

Ultra-Large-Scale Repository Analysis via Graph Compression

2020 IEEE 27th International
Conference on Software Analysis,
Evolution and Reengineering (SANER)

紹介者 森岡 勇介

背景

- 現在入手可能なソースコードの量は莫大なものになっている
- 開発成果物の分析するアプローチはビッグデータを複数のマシンに分割する方法に頼っている
- リソース消費の少ない分析手法に対する研究が少ない

目的・リサーチクエスチョン

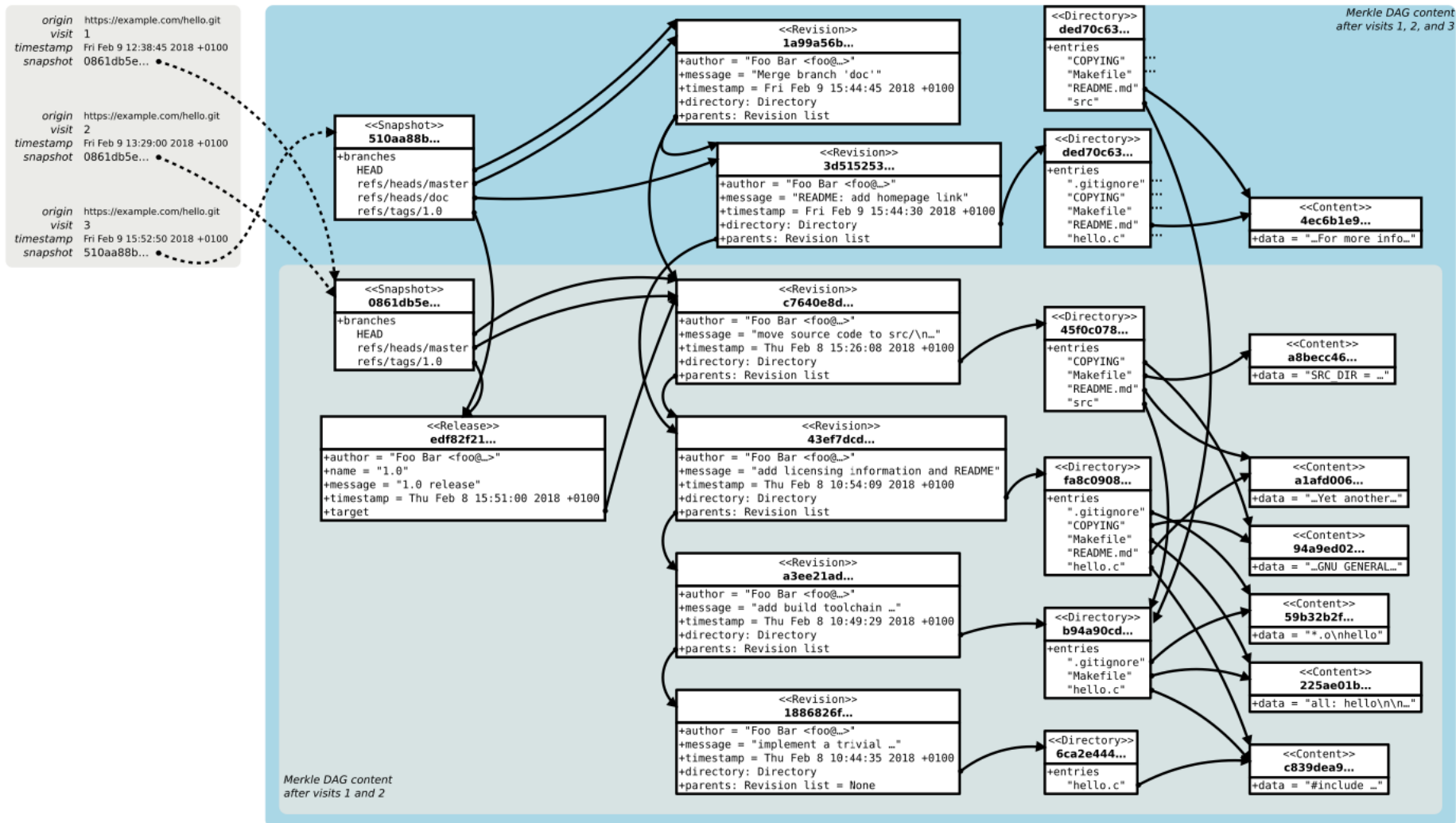
目的

大規模なソフトウェアコレクションの開発履歴の分析の
リソース消費の少ないアプローチの実現可能性を評価する

リサーチクエスチョン

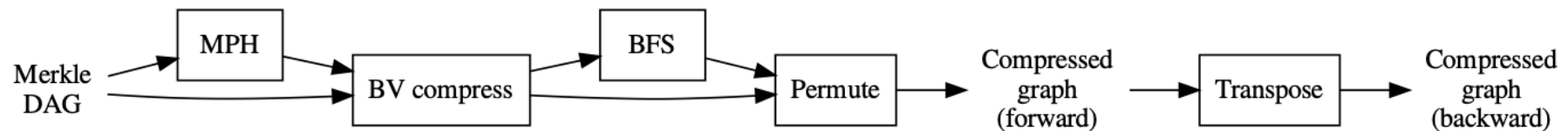
RQ:比較的安価な単一のマシンで超大規模ソフトウェア開発履歴の分析を
効率的に実行することは可能ですか

データモデル



アプローチ

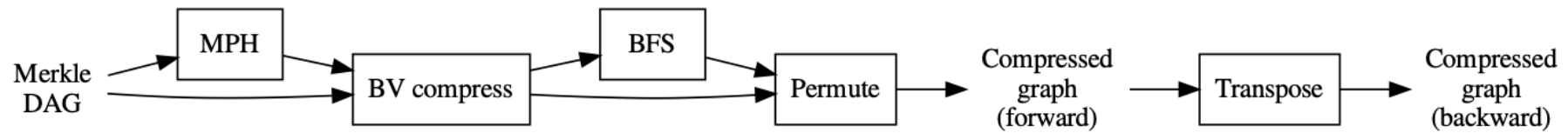
次の手順で**50億**以上のソフトウェアプロジェクトから収集された**50億**の固有のソースコードファイルと**10億**の固有のコミットで構成されたソフトウェアコレクションを圧縮する



Merkle DAG : 暗号化認証された有効非巡回グラフ

重複が削除される

アプローチ | MPH



MPH：すべてのノードに最小のハッシュ値をつける

すべてのノードに対して整数を一意に結びつけている

アプローチ | BVcompress

Node	Outdegree	Successors
...	...	...
15	11	13, 15, 16, 17, 18, 19, 23, 24, 203, 315, 1034
16	10	15, 16, 17, 22, 23, 24, 315, 316, 317, 3041
17	0	
18	5	13, 15, 16, 17, 50
...	...	...

CopyList: 基準としたノードのSuccessorsが自身のSuccessorsにもあれば1,なければ0を出力するリスト

圧縮前

Node	Outd.	Ref.	Copy list	Extra nodes
...	...	...	...	...
15	11	0		3,1,0,0,0,0,3,0,178,111,718
16	10	-1	01110011010	6, 293, 0, 2723
17	0			
18	5	-3	11110000000	32
...	...	...	...	...

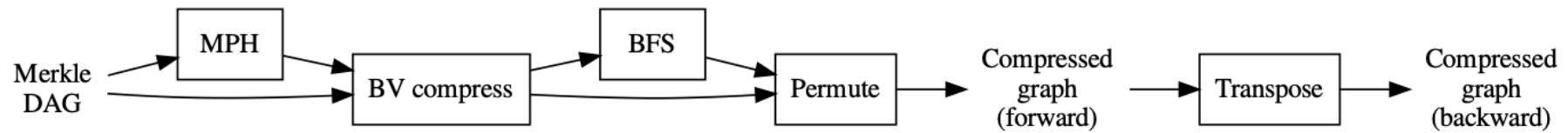
Extra Node:基準としたノードにないSuccessorsをgap compressしたリスト

gap compress:

$y_1 - x, y_2 - y_1 - 1, \dots, y_k - y_{(k-1)} - 1$
隣の要素からどれだけ離れているかを元に計算している

圧縮後

アプローチ | BFSとPermute

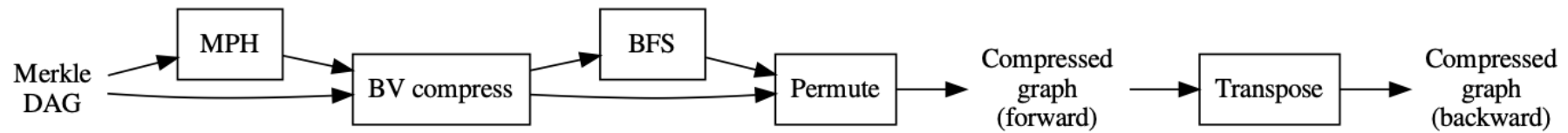


BFS: 局所性を最大化するノード順序を見つける

局所性を最大化することが優れた圧縮の鍵となる

Permute: BFSのノード順書に並び替える

アプローチ | Transpose



有向グラフなので逆向きが必要な場合を想定し、転置する

以上で圧縮が完了

結果 | 圧縮時間

TABLE IV
COMPRESSION TIME BREAKDOWN

Step	Wall time (hours)
MPH	2
BV Compress	84
BFS	19
Permute	18
Transpose	15
<i>Total</i>	138 (≈ 6 days)

全体の圧縮には6日かかった

利点:

- 1) 圧縮したものは再利用できる
- 2) 圧縮に必要なリソースは借りることができる

6日の圧縮時間が問題にならない場合は有効である

結果 | 探索時間

並列処理を行わずに幅優先探索をRAM内に保存して行った結果

TABLE VI
FULL GRAPH VISIT BENCHMARKS FOR A SINGLE-THREADED BFS VISIT

Forward (original) graph		Backward (transposed) graph	
wall time	1h48m	wall time	3h17m
throughput	1.81 M nodes/s (553 ns/node)	throughput	988 M nodes/s (1.01 μ s/node)

高速な探索が可能

結果 | サイズ

TABLE V
COMPRESSION RESULTS. COMPRESSION RATIOS ARE W.R.T. THE
INFORMATION-THEORETICAL LOWER BOUND FOR GRAPHS WITH THE
SAME DENSITY.

Forward (original) graph		Backward (transposed) graph	
total size	91 GiB	total size	83 GiB
bits per arc	4.91	bits per arc	4.49
compression ratio	15.8%	compression ratio	14.4%

圧縮率もよく91GiBに収めることができた
現在の市場では100GiBのRAMは300ドルほどで買えるので
RQに対して可能であると答えることができる

まとめ

- 研究内容によっては6日の圧縮の間に増える情報を無視できないので制限にあげられる
- 圧縮の間に増えた部分を動的に圧縮する方法が今後の展望にあげられる
(サイズの効率とトレードオフの関係にある)
- これまでにない大規模な分析が少ないリソースで可能になる