Project Implementation Plan

Comparative Analysis of Graph Clustering Features

Team: Leonardo Gusson, Luca Rao, Chiara Frizzarin

1. Role Division

Chiara *Data & Integration*	**Luca** *Graph Engineering*	**Leonardo** *ML & Experiments*
Data ingestion Cleaning Hybrid Vectors	Structural Metrics PageRank, BC, CC Approximations	Node2Vec Dimensionality Reduction Clustering

2. Recommended Tech Stack

- **Graph Library:** `igraph` (C++) or `networkx`
- **Embeddings:** `node2vec` or `gensim`
- **ML:** `scikit-learn`
- **Data:** `pandas`, `numpy`

3. Implementation Roadmap

Phase 1: Setup & Data Ingestion (Chiara)

☑ Repo Setup: GitHub repo with `.gitignore`

□ Parser: Extract ASIN, Group, and Edges from `amazon-meta.txt`

□ Filter: Keep only Book, DVD, Video, Music groups

□ Graph: Build Undirected Graph object

Phase 2: Feature Engineering (Parallel)

Structural (Luca)

□ PageRank: Compute and normalize

□ Clustering Coeff: Compute and normalize

□ Approx. Betweenness: Sampling-based BC ($k = 1000$)

□ Approx. Closeness: Sampling-based CC

Topological (Leonardo)

□ Node2Vec: Setup random walks

□ Training: Train for $d = 128$ dimensions

□ Storage: Save to `.npy`

Phase 3: Hybridization (Chiara)

□ Merge: Master DataFrame indexed by Node ID

□ Hybrid Vector: Concatenate Structural (6 dims) + Embedding (128 dims)

□ Reduce: Apply UMAP/PCA

Phase 4: Experiments (Leonardo)

- □ Cluster: K-Means ($k = 4$) on all 3 datasets
- □ Validate: Compute ARI and NMI scores
- □ Visualize: t-SNE scatter plots
- □ Profile: Measure execution time