# News Clustering Project

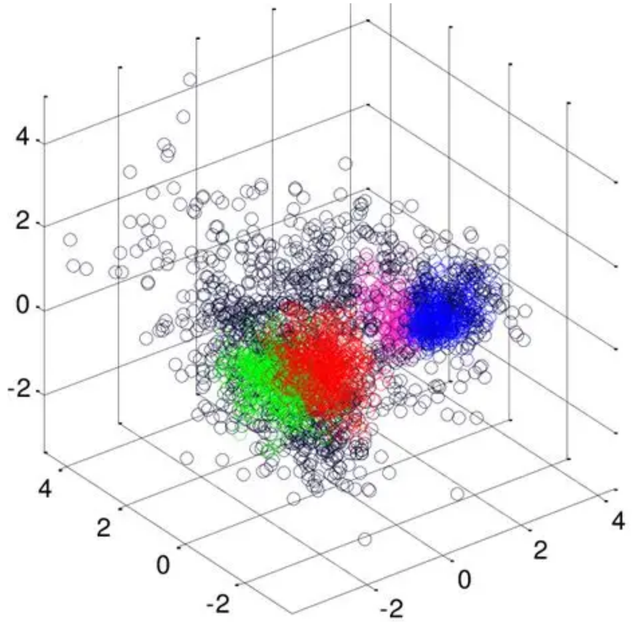
## Bundle AI News Clustering Pipeline

Ata Tan Dağıdır - Intern

Bundle

# Problem: Online Event Clustering

- We have a categorization system but no event clustering system.
- Event Clustering (all the news about an event)




Bundle

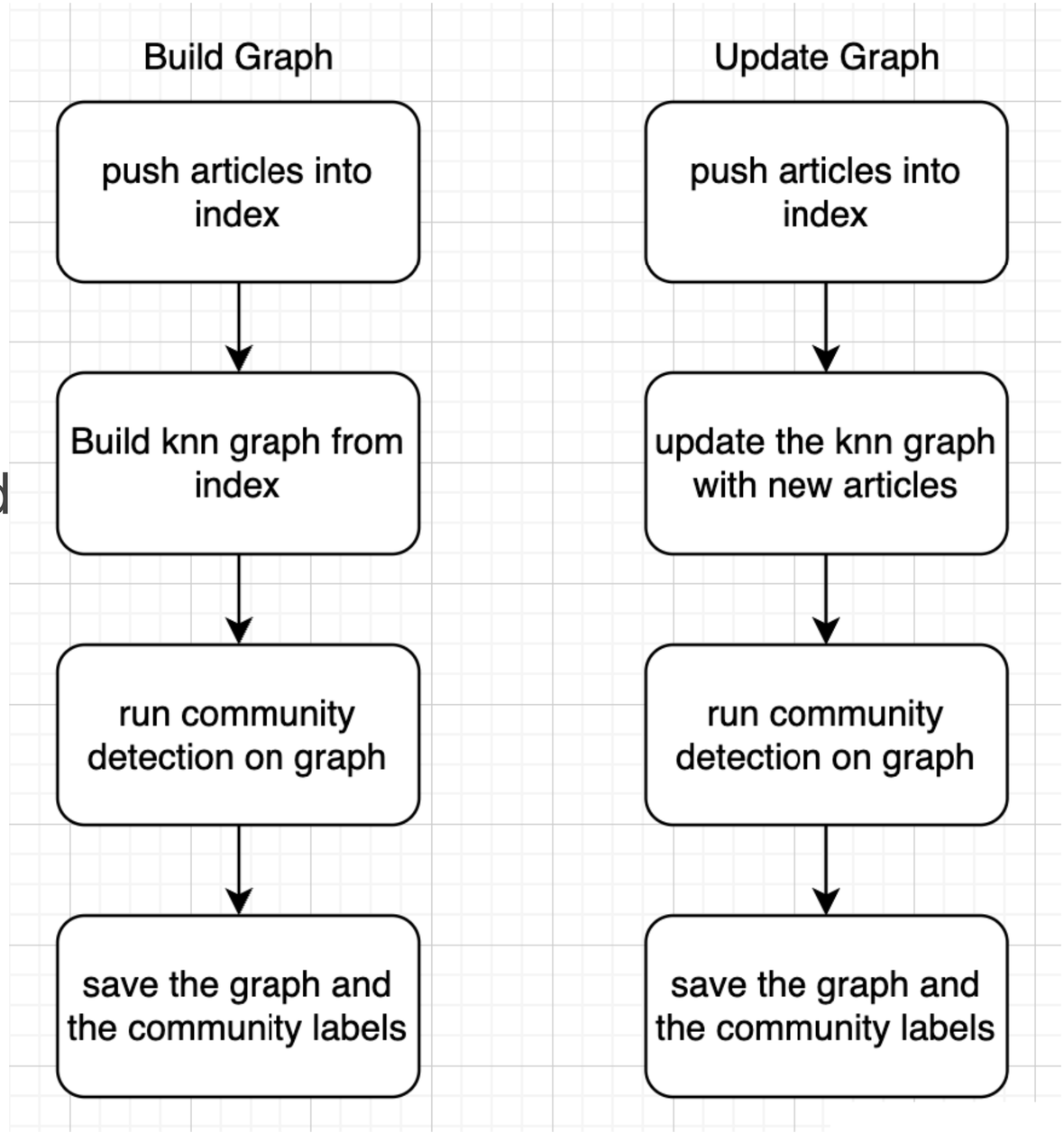
# Potential Benefits

- Event Clustering, avoiding duplicates on feed
- Breaking news detection
- Content-Based Recommendation Engine (and personalized ads, NO user data needed)
- Fake news elimination (flag events as fake)
- Insight Mining (for businesses)
- Listing events about entities or other features etc.

# Our Solution

## KNN Graph + Community Detection

- Existing infrastructure (Pinecone, Neo4j, leidenalg, igraph)
- Events are irregular in size (k and centroid methods fail)
- Blazingly fast clustering (O(logN) for online)
- Scalability(DBSCAN is unstable in high-dim, curse of dimensionality)
- Delete articles weekly(?)

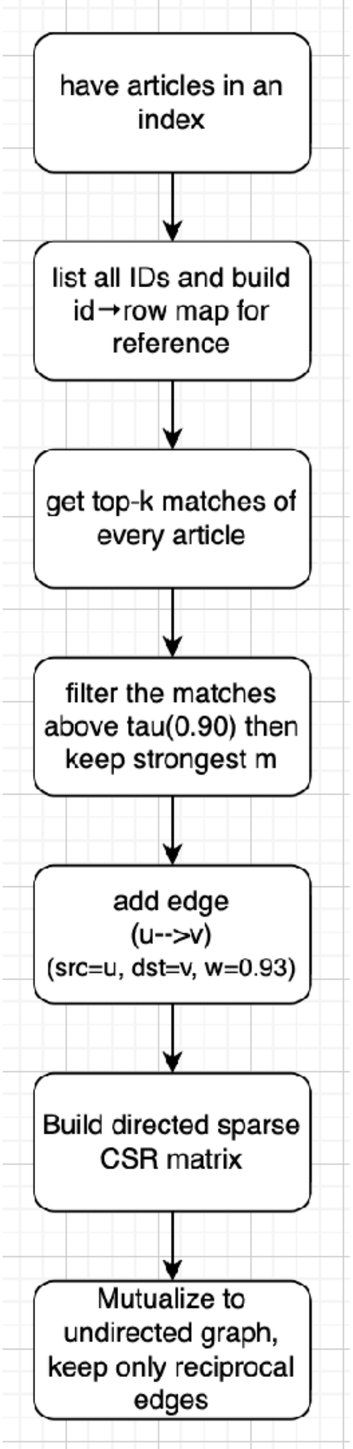



Bundle

# Building The Graph

## k-NN Graph

- Leveraging existing DBs systems
- HNSW indexing, fast queries
- Keep the CSR matrix in a graph DB or local file (npz)
- O(NlogN + Nk + Nm) -> O(N logN) time complexity
- O(N*m) memory complexity -> 100,000 articles, 30MB



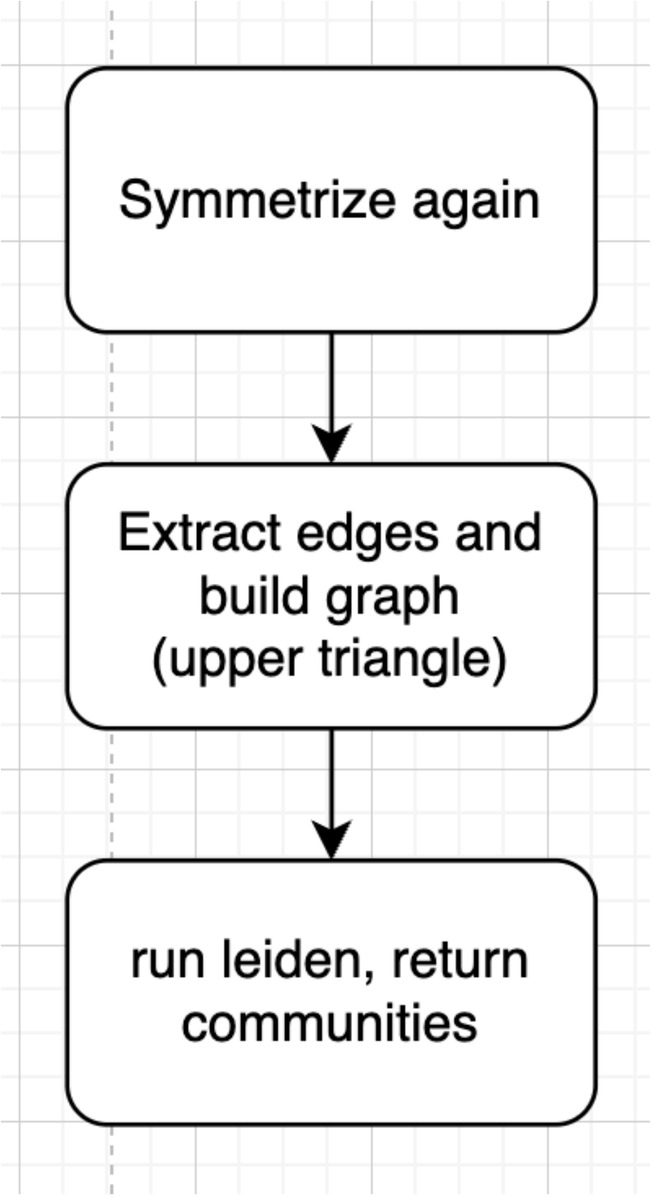

Bundle

# Clustering: Community Detection

## Leiden Algorithm

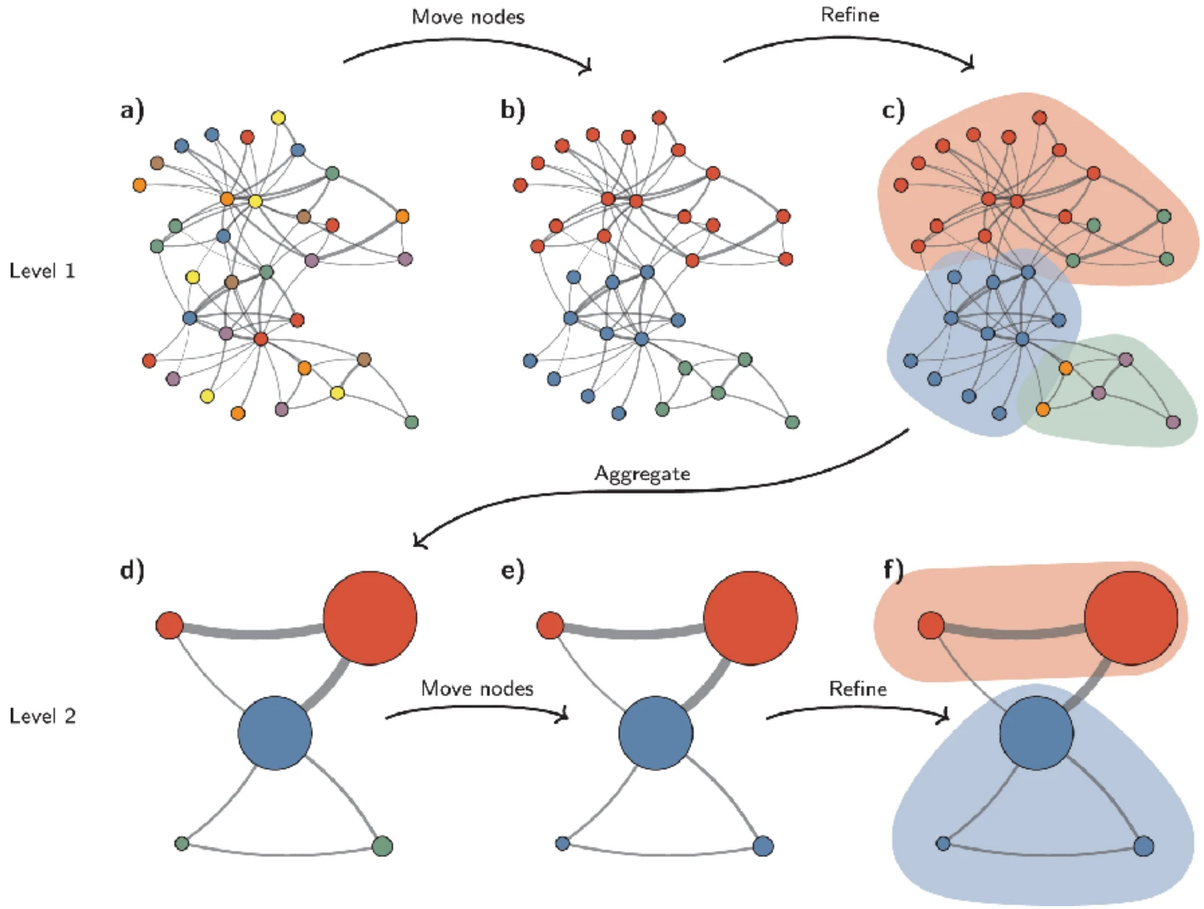
- Resolution Parameter (sensitivity)
- O(E) -> O(Nm) time complexity for our sparse graph, almost instant in practice (0.05s)
- Hungarian Matching for keeping previous community ids
- leidenalg, igraph

```
part = la.find_partition(
    g,
    la.RBConfigurationVertexPartition,
    weights="weight",
    resolution_parameter=resolution,
    seed=seed,
    n_iterations=-1,
    initial_membership=init
)
```



# Leiden Algorithm

- Modularity: $Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$ Compare actual edges inside communities vs random
- Goal: find groups with more internal edges than random expectation.

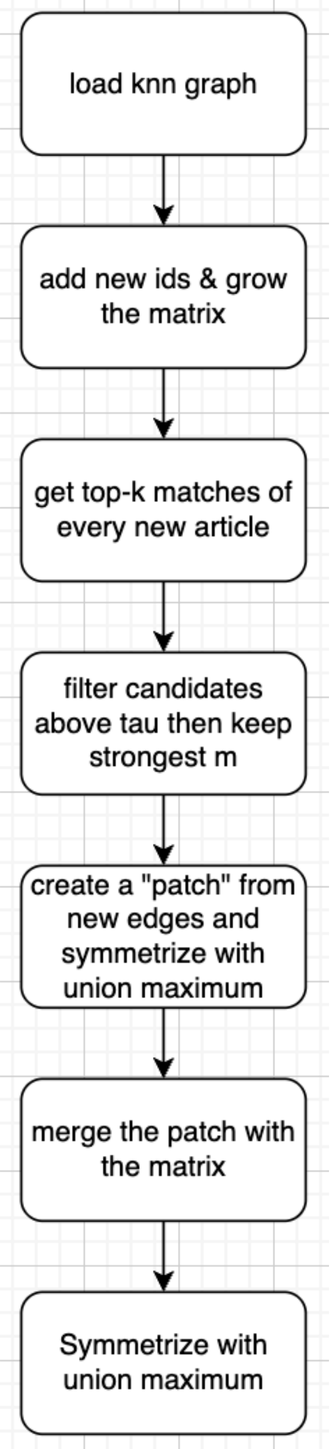



Bundle

# Online Clustering

## Updating the k-NN Graph

- Neo4j abstraction vs current system
- Very math-heavy, hard-to-maintain code
- O(b*logN) time complexity (b=50)




Bundle

# Current System

## Pinecone Cloud as Index

- **Bottleneck**: 140ms per query due to network latency (parallelization, local index)
- **Unstable Indexing**: Pinecone Cloud takes unreliable times to index.
  Every time we INCREMENT new vectors into the Pinecone index, it takes a while to get indexed, and there is no clear indication how long does it take to build this index.
  **https://community.pinecone.io/t/pinecone-takes-about-2-3-minutes-for-upsert-vector-to-be-available-for-query/5701**
- **Workaround "Fix"**: green-index, blue-index, time.wait

# Pinecone Bug

```python
print("Deleting existing data in green index...")
green_index.delete(namespace=namespace, delete_all=True)

dl.upsert_data_to_pinecone(initial_data, green_index, namespace)
time.sleep(15) # the pinecone bug
# Different tau
A, id2row, ids = gb.build_knn_graph_pinecone(green_index, namespace, k_search=k_search, m=max_neighbors, tau=tau, metric=metric)
labels = cluster.run_leiden(A, resolution=1.0, seed=123, initial_labels=None)
dl.save_snapshot(snapshot_file, A, ids, labels)

print("Graph built and saved\nMoving on to incremental update...")

# Extract IDs from new documents
new_records = json.load(open(new_data))
new_ids = [d['_id'] for d in new_records if d['country_id'] != 228]

dl.upsert_data_to_pinecone(new_data, green_index, namespace)
print("Deleting existing data in blue index...")
blue_index.delete(namespace=namespace, delete_all=True) # This is weird logic caused by the pinecone bug

dl.upsert_data_to_pinecone(new_all_data, blue_index, namespace)
time.sleep(15) # the pinecone bug

print("updating snapshot...")
A_curr, curr_id2row, curr_ids, curr_labels = inc.update_knn_snapshot_pinecone(snapshot_file, blue_index, namespace, new_ids, k_se

new_labels = cluster.run_leiden(A_curr, resolution=1.0, seed=123, initial_labels=curr_labels)

dl.save_snapshot(snapshot_file, A_curr, curr_ids, new_labels)
```

Clean Start

Unstable Indexing Bug

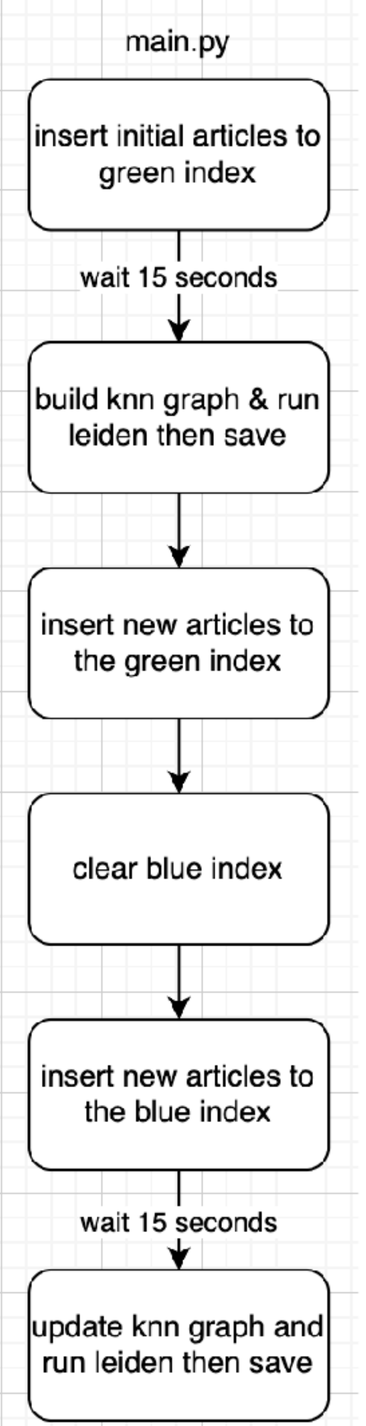
We first insert to green index

We insert only new to green index

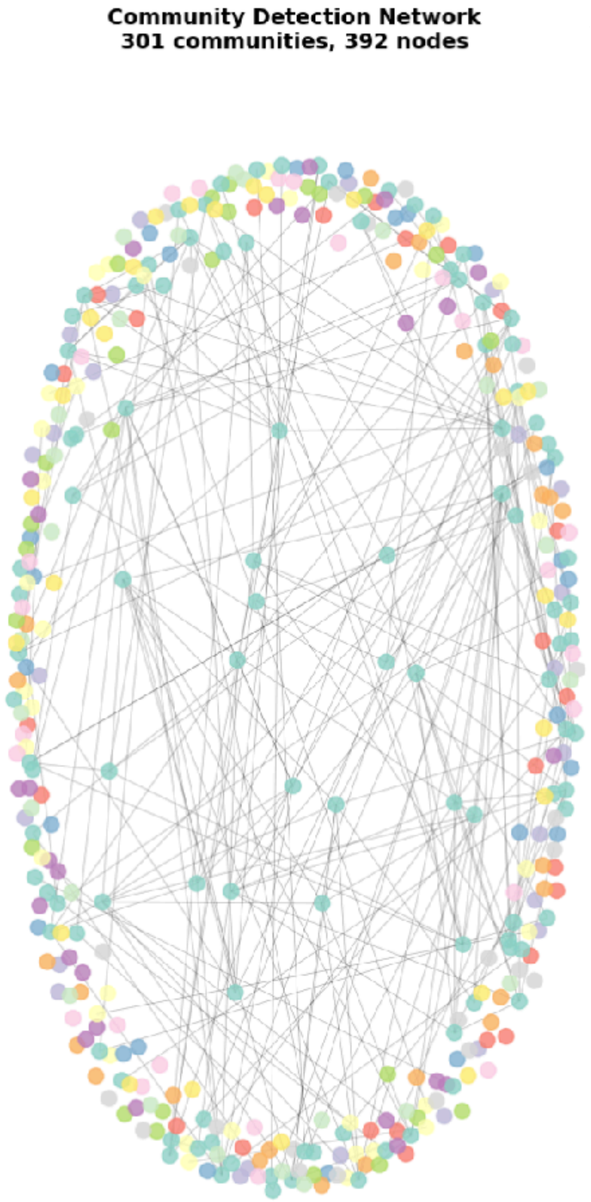
Clean start

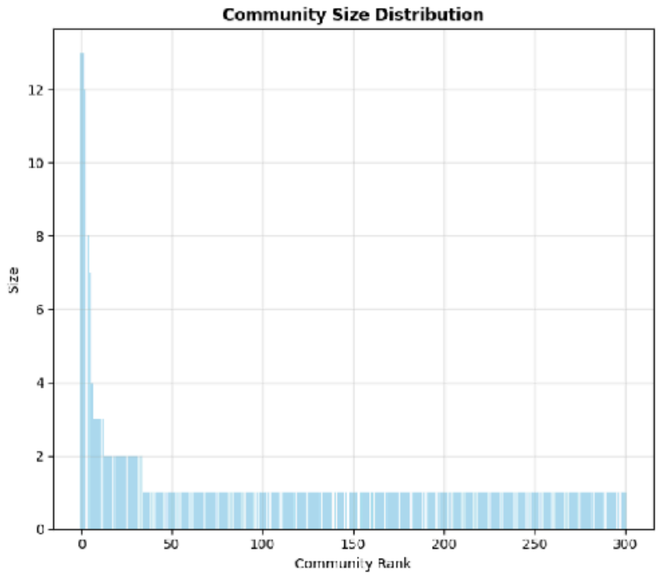
Then we insert new + old to green index

We use newly built blue index to update the knn



# Statistics




**Top Communities by Size**

| Community | Size | Dominant Category | Sample Text |
|---|---|---|---|
| Community 1 | 13 | Unknown | |
| Community 0 | 13 | Unknown | |
| Community 2 | 12 | Unknown | |
| Community 3 | 8 | Unknown | |
| Community 4 | 8 | Unknown | |
| Community 5 | 7 | Unknown | |
| Community 6 | 4 | Unknown | |
| Community 7 | 3 | Unknown | |
| Community 8 | 3 | Unknown | |
| Community 10 | 3 | Unknown | |

```
Graph Statistics:

Nodes: 392
Edges: 212
Communities: 301
Modularity: 0.870
Avg Clustering: 0.170
Density: 0.003

Largest Community: 13 nodes
Smallest Community: 1 nodes
Average Community Size: 1.3
```

# Sample Results

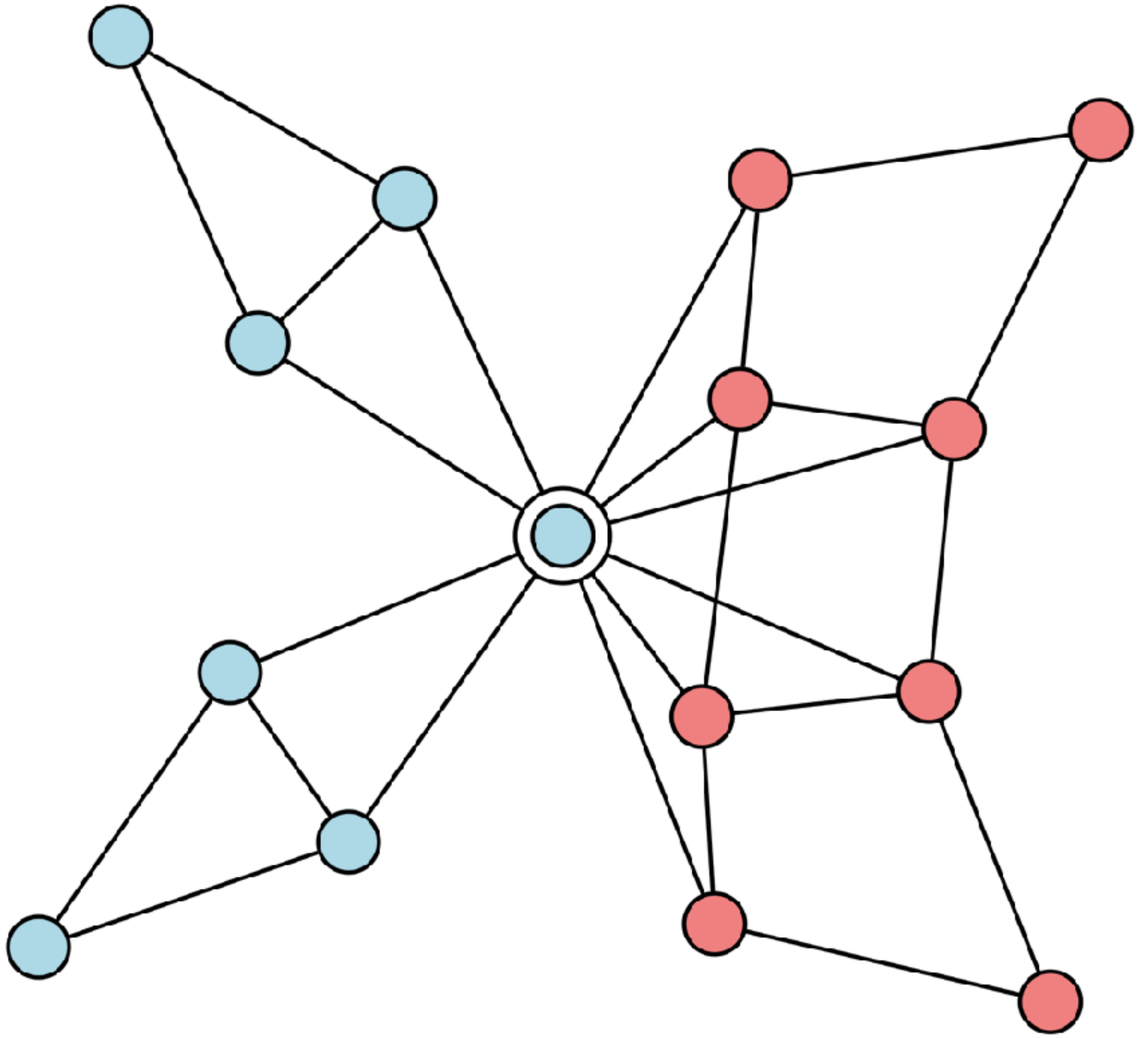
```
Community 0 (13 members):
--------------------------------------
  • 38393D06-125E-48FC-BA76-A180AE9D6C10: Meghan Markle's Netflix series With Love, Meghan returns for a second season that closely fol
  • 4AB0ABB2-30B3-41C3-8691-A1603411F254: Netflix released season 2 of Meghan Markle's lifestyle series With Love, Meghan, featuring ce
  • 4B951B68-2004-4A57-8A7C-32BE641E11CC: Meghan Markle discusses the early days of her relationship with Prince Harry while promoting
  • 5A075231-A052-44C1-84DB-989656CCEFBF: Prince Harry, who briefly appeared at the end of season one of Meghan Markle's Netflix series
  • 6A0467F3-F347-4E1A-9F82-1932E9E9743E: Meghan Markle said she became "not well" after spending nearly three weeks apart from her chi
  • 6F59EB35-274E-4F70-A373-6993135E21B7: Meghan Markle showcased her backyard garden and prepared apple butter in the fourth episode o
  • 8ABFBDF5-7F5C-4BAA-B2BA-F67D58F4DB69: Meghan Markle says she misses the UK three years after stepping back from royal duties and mo
  • 8EA69A2B-2F75-44A2-AFED-4664DB7B3C56: Meghan Markle's lifestyle series With Love returned for a second season on Netflix, featuring
  • 99C95E51-C3CB-483B-AD66-C6BC4C6A4127: Meghan Markle revealed on the new season of her Netflix show With Love that Prince Harry was
  • 9E92B6C9-A486-4443-A31C-B43EFDA12066: Meghan Markle has rejected claims that her Netflix lifestyle series glamorizes the "trad wife
```

```
Community 6 (4 members):
--------------------------------------
  • 166E7BB7-9E61-45EA-9F0A-EE25ECCF4CED: Hundreds of Israeli demonstrators gathered near Tel Aviv on Tuesday, August 26, calling for ar
  • 267DAD75-3F15-4D26-832C-BEE890A67F80: Israel faces mounting domestic pressure over the fate of hostages and the conduct of its Gaza
  • 456F987A-E829-4B6B-81F1-2DDE84CD7975:(new) Israeli protesters in Tel Aviv and other cities set fires, blocked roads and held large
  • A1D2FEB6-9325-4F2F-9F17-567289E6532A:(new) Massive protests erupted across Israel as families of hostages held by militants demande
```

# Future Upgrades

- Hungarian Matching (SciPy package)
- Neo4j
- Testing the clustering quality over time with huge data
- Deployment options(rebuild triggers/cron etc.)

# Thank You



- Ata Tan Dağıdır

Bundle