Approximating Statistics in Large Knowledge Graphs

KG Mini-Project

Alexander Hetzer & Tanja Tornede

January 31, 2018

University of Paderborn

Table of contents

Task

Task: Statistics in Large Knowledge Graphs

Number of Node Triangles

- Computation of number of node triangles can be a very time-consuming task for large graphs
  1. Implement and evaluate 3 different approaches in the Lemming[1] framework
  2. Implement a heuristic choosing the best approach based on the topology of a given graph
- Lemming Fork: `https://github.com/BlackHawkLex/Lemming`

[1] `https://github.com/dice-group/Lemming`

Classification from [2]

- Counting algorithm: output number of triangles
- Listing algorithm: output members of each triangle
  - Requires at least one operation per triangle
  - $\Rightarrow$ worst case lower bounds of $\Omega\left(n^3\right)$ (n nodes) or $\Omega\left(m^{3/2}\right)$ (m edges) [2]

Solution

Overview

Overview

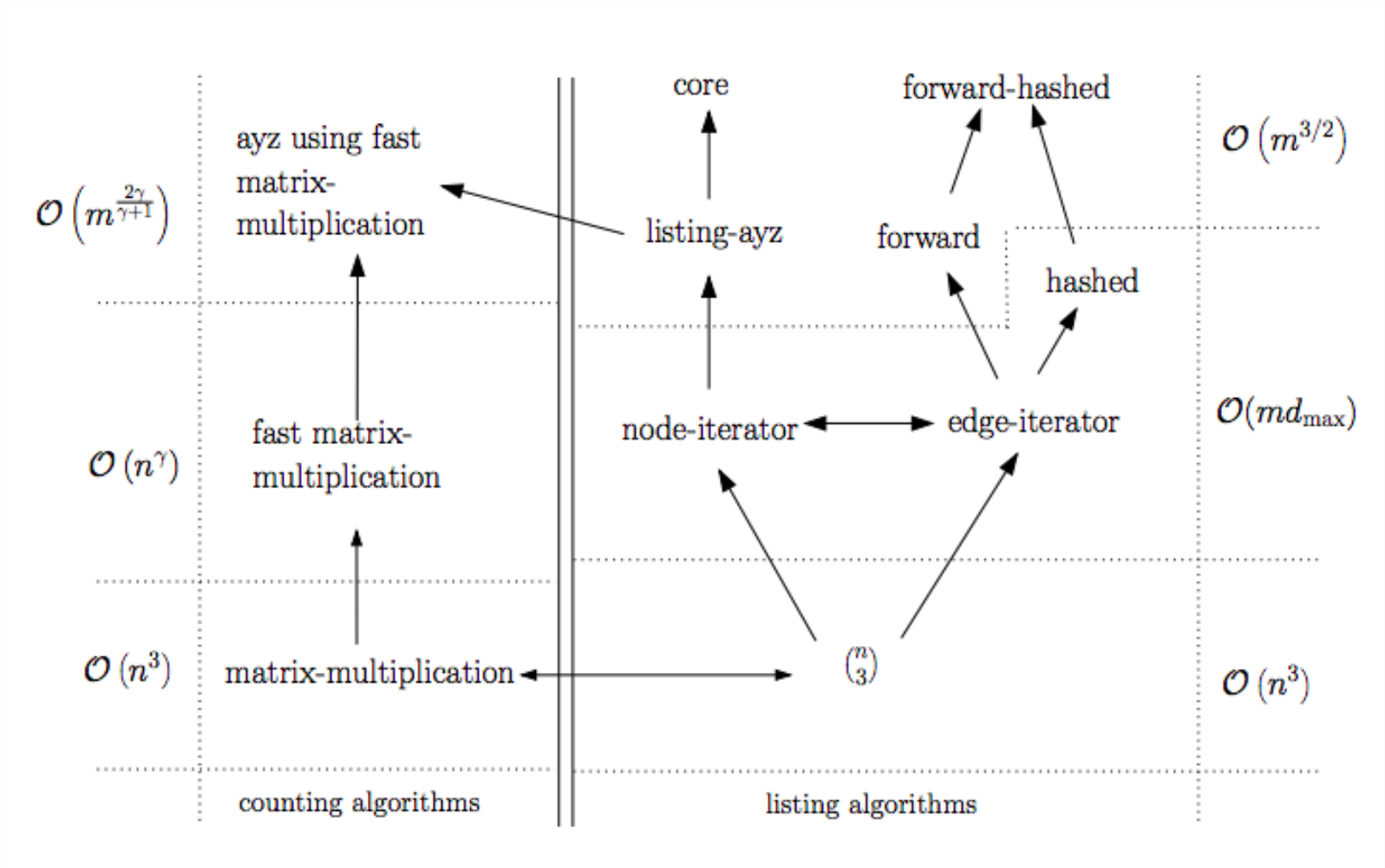

Figure 1: From [2] ($\gamma \leq 2.376 :=$ matrix multiplication exponent)

Algorithms of Choice

- forward[2]
  - Iterate over edges and check adjacency datastructure of both incident nodes
  - Makes use of a very clever adjacency datastructure
  - Reason: Good for skewed degree distributions [2]
- duolion[3]
  - Reduce graph by randomly removing edges with probability p
  - Count triangles in reduced graph
  - Multiply count by value depending on p
  - Reason: Scalable approach for large graphs
- matrix multiplication[2]
  - Sum of diagonal of cubic adjacency matrix A^3
  - Reason: Structure of graph of less importance

Algorithms of Choice

- node-iterator[2]
  - Iterate over nodes and test for each pair of neighbors if they are connected by an edge
  - Reason: Generalization of node-iterator-core
- node-iterator-core[2]
  - Similar to node-iterator, but make use of concept of *cores* (special subgraphs)
  - Reason: very efficient with respect to number of triangle operations [2]
- ayz[1]
  - Divide V into V^- (low degree vertices) and V^+ (high degree vertices)
  - Use node-iterator on V^- and matrix multiplication on V^+
  - Reason: interesting combination of node-iterator and matrix multiplication

Evaluation

Evaluation Setup

- Real world data:
  - Semantic Dog Food dataset[2]
    - 15 graphs of different sizes
  - Stanford Large Network Dataset Collection[3]
    - email-Eu-core network[4]
      ($V = 1005$, $E = 25571$, $\#triangles = 105461$)
    - EU email communication network[5]
      ($V = 265214$, $E = 420045$, $\#triangles = 267313$)
- Reference graph based on real world graphs
  (Star, Grid, Ring, Clique, Complete bipartite graph)

---

[2] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/de/dataset/semantic-web-dog-food

[3] https://snap.stanford.edu/data/

[4] https://snap.stanford.edu/data/email-Eu-core.html

[5] https://snap.stanford.edu/data/email-EuAll.html

Evaluation Setup (1)

- Runtimes reported in seconds
- Timeout of 60s except
  - matrix multiplication approach (15000s)
  - on complete bipartite graph (15000s)
- Timeouts reported as ?
- Tables missing algorithms $\Rightarrow$ timeouts on all graphs
- Legend:

Algorithm	Abbreviation
forward	f
node-iterator	ni
node-iterator-core	nic
matrix multiplication	mm
duolion with forward	df

Evaluation Results

- Fastest algorithm: forward
- Worst algorithm: matrix multiplication (due to inefficiency of IntMatrix.multiplication)
- Matrix multiplication approach scales incredibly bad (due to inefficiency of IntMatrix.multiplication)
- Complete bipartite graph is hardest reference graph (by far)

Evaluation

Table 1: email-Eu-core(V=1005, E=25571, T=105461)

A	Orig.	Star	Grid	Ring	Clique
f	0.134	0.002	0.010	0.002	0.000
ni	0.727	0.010	0.006	0.003	0.005
nic	0.713	0.225	0.008	0.003	0.005
df	0.092	0.005	0.009	0.010	0.001

Evaluation

Table 2: email-EuAll(V=265214, E=420045, T=267313)

A	Orig.	Star	Grid	Ring	Clique
f	3.619	0.504	1.172	0.894	0.793
ni	15.430	?	?	?	?
nic	?	?	?	?	?
df	2.693	1.707	3.060	1.921	0.878

Evaluation

Table 3: SemanticWebDogFood-Year:2001(V=1112, E=3994, T=204)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.163	0.017	0.022	0.010	0.003	0.075
ni	0.065	0.087	0.019	0.008	0.015	1.009
nic	0.088	0.254	0.011	0.006	0.007	0.892
mm	5.346	5.312	5.089	5.205	0.001	0.077
ayz	0.028	0.019	0.010	0.006	0.008	1.003
df	0.032	0.008	0.014	0.008	0.002	0.110

Evaluation

Table 4: SemanticWebDogFood-Year:2002(V=1833, E=6957, T=510)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.032	0.004	0.012	0.006	0.002	0.089
ni	0.034	0.031	0.012	0.006	0.011	3.922
nic	0.165	0.670	0.015	0.008	0.015	4.288
mm	32.736	37.110	33.276	32.043	0.000	0.382
ayz	0.042	0.034	0.014	0.008	0.012	4.072
df	0.040	0.011	0.021	0.013	0.003	0.296

Evaluation

Table 5: SemanticWebDogFood-Year:2003(V=2762, E=10948, T=920)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.058	0.007	0.016	0.008	0.002	0.167
ni	0.062	0.068	0.018	0.008	0.024	12.109
nic	0.326	1.503	0.023	0.013	0.024	13.057
mm	224.930	219.665	213.438	230.257	0.000	1.340
ayz	0.068	0.072	0.024	0.011	0.024	12.101
df	0.044	0.014	0.027	0.017	0.005	0.647

Evaluation

Table 6: SemanticWebDogFood-Year:2004(V=3890, E=15779, T=1404)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.074	0.018	0.022	0.011	0.004	0.390
ni	0.101	0.132	0.025	0.013	0.037	38.127
nic	0.595	2.981	0.034	0.020	0.038	40.433
mm	797.632	2497.347	647.531	792.651	0.000	3.620
ayz	0.116	0.141	0.032	0.016	0.039	38.744
df	0.062	0.017	0.036	0.024	0.006	1.579

Evaluation

Table 7: SemanticWebDogFood-Year:2005(V=5193, E=21646, T=1926)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.104	0.010	0.026	0.017	0.005	0.649
ni	0.156	0.231	0.034	0.016	0.058	83.643
nic	1.043	5.318	0.050	0.029	0.057	90.328
mm	1843.026	1777.603	1673.251	1784.795	0.001	8.974
ayz	0.177	0.249	0.047	0.026	0.061	84.809
df	0.085	0.025	0.051	0.040	0.008	2.922

Evaluation

Table 8: SemanticWebDogFood-Year:2006(V=7239, E=27130, T=2521)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.138	0.012	0.036	0.019	0.007	1.526
ni	0.204	0.442	0.047	0.022	0.091	259.192
nic	1.449	10.440	0.077	0.048	0.100	274.418
mm	5622.309	6294.610	5404.933	5499.206	0.001	30.881
ayz	0.222	0.451	0.092	0.029	0.093	260.294
df	0.115	0.047	0.067	0.042	0.012	7.115

Evaluation

Table 9: SemanticWebDogFood-Year:2007(V=12942, E=44378, T=4419)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.229	0.022	0.061	0.033	0.014	5.151
ni	0.388	1.391	0.085	0.039	0.233	1610.187
nic	2.996	33.526	0.174	0.126	0.245	1668.783
mm	?	?	?	?	?	406.775
ayz	0.414	1.431	0.110	0.056	0.244	1616.734
df	0.187	0.070	0.123	0.077	0.023	33.151

Evaluation

Table 10: SemanticWebDogFood-Year:2008(V=21731, E=83901, T=8078)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.638	0.047	0.115	0.311	0.026	13.182
ni	1.448	4.005	0.143	0.064	0.493	6297.660
nic	10.064	89.210	0.366	0.369	0.499	6707.209
ayz	1.167	4.041	0.178	0.083	0.492	6278.028
df	0.465	0.133	0.272	0.169	0.051	127.716

Evaluation

Table 11: SemanticWebDogFood-Year:2009(V=24766, E=97691, T=9687)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.535	0.044	0.121	0.064	0.031	21.363
ni	1.373	5.311	0.162	0.559	0.576	11411.384
nic	14.947	116.632	0.457	0.387	0.584	12011.483
ayz	1.457	5.324	0.209	0.097	0.605	11429.231
df	0.450	0.113	0.249	0.148	0.054	211.495

Evaluation

Table 12: SemanticWebDogFood-Year:2010(V=29561, E=119050, T=11536)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	1.127	0.051	0.138	0.076	0.036	28.374
ni	1.950	8.198	0.194	0.091	0.743	?
nic	22.531	167.911	0.625	0.476	0.767	?
ayz	2.207	8.001	0.241	0.117	0.741	?
df	0.524	0.135	0.288	0.172	0.061	322.517

Evaluation

Table 13: SemanticWebDogFood-Year:2011(V=34186, E=144348, T=13759)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	1.443	0.060	0.160	0.088	0.044	44.434
ni	2.666	10.264	0.222	0.102	0.943	?
nic	27.568	221.482	0.756	0.608	0.916	?
ayz	2.732	10.206	0.280	0.131	0.927	?
df	0.612	0.157	0.331	0.192	0.073	487.830

Evaluation

Table 14: SemanticWebDogFood-Year:2012(V=39896, E=179340, T=17500)

A	Orig.	Star	Grid	Ring	Clique	Com.Bi.
f	0.974	0.067	0.191	0.106	0.062	51.427
ni	3.674	14.007	0.292	0.115	1.404	?
nic	?	?	?	?	4.893	?
ayz	?	?	?	?	4.910	?
df	?	?	?	?	0.102	701.297

Evaluation

Table 15: SemanticWebDogFood-Year:2013(V=42736, E=193590, T=18972)

A	Orig.	Star	Grid	Ring	Clique
f	1.603	0.090	0.211	0.116	0.063
ni	4.321	15.811	0.458	0.124	1.522

Evaluation

Table 16: SemanticWebDogFood-Year:2014(V=44447, E=202273, T=19756)

A	Orig.	Star	Grid	Ring	Clique
f	1.249	0.077	0.207	0.112	0.063
ni	4.398	15.907	0.293	0.253	1.591

Evaluation

Table 17: SemanticWebDogFood-Year:2015(V=45387, E=207262, T=20166)

A	Orig.	Star	Grid	Ring	Clique
f	1.175	0.278	0.212	0.114	0.066
ni	4.546	16.828	0.329	0.162	1.998

Discussion

Discussion

- Grph[6] is well optimized
- BUT: Documentation is horrible + has some bugs (e.g. making a graph undirected)
- Efficient triangle counting without graph manipulation can be tricky
- Decision not to implement heuristic due to dominance of forward approach

[6] `http://www.i3s.unice.fr/~hogie/software/index.php`

Organization

Documentation & Code Quality

- README[7] with:
  - List of implemented approaches
  - Instructions on how to run an evaluation
- Classes, methods and variables with speaking names
- Algorithm classes feature Javadoc including link to paper
- Keep classes small & responsible for only one task

[7] `https://github.com/BlackHawkLex/Lemming`

Task Distribution

- Algorithms:
  - forward, duolion, matrix multiplication: Alexander
  - node-iterator, node-iterator-core, ayz: Tanja
- Evaluation: Both
- Unit tests: Tanja

Questions?

References I


N. Alon, R. Yuster, and U. Zwick.
Finding and counting given length cycles.
Algorithmica, 17(3):209–223, 1997.


T. Schank and D. Wagner.
Finding, counting and listing all triangles in large graphs, an experimental study.
In *WEA*, pages 606–609. Springer, 2005.


C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos.
Doulion: counting triangles in massive graphs with a coin.
In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 837–846. ACM, 2009.