On Computing the Diameter of Real World Graphs

Andrea Marino

ALGORITMI E PROGRAMMAZIONE PER L'ANALISI DEI DATI,
Università di Firenze

Firenze, a.a. 2020-21

Small world effect

The small average distance observed in the complex networks is referred as small world effect:

- If the size of the network is n, the average distance and the diameter have at most the order of magnitude of $\log(n)$.

The average distance in Facebook (721.1M nodes and 68.7G edges) is 5.7 and the diameter is 41.


The New York Times

Business Day

Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH

M·COIN Mobile payments

75 COUNT

Separating You and Me? 4.74 Degrees

By JOHN MARKOFF and SOMINI SENGUPTA

Published: November 21, 2011

- Average distance has been computed by applying HyperANF tool.
  - Boldi, Rosa, and Vigna. Hyperanf: approximating the neighbourhood function of very large graphs on a budget. In WWW 2011.
- Diameter has been computed by applying *i*FUB.
  - Crescenzi, Grossi, Habib, Lanzi, and Marino. On computing the diameter of real-world undirected graphs. Theoretical Computer Science, 2012.

Given a graph $G = (V, E)$ undirected connected.

Definition (Distance)

The distance $d(u, v)$ is the number (sum of the weights) of edges along shortest path from u to v.

Definition (Diameter)

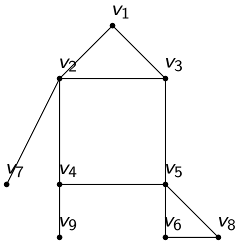
$$D = \max_{u,v \in V} d(u, v)$$

An Example Graph

Definition

- The eccentricity of a node u, $\mathrm{ecc}(u) = \max_{v \in V} d(u, v)$: in how many hops u can reach any node?

$$D = \max_{u \in V} \mathrm{ecc}(u)$$



	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	ecc
v_1	0	1	1	2	2	3	2	3	3	3
v_2	1	0	1	1	2	3	1	3	2	3
v_3	1	1	0	2	1	2	2	2	3	3
v_4	2	1	2	0	1	2	2	2	1	2
v_5	2	2	1	1	0	1	3	1	2	3
v_6	3	3	2	2	1	0	4	1	3	4
v_7	2	1	2	2	3	4	0	4	3	4
v_8	3	3	2	2	1	1	4	0	3	4
v_9	3	2	3	1	2	3	3	3	0	3

BFS [O(m) time]

For any i, $F_i(u)$ are the nodes at distance i from u (and vice versa).

Motivations

Communication: completion time of broadcast protocols based on flooding

Social: how quickly information reaches every individual

Web: how quickly, in terms of mouse clicks, any page can be reached

State of the Art

- Textbook Algorithm ($n = |V|$, $m = |E|$). Too expensive.
  - Perform n BFS and return maximum ecc.
    - A BFS from x returns all the distances from x and takes $O(m)$ time.
- Several other approaches (see [Zwick, 2001]) that solves all pairs shortest path. Still too expensive.
  - $O(n^{(3+\omega)/2} \log n)$ where ω is the exponent of the matrix multiplication.
- Empirically finding lower bound L and upper bound U
  - That is, $L \leq D \leq U$
  - D found, when $L = U$

State of the Art: Negative Results

- Unless the so-called Strong Exponential Time Hypothesis (SETH) is false, deciding whether a graph has diameter 2 or 3 requires $\Omega(n^2)$.
  - Informally, SETH says that SAT cannot be solved in sub-exponential time.
- By this reduction, unless SETH fails, $\Omega(n^2)$ time is required to get a $(3/2 - \epsilon)$-approximation algorithm for computing the diameter even in the case of sparse graphs.

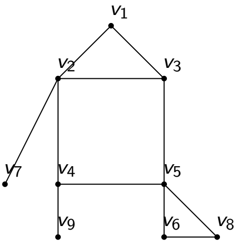
Liam Roditty, Virginia Vassilevska Williams: Fast approximation algorithms for the diameter and radius of sparse graphs. STOC 2013: 515-524

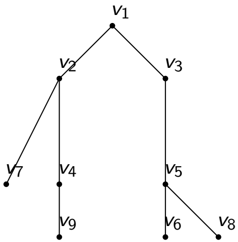
Part I

Computing lower and upper bound

Computing lower and upper bounds (undirected graph)

By using Single source (BFS) Shortest Path





Lower bound The eccentricity, ecc (height of the BFS tree) of a node.
In the example 3: at least a pair is at distance 3.

Upper bound The double of the eccentricity ecc of a node.
In the example 6: every node can reach another node going to v_1 by ≤ 3 edges and going to the destination in ≤ 3 edges.

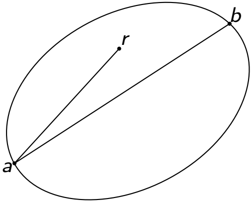
$x : d(x, u) = i$ and $y : d(u, y) = j \implies d(x, y) \leq i + j$
$i + j$ is the length of a path from x to y passing through u.

- Bounds by sampling but very often $L < D < U$ (see SNAP experiments)
  In the example diameter is 4: $d(v_7, v_8) = 4$.

Good lower bounds in undirected graphs: 2-Sweep

2-Sweep

1. Run a BFS from a random node r: let a be the farthest node.
2. Run a BFS from a: let b be the farthest node.
3. Return the length of the path from a to b.

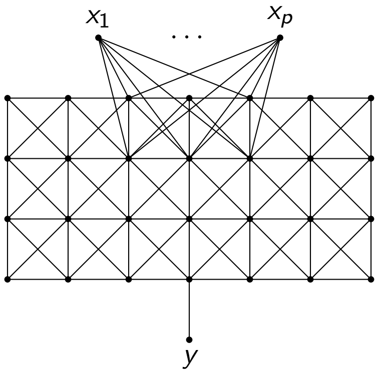


Return $d(a, b)$.

Experiments: effectiveness of 2-SWEEP

By starting from the highest degree node.

Category	# of Networks	2-dSWEEP *HdOut*	
		# of Networks in which *lb* is tight	Maximum error
PROTEIN-PROTEIN INTERACTION	14	11	1
COLLABORATION	14	12	1
UNDIRECTED SOCIAL	4	4	0
UNDIRECTED COMMUNICATION	36	34	2
AUTONOMOUS SYSTEM	2	1	1
ROAD	3	1	14
WORD ADJACENCY	7	4	1



In this modified grid with k rows and $1 + 3k/2$ columns. The algorithm can return k. The diameter of the network is instead $3k/2$.

Clemence Magnien, Matthieu Latapy, Michel Habib: Fast computation of empirically tight bounds for the diameter of massive graphs. ACM Journal of Experimental Algorithmics 13 (2008)

Pierluigi Crescenzi, Roberto Grossi, Claudio Imbrenda, Leonardo Lanzi, Andrea Marino: Finding the Diameter in Real-World Graphs - Experimentally Turning a Lower Bound into an Upper Bound. ESA (1) 2010: 302-313

Part II

Computing exactly the diameter

Bound Refinement: iterative fringe upper bound

Reminder

The *textbook* algorithm runs a BFS for any node and return the maximum ecc found.

Main Schema of a new algorithm

- Perform the BFSes one after the other specifying the order in which the BFSes have to be executed, while doing this:
  - refine a lower bound: that is the maximum ecc found until that moment.
  - upper bound the eccentricities of the remaining nodes.
  - stop when the remaining nodes cannot have eccentricity higher than our lower bound.

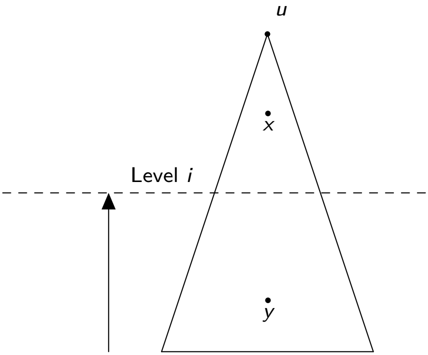
Good order can be inferred looking at some properties of BFS trees.

Notation

- Let u be any node in V and let us denote the set $\{v \mid d(u,v) = \mathrm{ecc}(u)\}$ of nodes at maximum distance $\mathrm{ecc}(u)$ from u as $F(u)$.
- Let $F_i(u)$ be the *fringe* set of nodes at distance i from u (note that $F(u) = F_{\mathrm{ecc}(u)}(u)$)
- Let $B_i(u) = \max_{z \in F_i(u)} \mathrm{ecc}(z)$ be the maximum eccentricity among these nodes.

Main observation

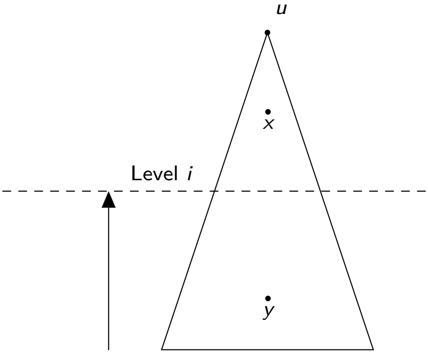
All the nodes x above the level i in $\mathrm{BFS}(u)$ having ecc greater than $2(i-1)$ have a corresponding node y, whose ecc is greater or equal to $\mathrm{ecc}(x)$, below or on the level i in $\mathrm{BFS}(u)$.



Main Theorem

Theorem

For any $1 \leq i < \mathrm{ecc}(u)$ and $1 \leq k < i$, and for any $x \in F_{i-k}(u)$ such that $\mathrm{ecc}(x) > 2(i-1)$, there exists $y \in F_j(u)$ such that $d(x, y) = \mathrm{ecc}(x)$ with $j \geq i$.



Proof: some observations

Observation

For any x and y in V such that $x \in F_i(u)$ or $y \in F_i(u)$, we have that $d(x, y) \leq B_i(u)$.

Indeed, $d(x, y) \leq \min\{\mathrm{ecc}(x), \mathrm{ecc}(y)\} \leq B_i(u)$.

Observation

For any $1 \leq i, j \leq \mathrm{ecc}(u)$ and for any $x \in F_i(u)$ and $y \in F_j(u)$, we have $d(x, y) \leq i + j \leq 2 \max\{i, j\}$.

Proof

Theorem

For any $1 \leq i < \mathrm{ecc}(u)$ *and* $1 \leq k < i$, *and for any* $x \in F_{i-k}(u)$ *such that* $\mathrm{ecc}(x) > 2(i-1)$, *there exists* $y_x \in F_j(u)$ *such that* $d(x, y_x) = \mathrm{ecc}(x)$ *with* $j \geq i$.

Proof.

Since $\mathrm{ecc}(x) > 2(i-1)$, then there exists y_x whose distance from x is equal to $\mathrm{ecc}(x)$ and, hence, greater than $2(i-1)$.
If y_x was in $F_j(u)$ with $j < i$, then from the previous observation it would follow that

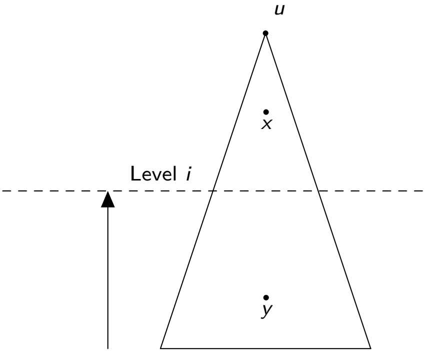
$$d(x, y_x) \leq 2 \max\{i-k, j\} \leq 2 \max\{i-k, i-1\} = 2(i-1),$$

which is a contradiction.
Hence, y_x must be in $F_j(u)$ with $j \geq i$. □

Implication

Corollary

Let lb be the maximum eccentricity among all the eccentricities of the nodes in or below the level i. The eccentricities of all the nodes above the level i is bounded by $\max\{lb, 2(i-1)\}$.



Implication

Corollary

Let lb be the maximum eccentricity among all the eccentricities of the nodes in or below the level i. The eccentricities of all the nodes above the level i is bounded by $\max\{lb, 2(i-1)\}$.

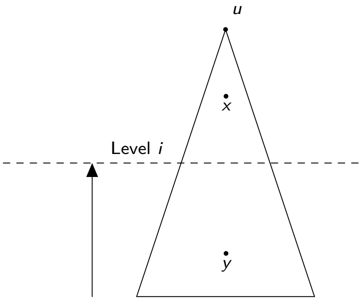
Proof.

For any node x above the level i, there are two cases:

- $ecc(x) \leq 2(i-1)$
- $ecc(x) > 2(i-1)$. The Theorem applies: there is a node y below the level i whose eccentricity $ecc(y)$ is greater than or equal to $ecc(x)$.

$$lb \geq ecc(y) \geq ecc(x)$$

$\square$



Back to the main schema

Perform the BFSes one after the others following the order induced by the BFS tree of a node u: starting from the nodes in $F(u)$, go in a bottom-up fashion.

- At each level i compute the eccentricities of all its nodes: if the maximum eccentricity found lb is greater than $2(i-1)$ then we can discard traversing the remaining levels, since the eccentricities of all their nodes cannot be greater than lb.
  - Since the eccentricity of the remaining nodes is bounded by $\max\{lb, 2(i-1)\}$

Given a node u.

- Set $i = \mathrm{ecc}(u)$ and $M = B_i(u)$.
- If $M > 2(i-1)$, then return M; else, set $i = i-1$ and $M = \max\{M, B_i(u)\}$, and repeat this step.

Algorithm 1: *i*FUB

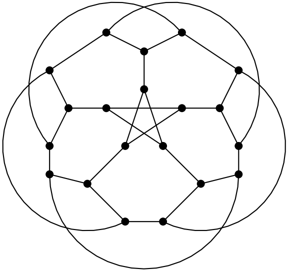
Input: A graph G, a node u
Output: The diameter D

```
i ← ecc(u);
lb ← ecc(u);
ub ← 2ecc(u);
while ub > lb do
    if max{lb, B_i(u)} > 2(i − 1) then
        return max{lb, B_i(u)};
    else
        lb ← max{lb, B_i(u)};
        ub ← 2(i − 1);
    end
    i ← i − 1;
end
return lb
```

Bad cases (1)

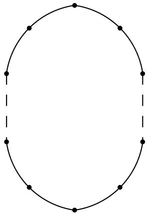
Two bad things can happen:

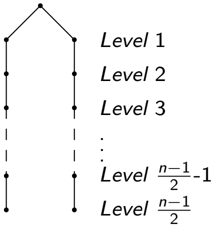
- The amount of nodes in $F(u)$ is linear: for instance if the BFS tree of the starting node is a binary tree.
  - In special regular graphs [such as Moore graphs] no good choice is possible.



Bad cases (2)

- Cases in which nodes have close eccentricity or the BFS trees are isomorphic, like the Moore graph, but even simpler: a cycle.
  - A cycle with n nodes (n odd) has diameter $\frac{n-1}{2}$, and each node has the same BFS tree.
  - The loop is repeated until $2(i-1) \geq \frac{n-1}{2}$, that is $i \geq \frac{n+3}{4}$, and stops the first time that $2(i-1) < \frac{n-1}{2}$.
  - The total number of iterations is equal to $\frac{n-1}{2} - \frac{n+3}{4} + 2 = \frac{n+3}{4}$.
  - The number of BFSes is $\frac{n+3}{2}$.





The choice of the starting node

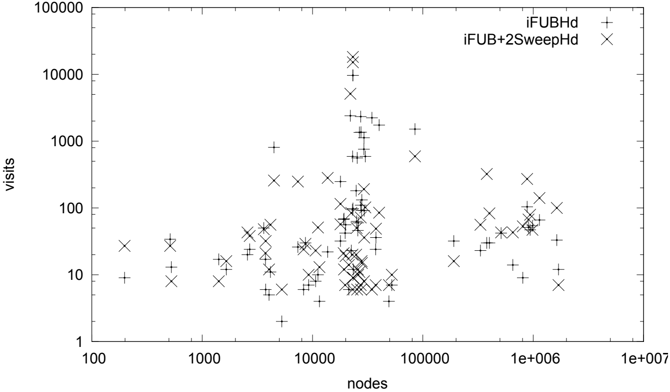
It heavily affects the performance (number of visits we will do).

Try to use a vertex with low eccentricity (i.e., central, well connected).

- High degree node (in-degree or out-degree).
- The node in the middle of the diametral path returned by 2-SWEEP.

Experiments for undirected graphs

The number of visits seems to be constant.



It has been used to compute the diameter of Facebook (721.1M nodes and 68.7G edges, Diameter 41) with just 17 BFSes.

More Experiments

Network name	n	m	Avg. Visits	Visits worst run
Wiki-Vote	1300	39456	17	17
p2p-Gnutella08	2068	9313	45.9	64
p2p-Gnutella09	2624	10776	202.1	230
p2p-Gnutella06	3226	13589	236.6	279
p2p-Gnutella05	3234	13453	60.4	94
p2p-Gnutella04	4317	18742	36.7	38
p2p-Gnutella25	5153	17695	85.1	161
p2p-Gnutella24	6352	22928	13	13
p2p-Gnutella30	8490	31706	255.4	516
p2p-Gnutella31	14149	50916	208.7	255
s.s.Slashdot081106	26996	337351	22.3	25
s.s.Slashdot090216	27222	342747	21.5	26
s.s.Slashdot090221	27382	346652	22.8	26
soc-Epinions1	32223	443506	6.1	7
Email-EuAll	34203	151930	6	6
soc-sign-epinions	41441	693737	6	6
web-NotreDame	53968	304685	7	7
Slashdot0811	70355	888662	40	40
Slashdot0902	71307	912381	32.9	40
WikiTalk	111881	1477893	13.6	19
web-Stanford	150532	1576314	6	6
web-BerkStan	334857	4523232	7	7
web-Google	434818	3419124	9.4	10

(`snap.stanford.edu` dataset)

More Experiments

Network name	n	m	Avg. Visits	Visits worst run
wordassociation-2011	4845	61567	412.5	423
enron	8271	147353	19	22
uk-2007-05@100000	53856	1683102	14	14
cnr-2000	112023	1646332	17	17
uk-2007-05@1000000	480913	22057738	6	6
in-2004	593687	7827263	14	14
amazon-2008	627646	4706251	136.3	598
eu-2005	752725	17933415	6	6
indochina-2004	3806327	98815195	8	8
uk-2002	12090163	232137936	6	6
arabic-2005	15177163	473619298	58	58
uk-2005	25711307	704151756	170	170
it-2004	29855421	938694394	87	87

(`webgraph.dsi.unimi.it` dataset)

The number of visits seems to be constant.
$\Rightarrow$ For any graph with more than 10000 nodes, D*i*FUB performs $0.001\%n$ visits in stead of n.

Why so well?

Suitable properties of the starting node u

(1) u has to be the node with minimum eccentricity, called radius R.

(2) Constant number of nodes in $F(u)$.

- If you are able to infer the node u such that (1) and $R = D/2$ you will stop after one iteration.
  - High degree node is very often a good choice.
  - If the lower bound path returned by 2-SWEEP is tight and $R = D/2$, the node in the middle of this path make us stop after one iteration.
- Almost always in real-world graphs $R = D/2$ (the minimum possible, maximum heterogeneity) and (2) is true if u is central.

Further development

- DiFUB can be generalized to weighted graphs, using Dijkstra Algorithm instead of BFS and sorting the nodes according to their distance from u. It works well, but not for Road Networks.
- Further optimization allow to do better than this and to compute also the diameter of weakly connected graphs (Borassi et al., TCS 2015).
- It is possible to prove that for the configuration model fixing the power law the number of BFSes is almost constant.

Open questions

- Why the *i*FUB method works so well in general.
  - Might be related to eccentricities distribution?

- Pierluigi Crescenzi, Roberto Grossi, Leonardo Lanzi, Andrea Marino: On Computing the Diameter of Real-World Directed (Weighted) Graphs. SEA 2012: 99-110
- Pilu Crescenzi, Roberto Grossi, Michel Habib, Leonardo Lanzi, Andrea Marino: On computing the diameter of real-world undirected graphs. Theor. Comput. Sci. 514: 84-95 (2013)
- Frank W. Takes, Walter A. Kosters: Determining the diameter of small world networks. CIKM 2011: 1191-1196
- Michele Borassi, Pierluigi Crescenzi, Michel Habib, Walter A. Kosters, Andrea Marino, Frank W. Takes: On the Solvability of the Six Degrees of Kevin Bacon Game - A Faster Graph Diameter and Radius Computation Method. FUN 2014: 52-63