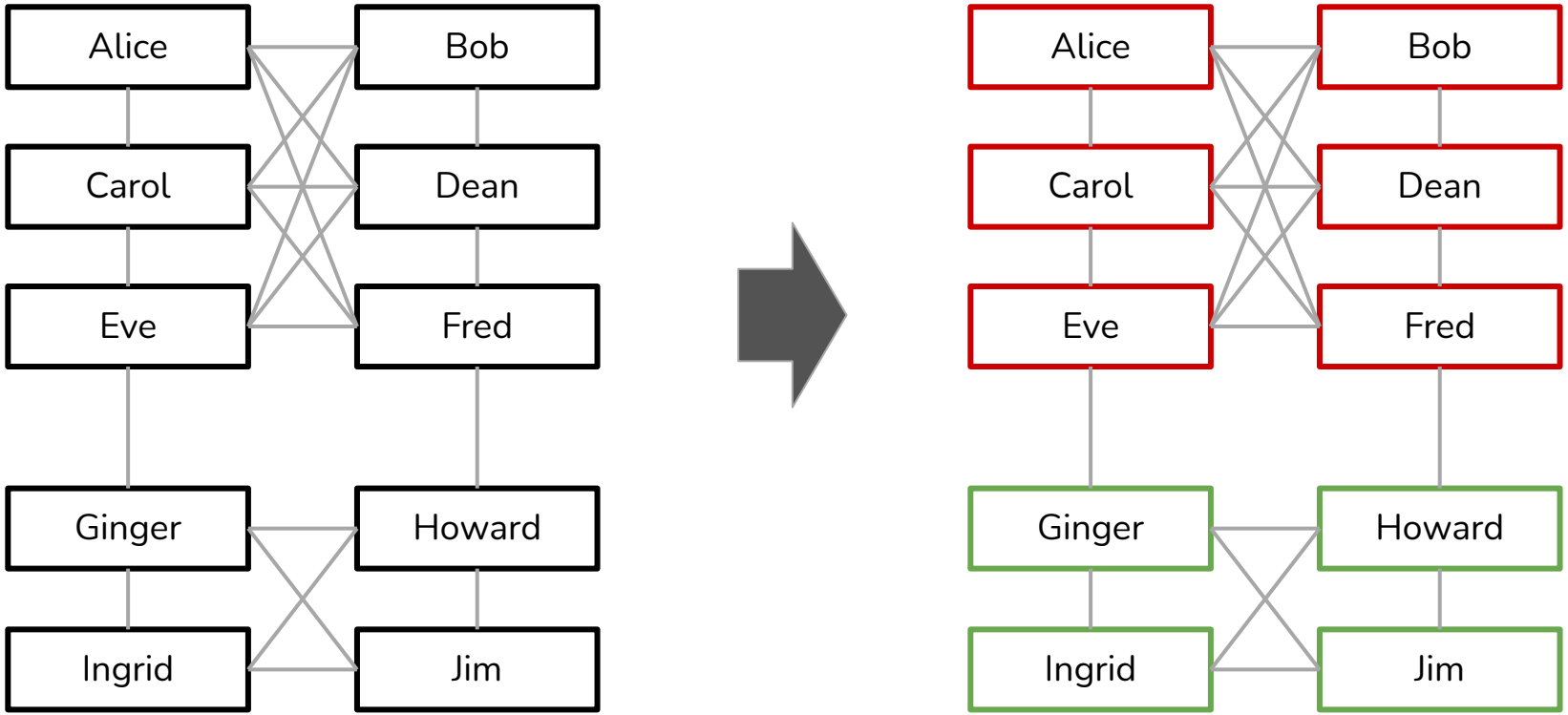
Use Case 2:
Customer clustering

Use Case 2: Customer clustering

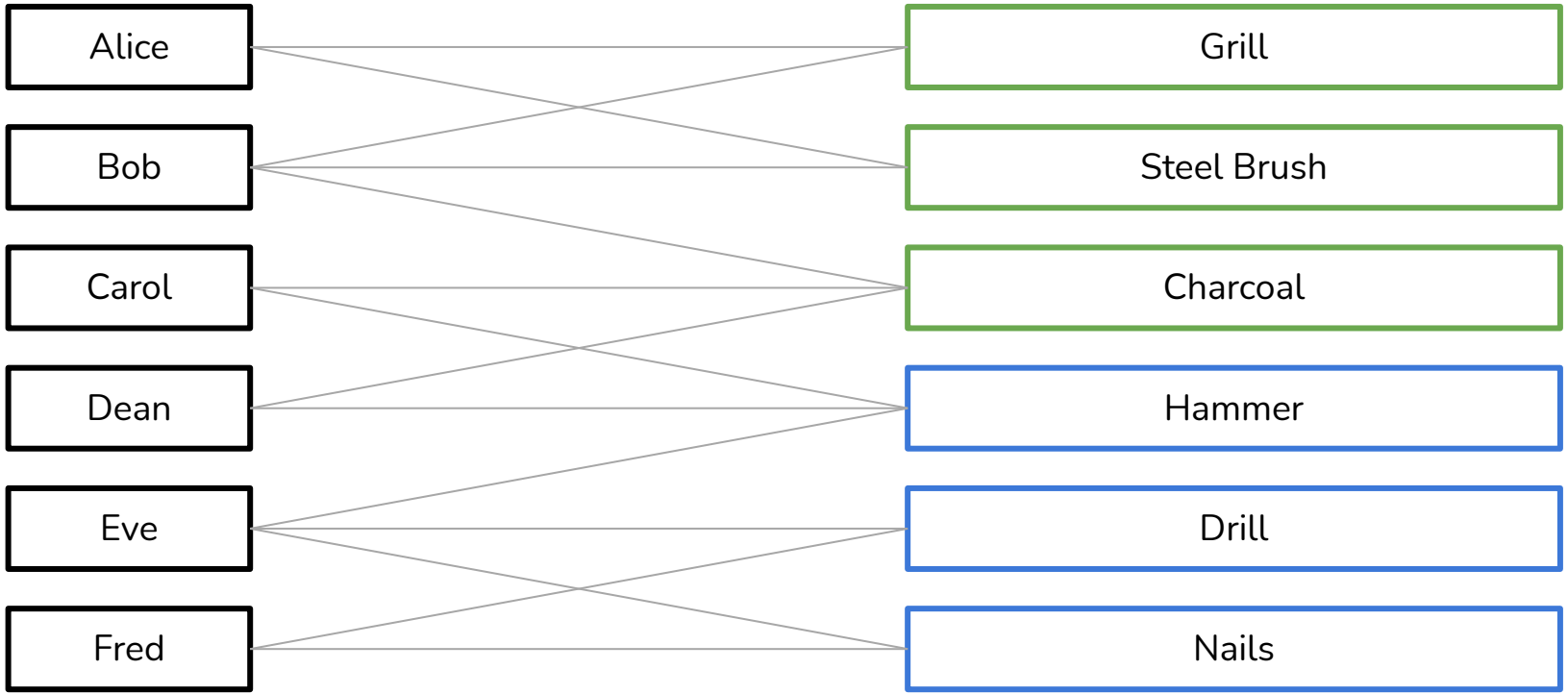
- Understanding what products buyers are likely to purchase is valuable.
- We'll show how to do this with graph analytics.
- We'll introduce 1 new algorithm: Louvain Community Detection

Louvain Community Detection

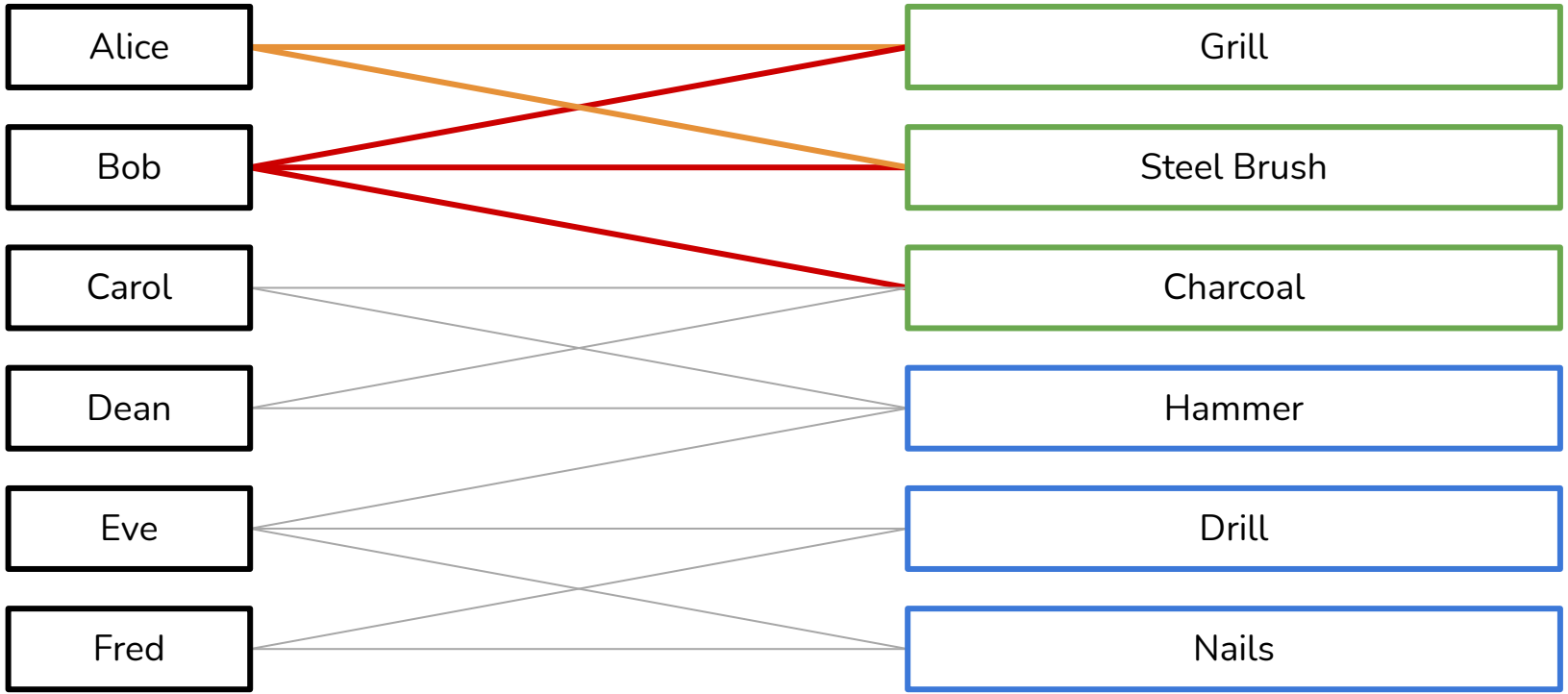
$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$



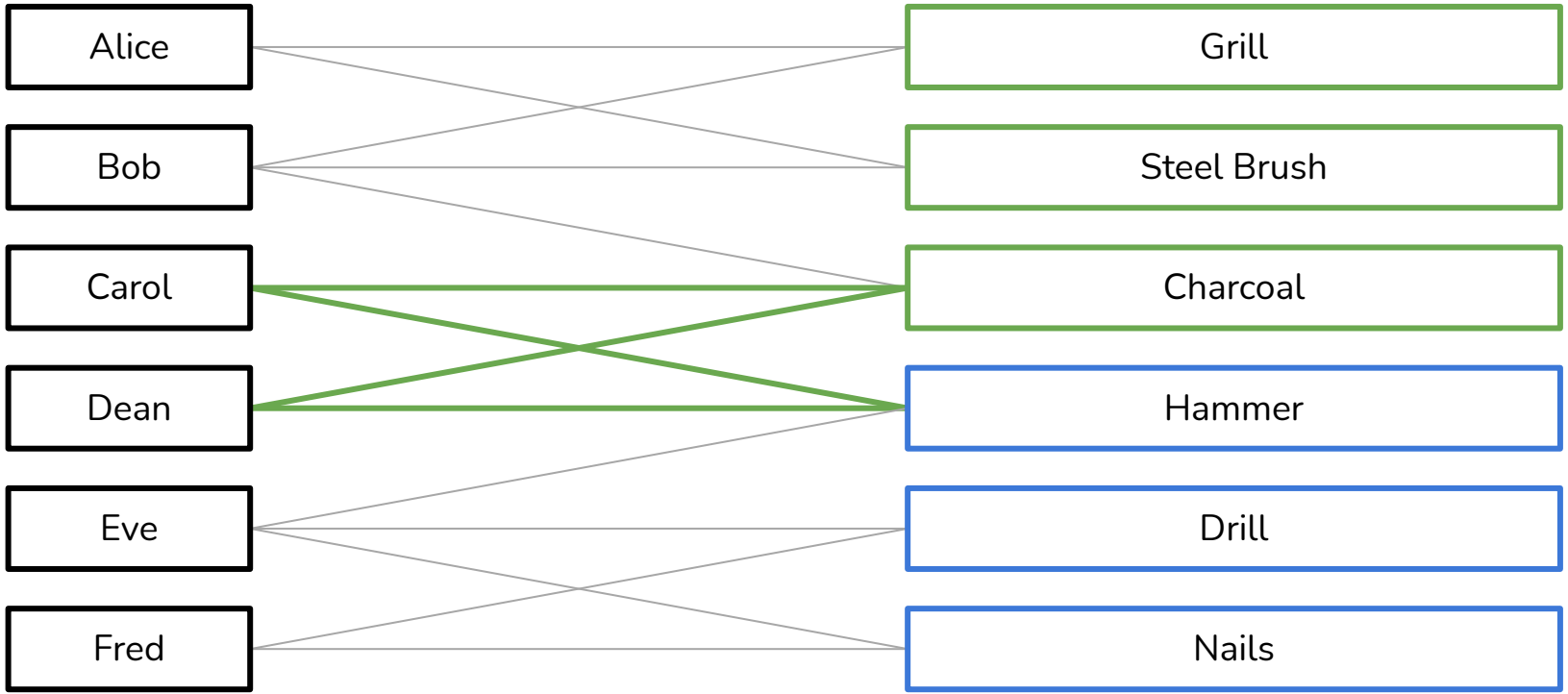
Use Case 2: Customer clustering



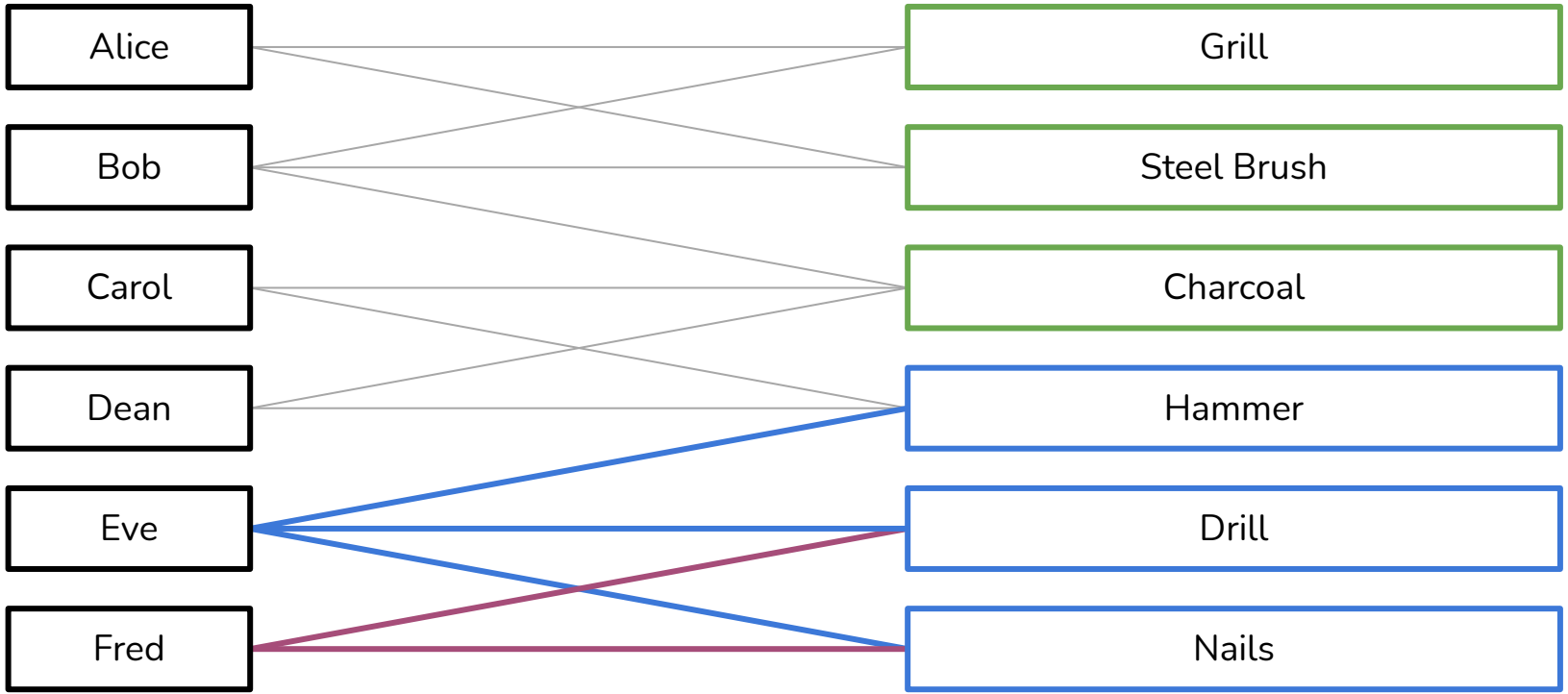
Use Case 2: Customer clustering



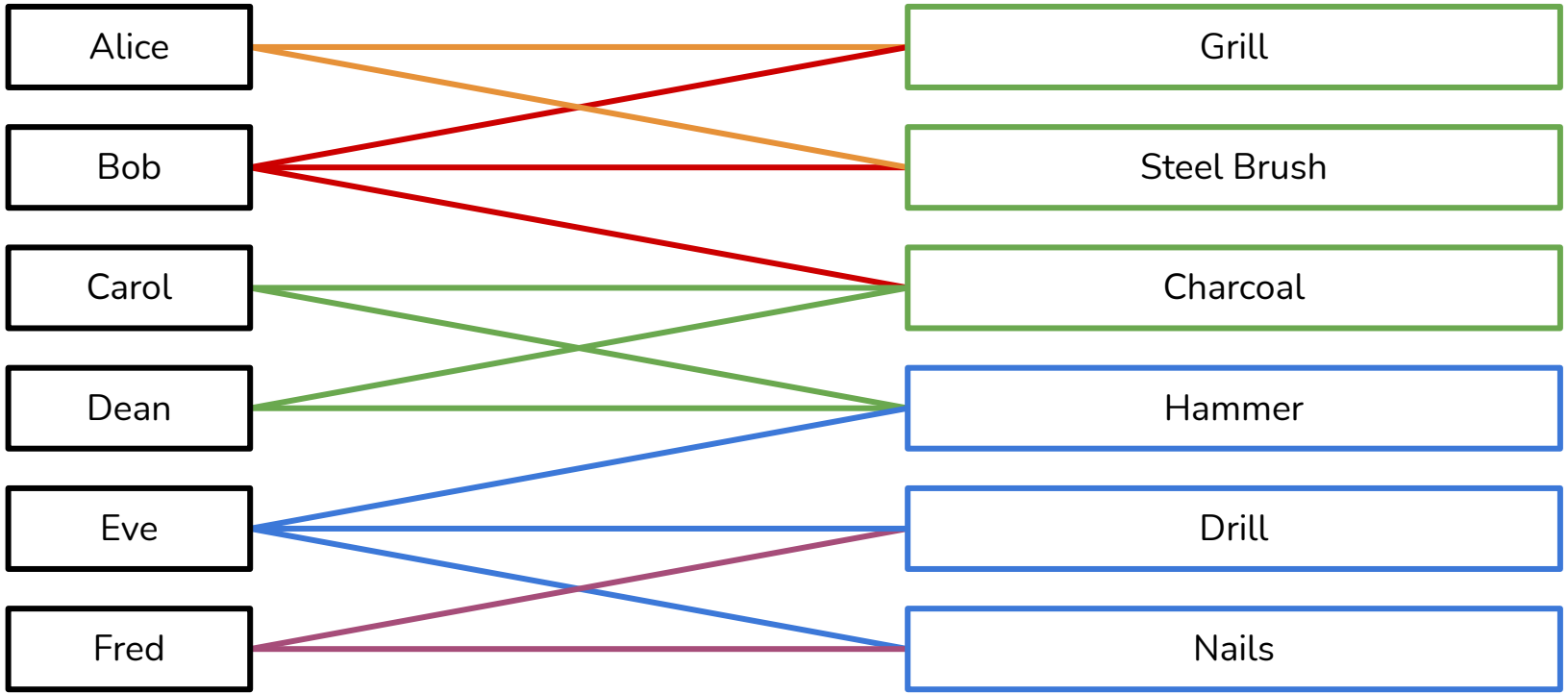
Use Case 2: Customer clustering



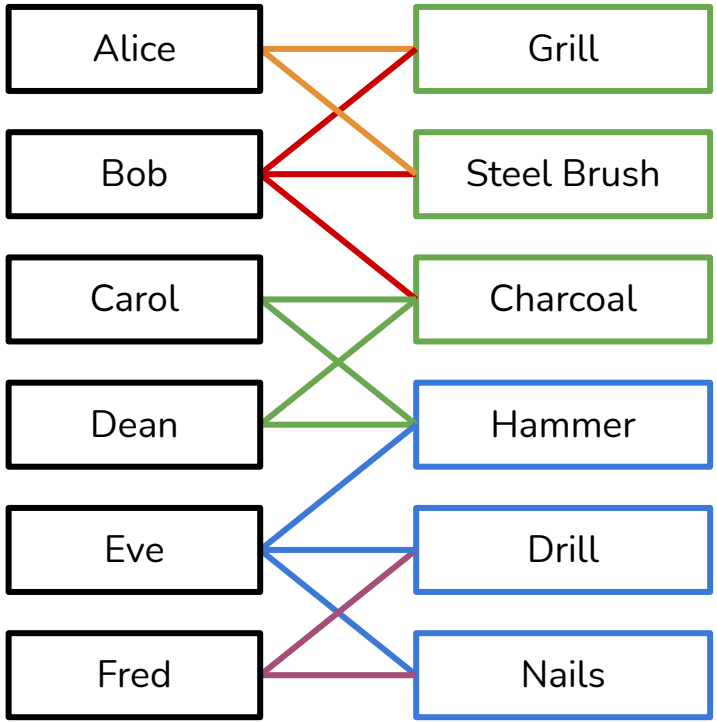
Use Case 2: Customer clustering



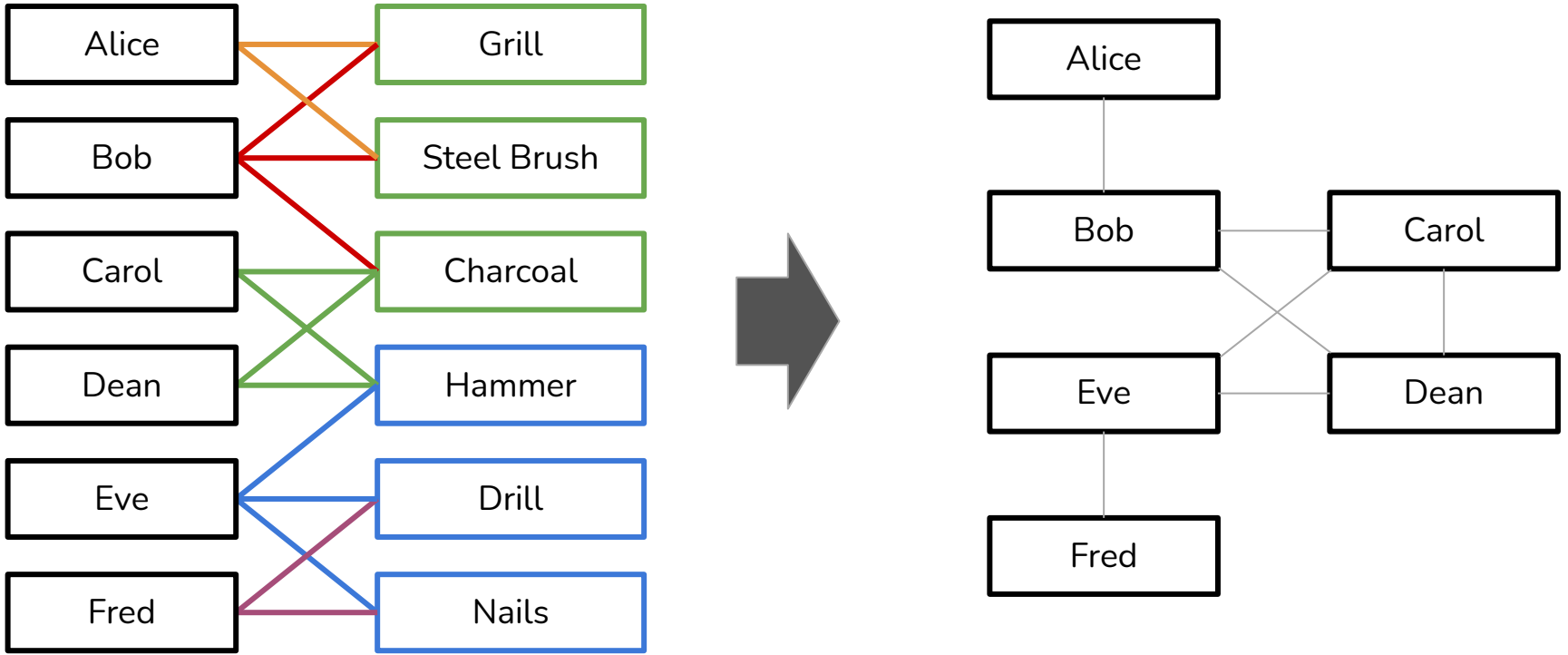
Use Case 2: Customer clustering



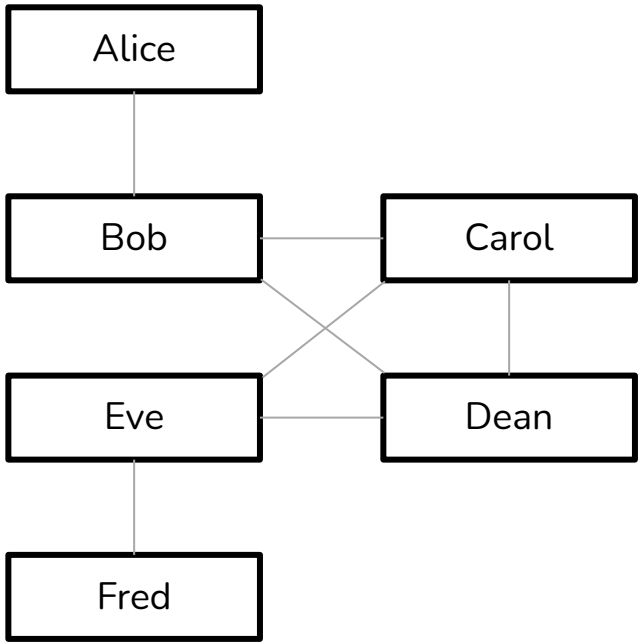
Step 1: Bipartite projection



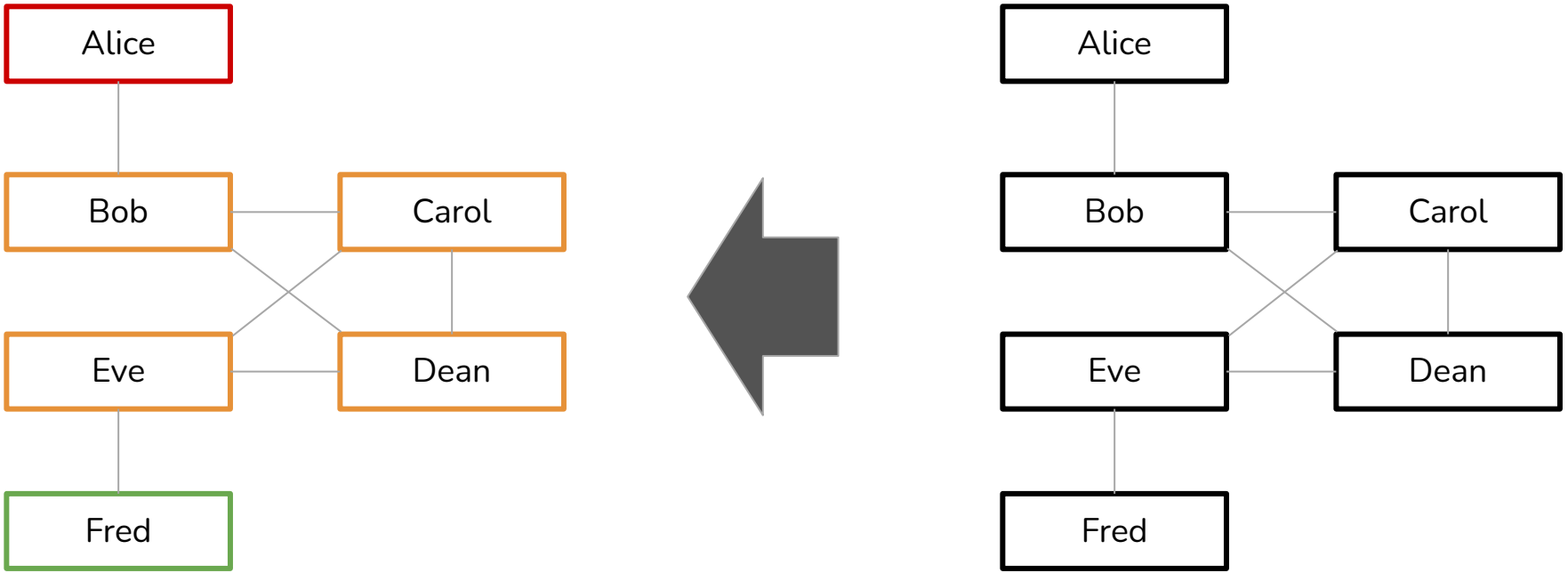
Step 1: Bipartite projection



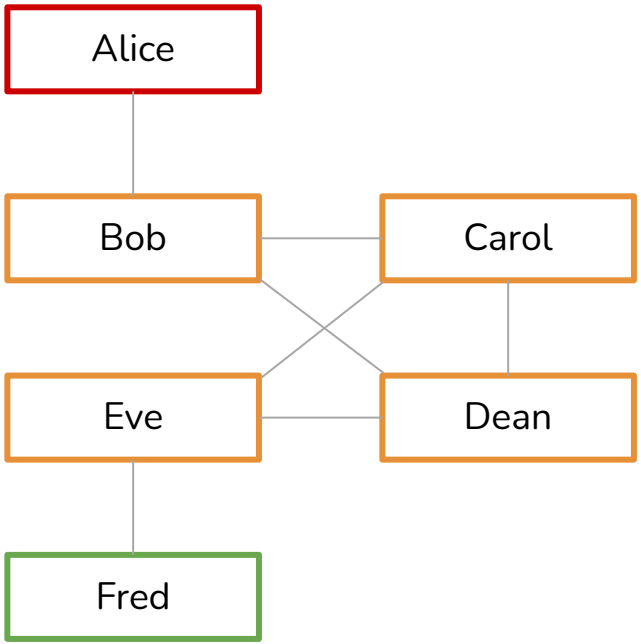
Step 2: Louvain Community Detection

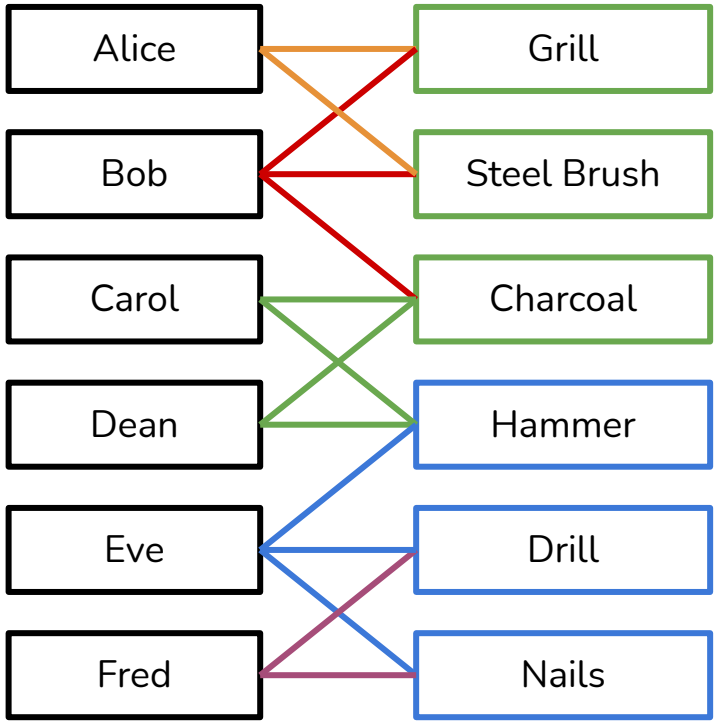


Step 2: Louvain Community Detection

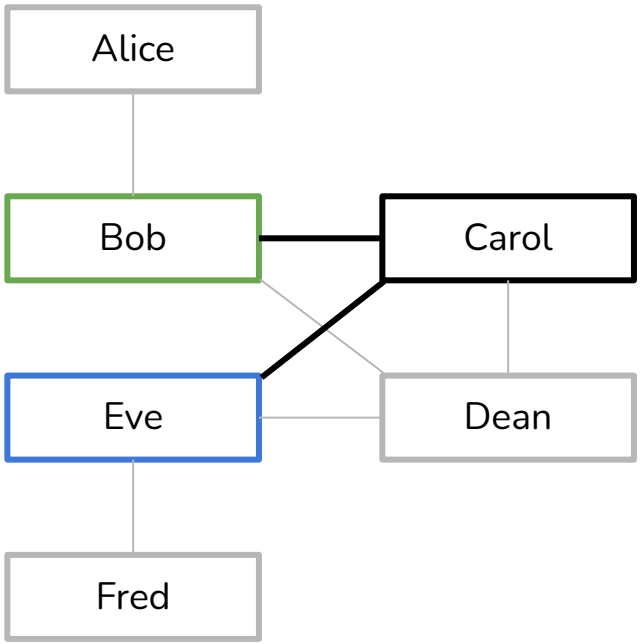


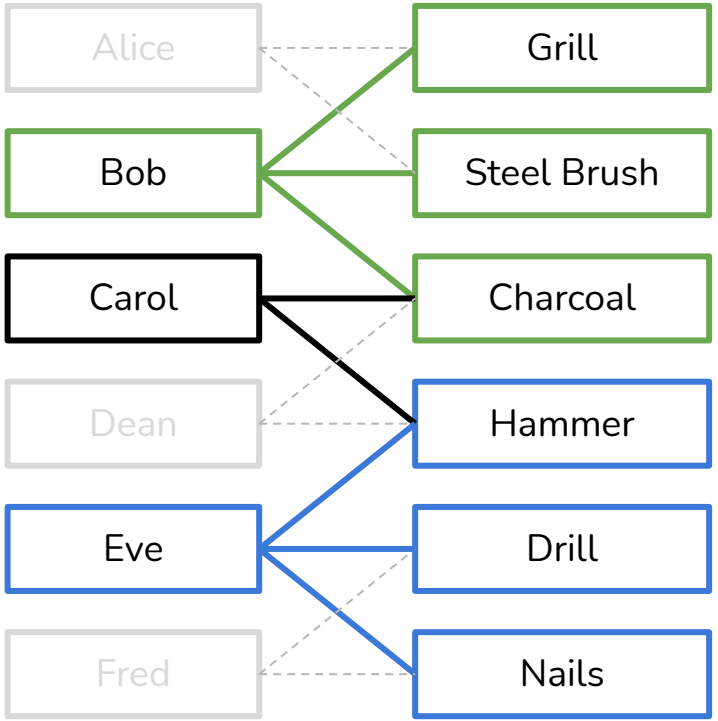
Step 3: Recommendation





Step 3: Recommendation





Use Case 3:
Identifying transportation network bottlenecks

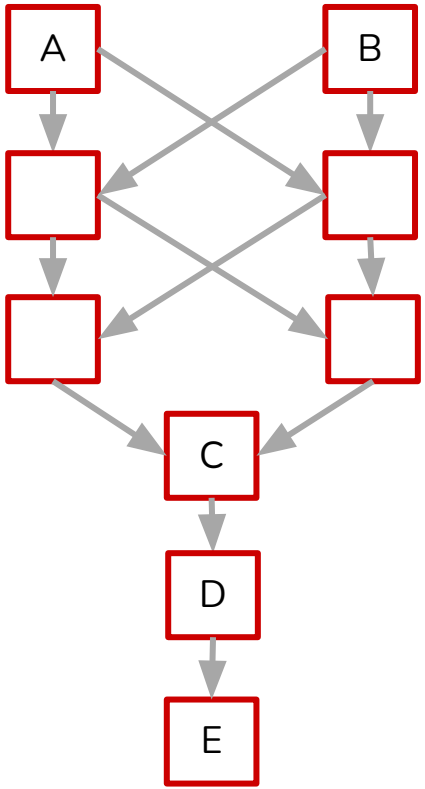
Use Case 3: Identifying transportation network bottlenecks

- Retailers seek to optimize delivery of products from distribution centers to store fronts.
- Identifying the bottlenecks can help identify regions needing improved transportation.
- We'll introduce 1 new algorithm: Betweenness Centrality
- We'll show how to implement this in Python.

Betweenness centrality

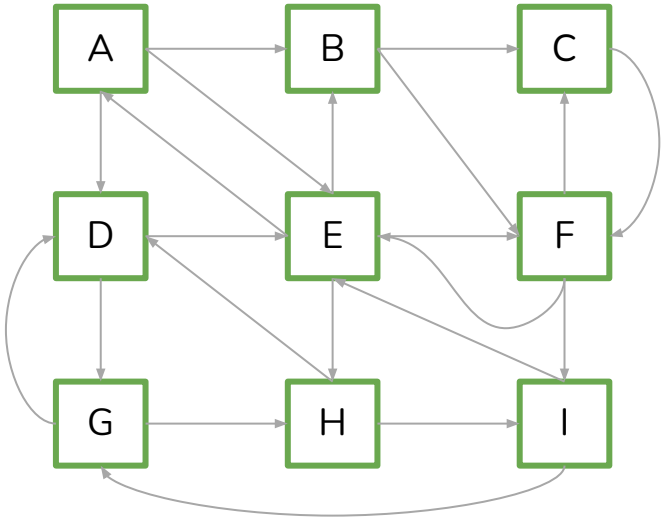
- Betweenness centrality measures how “travelled between” each node is.
- For a node v, the betweenness centrality of v is the number of shortest paths that pass through v.
- The term “shortests paths” refers to all the shortest paths between every pair of points in the graph.

Betweenness centrality



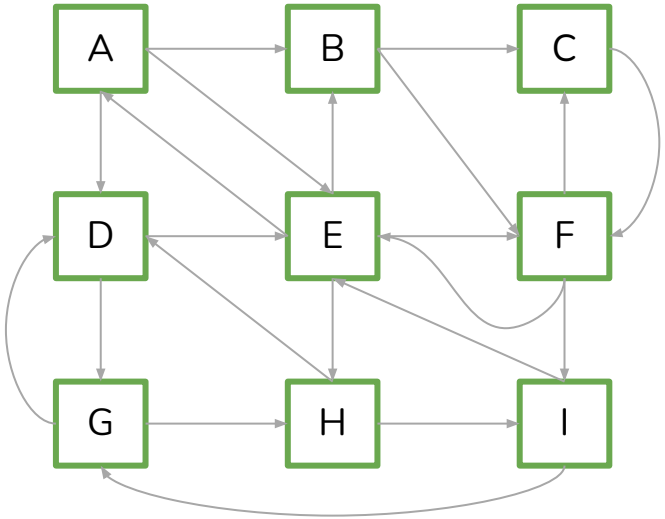
- C is part of many shortest paths to D and E.
- C has a high betweenness centrality.

Use Case 3: Identifying transportation network bottlenecks



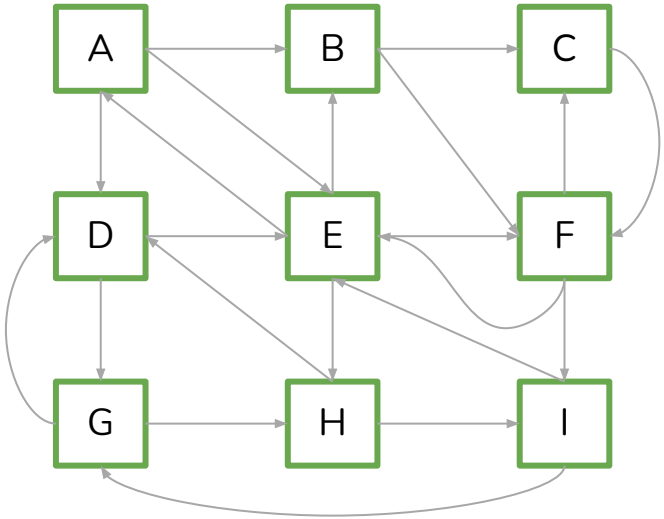
A→B: 40	D→G: 40	F→I: 42
A→E: 84	E→A: 92	D→G: 92
A→D: 78	E→B: 51	G→H: 40
B→C: 78	E→F: 62	H→D: 53
B→F: 96	E→H: 51	H→I: 47
C→F: 55	F→C: 50	I→E: 50
D→E: 97	F→E: 55	I→G: 49

Use Case 3: Identifying transportation network bottlenecks



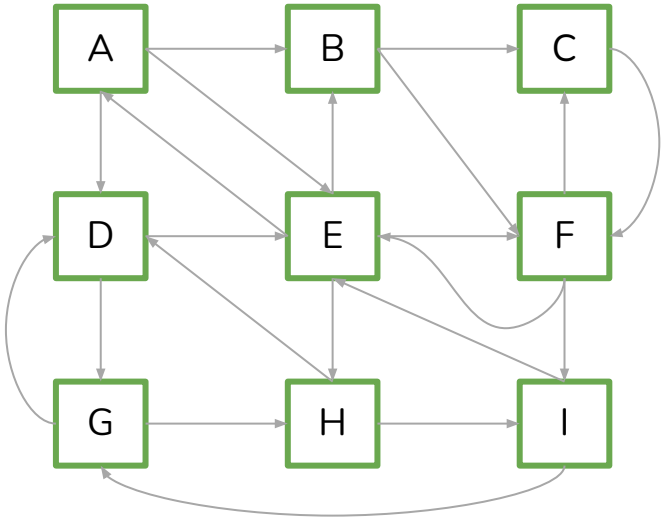

```
>>> import pandas as pd
>>>
>>> df = pd.DataFrame(...)
>>> df
   src dst  weight
0    A   B      40
1    A   E      84
..  ..  ..     ...
20   I   G      49
```

Use Case 3: Identifying transportation network bottlenecks




```
>>> df
   src dst  weight
0    A   B      40
..  ..  ..     ...
20   I   G      49
>>>
>>> import networkx as nx
>>>
>>> graph = nx.from_pandas_edgelist(df,
'src', 'dst', 'weight')
<networkx.classes.graph.Graph object at
0x7fe4af19ffd0>
>>>
>>> graph['A']['B']['weight']
40
```


Use Case 3: Identifying transportation network bottlenecks




```
>>> nx.betweenness_centrality(graph,
normalized=False, weight='weight')
{'A': 1.0,
 'B': 4.0,
 'E': 4.0,
 'D': 2.0,
 'C': 0.0,
 'F': 5.0,
 'G': 0.0,
 'H': 5.0,
 'I': 6.0}
>>>
```


Use Case 3: Identifying transportation network bottlenecks

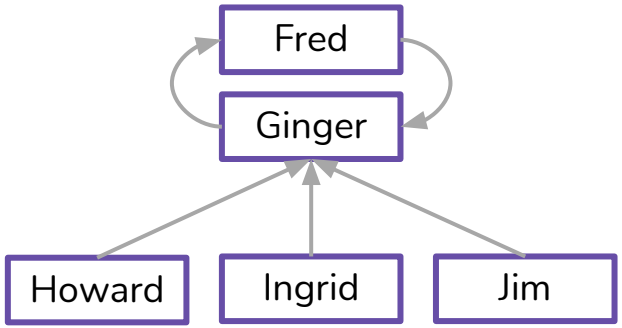
- There are many different measures of “importance” of a node that we could’ve also used that we didn’t cover here, e.g. PageRank and Eigenvalue Centrality.
- Also worth looking into are Max-Flow algorithm variants.

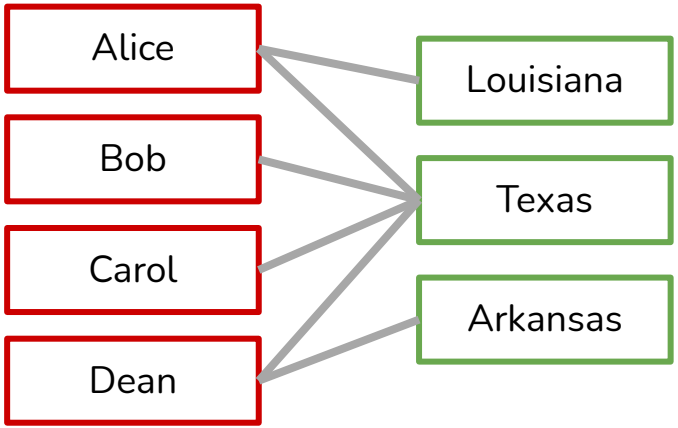
Practical Graph Analytics Tips

Practical tips on when/how to use graph analytics

When to use graph analytics:

- The most intuitive way to visually present the data is on a whiteboard as a graph or network.

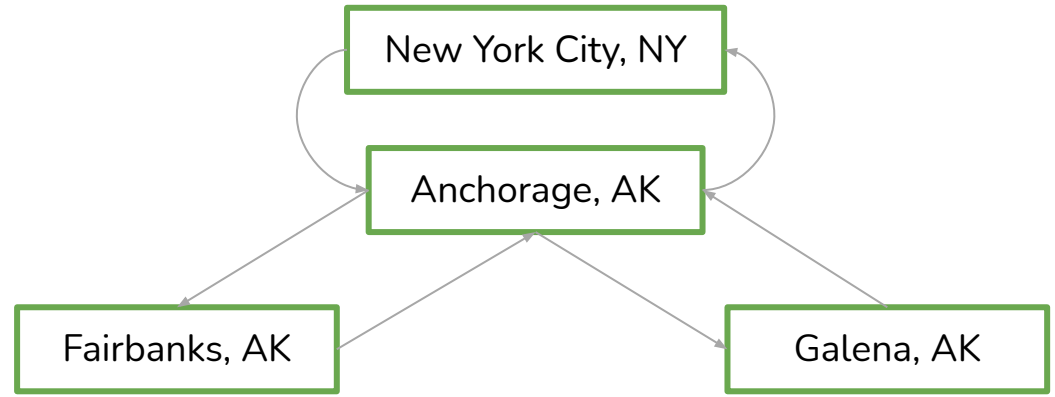




Practical tips on when/how to use graph analytics

When to use graph analytics:

- If you take columns of a table and can potentially view those as an edgelist, then a graph representation might be appropriate.

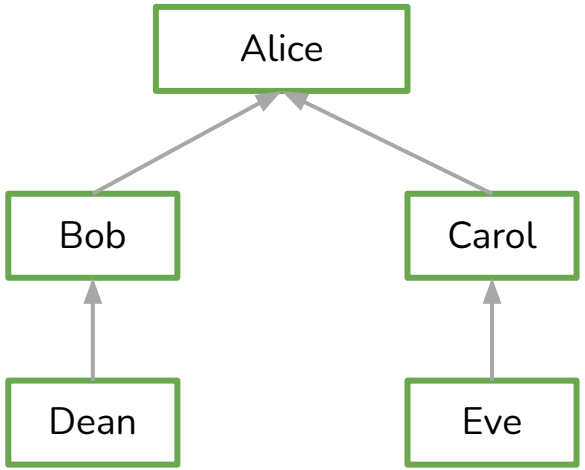


Source	Destination
NYC	Anchorage
Anchorage	NYC
Anchorage	Fairbanks
Fairbanks	Anchorage
Anchorage	Galena
Galena	Anchorage

Practical tips on when/how to use graph analytics

When to use graph analytics:

- There's interdependence between data points.

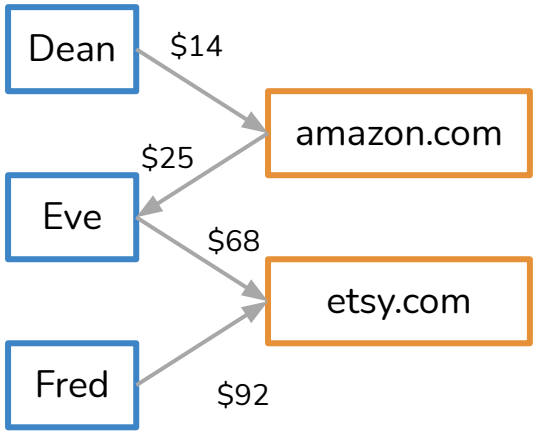


	Name	Age	Location	Manager
0	Alice	21	Texas	NULL
1	Bob	33	Louisiana	Alice
2	Carol	33	Texas	Alice
3	Dean	37	Florida	Bob
4	Eve	22	Texas	Carol

Practical tips on when/how to use graph analytics

When to use graph analytics:

- In a normalized DB, different tables can correspond to different partitions of a bipartite/multipartite graph.



	Name	Age	Location
0	Dean	21	AR
1	Eve	37	TX
2	Fred	27	LA

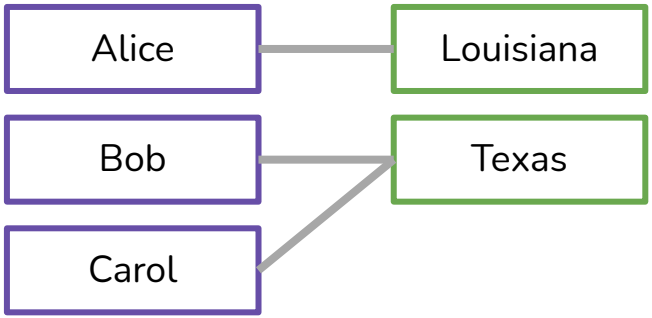
	Retailer
0	amazon.com
1	etsy.com

	Retailer	Customer	Amount
0	amazon.com	Dean	14
1	amazon.com	Eve	-25
2	etsy.com	Eve	68
3	etsy.com	Fred	92

Practical tips on when/how to use graph analytics

When to use graph analytics:

- A categorical column can also indicate a bipartite relationship.

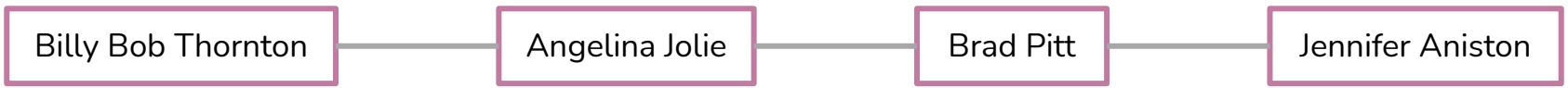


	Resident	State	Age
0	Alice	Louisiana	38
1	Bob	Texas	35
2	Carol	Texas	29

Practical tips on when/how to use graph analytics

- Sometimes, the graph algorithms might get you some insight, but not insight that's necessarily relevant towards your goal.

	Actor	Spouse	Marriage Year
0	Billy Bob Thornton	Angelina Jolie	2000
1	Angelina Jolie	Brad Pitt	2014
2	Brad Pitt	Jennifer Aniston	2000



Resources to Dive Deeper

Resources to Dive Deeper

- For the uninitiated, introductory algorithms textbooks:
  - Introduction to Algorithms - Cormen, Leiserson, Rivest, Stein
  - Algorithm Design - Kleinberg & Tardos
- Graph Algorithms in Apache Spark and Neo4j - Needham & Hodler
  - https://neo4j.com/blog/new-oreilly-book-graph-algorithms-spark-neo4j/
- Neo4j Graph Gists: https://neo4j.com/graphgists/
- NetworkX: https://networkx.org/
- Julian Shun: https://people.csail.mit.edu/jshun/graph.shtml