Summary

Metric	Ground Truth	My Compute	Notes on Difference
Nodes	7,115	7,115	Perfect match → dataset parsed correctly.
Edges	103,689	103,689	Perfect match → no edges lost/skipped.
Largest WCC (nodes)	7,066 (0.993)	7,066	Exact match → connected components computed correctly.
Largest WCC (edges)	103,663 (1.000)	103,663	Exact match.
Largest SCC (nodes)	1,300 (0.183)	1,300	Exact match.
Largest SCC (edges)	39,456 (0.381)	39,456	Exact match.
Avg. clustering coefficient	0.1409	0.13865	Slightly lower (~1.6% diff). Likely due to floating-point precision or implementation (GraphFrames vs. SNAP definition).
Number of triangles	608,389	608,389	Perfect match.
Fraction of closed triangles	0.04564	0.04146	Noticeable difference (~9%). Maybe due to formula implementation: some libraries normalize differently (by triplets vs connected triples).
Diameter	7	9	Overestimate → Spark `shortestPaths` only samples landmarks. To get the exact diameter, need BFS from all nodes (a two-pass heuristic would get closer).
Effective diameter (90%)	3.8	4	Close match (within rounding). Using `approxQuantile` explains a slight float→int difference.