1 Introduction

Dealing with Maximum Flow Problem regarding a directed graph is interesting because of its potential to solve some real problem in our life such as studying communication networks. in this project three main algorithms including Ek, MPM and Dinic are considered to deal with the problem of finding the maximum flow between source and target vertex. The main goal is runtime performance investigation of mentioned algorithms. Achieved results will be investigated regarding outliers with the help of Box Plots, then Histograms and Density curves will be drawn to inspect distribution of results. In the final stage, R2 will be calculated for different regression models and best one will be considered as regression model for both Max Flow and CPU time separately. At the end, Q-Q and residual diagrams will be discussed. R project is considered our main tool for all calculation and diagrams.

2 Variable Identification

Dependent (outputs) and independent (inputs) variables are described in table 1 and table2.

Table 1: Description of inputs can be assigned for an arc

Description	Abbreviation
Number of vertices	N
Probability of generating an arc between two vertices	P
Maximum capacity	R

* To apply "gen.py" number of seeds and the name of text files are other arguments.
** Number of seeds are considered a constant in all scenarios (it is extracted randomly)

Table 2: Description of outputs

Description	Abbreviation
Maximum flow between source and target vertex	Max Flow
Performance runtime to calculate the maximum flow	CPU time

3 Scenarios:

Three main scenarios are defined for doing experiments. In the first, we consider five different R (Maximum capacity) and constant number for P (Probability of generating an arc between two vertices) and N (Number of vertices). In the second and third, P and N will be five different numbers respectively. In table 3, assigned numbers for each scenario are shown which will result fifteen *.txt files for deploying in three complied files.

Table 3: List of values assigned to fifteen scenarios and related text files' names.

Scenarios	N	P	R	File Name
1	1000	70	200	N1000P70R800S1972.txt
	1000	70	400	N1000P70R800S1972.txt
	1000	70	600	N1000P70R800S1972.txt
	1000	70	800	N1000P70R800S1972.txt
	1000	70	1000	N1000P70R800S1972.txt
2	1000	10	800	N1000P10R800S1972.txt
	1000	30	800	N1000P3/0000R800S1972.txt
	1000	50	800	N1000P50R800S1972.txt
	1000	70	800	N1000P70R800S1972.txt
	1000	90	800	N1000P90R800S1972.txt
3	200	70	800	N200P70R800S1972.txt
	400	70	800	N400P70R800S1972.txt
	600	70	800	N600P70R800S1972.txt
	800	70	800	N800P70R800S1972.txt
	1000	70	800	N1000P70R800S1972.txt

4 Implementation

All above mentioned files are applied C++ files, in another words, there would be 45 times application that generate 45 numbers as Max Flows and 45 numbers as CPU time. Results are putted into nine "*.csv" files to deploy in R program for results' investigation procedure. "*.csv" files are basically separated by algorithms and constants of the scenarios and includes Max Flow and CPU time. For instance, "EKN1000PdifR800.csv" includes results of "*.txt" files for first scenario in "EK.exe" file.

5 Outliers

To analyze outliers, we divided results based on algorithms (EK, MPM, Dinic: 15 values for each one); however, regarding Max Flow, values are the same so only one Box Plot is drawn. As it is shown there is not any outlier to be shown in the figures.

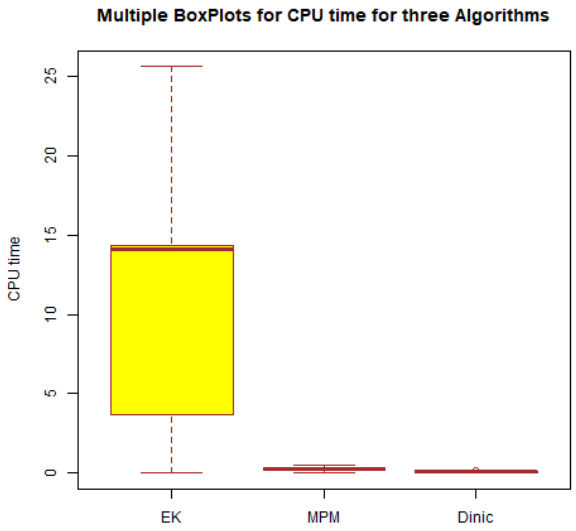


Figure 1: Box Plots for CPU times for three mentioned algorithms in three scenarios

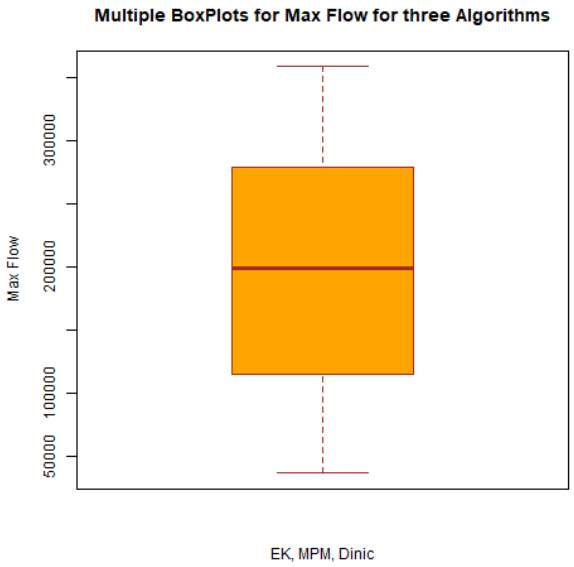


Figure 2: Box Plots for Maximum Flow for three mentioned algorithms in three scenarios (they are the same so one Box Plot is drawn)

6 Histograms and Density Curves

As previous, results are divided in three groups based on its calculation algorithms. Histograms show the distribution of a variable's set of values. It can be seen distribution of CPU time (Figure 3~5) and Max Flow (Figure 6) for bin size:5 based on EK, MPM and Dinic codes.

Regarding CPU times, results from all algorithms are bimodal, EK distribution is nearer to normal than others in Fig 3. Distribution of values are more or less as the same as EK but not as sharp as that. Results from Dinic have bimodal distribution with right skew. Comparison of time either domain or density curve, among figures 3,4 and 5 makes it clear that MPM is the fast one and EK is the lowest regarding CPU time which is representor of performance time.

We have only one histogram due to similarity of calculated Max Flow values for all three algorithms. Distribution of Max Flow values are symmetric around 200000, the minimum is a bit less than 50000 and the maximum is a bit more then 350000 (Fig 6).

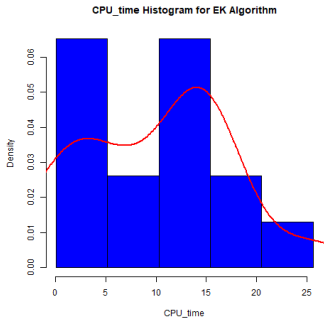


Figure 3: CPU time Histogram for EK

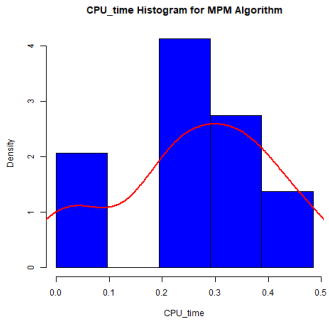


Figure 4: CPU time Histogram for MPM

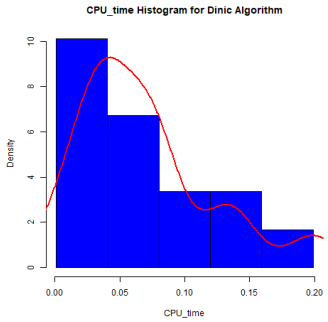


Figure 5: CPU time Histogram for Dinic

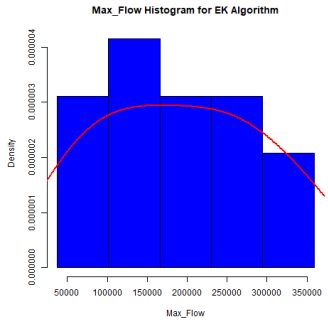


Figure 6: Max Flow Histogram for EK, MPM and Dinic

7 Regression model selection based on R2

It is necessary to evaluate regression models for both out puts e.g. CPU time and Max Flow to choose the best model for analyzing. Considering mentioned goals, several datasets are extracted from the results which their names mentioned in Table 4 & 5. For instance, dataset of "Dinic N1000 P70 Rdif" shows that results belong to experiments in which the algorithm is "Dinic" and number of vertices is 1000, probability is 70% and maximum capacity is different (200, 400, 600, 800, 1000 regarding table 3). Five different regression models are applied to calculate R2 (calculated 90 times) which are mentioned in columns' names of both table 3&4. In some cases, performance of regressions are similar to each other. For example, R2 from EK N1000 Pdif R807 and EK Ndif P70 R807 are more that .90 for last three regression models. To have better judgment regarding R2 values we calculate average for each model. Maximum average of different regression models for CPU time and Max Flow are shown in green cells in tables 4&5. It means that regrading regression of CPU time and Max Flow we have to apply (1/y)^2 = a + bx and y = a + bx respectively.

Table 4: Comparison of R2 for 5 regression models considering CPU time in 9 datasets

Dataset	$\frac{1}{y} = a + bx$	$y = a\log(x)$	$y = ax^b$	$y = a + bx$	$\sqrt{y} = a + bx$
Dinic N1000 P70 Rdif	0.41142240	0.42400340	0.39896820	0.39365110	0.39502830
Dinic N1000 Pdif R807	0.09399972	0.00085545	0.00878868	0.01268020	0.00027443
Dinic Ndif P70 R807	0.52407180	0.41901010	0.55473020	0.27893440	0.43215940
EK N1000 P70 Rdif	0.23437740	0.21934560	0.23500390	0.23552030	0.23527600
EK N1000 Pdif R807	0.58815830	0.72551690	0.90243520	0.92739410	0.99067300
EK Ndif P70 R807	0.56412800	0.73821190	0.92864040	0.90825740	0.98877080
MPM N1000 P70 Rdif	0.05708549	0.00001280	0.05356618	0.05106550	0.05217256
MPM N1000 Pdif R807	0.62433980	0.79956270	0.70502330	0.75552510	0.73674300
MPM Ndif P70 R807	0.50004580	0.80602250	0.68028210	0.88117530	0.89561830
R2 Average	0.39973652	0.45917126	0.49638202	0.49380038	0.52519064

Table 5: Comparison of R2 for five regression model considering Max Flow in 9 datasets

Dataset	$\frac{1}{y} = a + bx$	$y = a\log(x)$	$y = ax^b$	$y = a + bx$	$\sqrt{y} = a + bx$
Dinic N1000 P70 Rdif	0.81436570	0.94612880	0.94825710	0.99997740	0.98765410
Dinic N1000 Pdif R806	0.68979610	0.91571970	0.89497100	0.99976760	0.97230320
Dinic Ndif P70 R806	0.80347840	0.94309520	0.94624730	0.99967060	0.98775830
EK N1000 P70 Rdif	0.81436570	0.94612880	0.94825710	0.99997740	0.98765410
EK N1000 Pdif R806	0.68979610	0.91571970	0.89497100	0.99976760	0.97230320
EK Ndif P70 R806	0.80347840	0.94309520	0.94624730	0.99967060	0.98775830
MPM N1000 P70 Rdif	0.81436570	0.94612880	0.94825710	0.99997740	0.98765410
MPM N1000 Pdif R806	0.68979610	0.91571970	0.89497100	0.99976760	0.97230320
MPM Ndif P70 R806	0.80347840	0.94309520	0.94624730	0.99967060	0.98775830
R2 Average	0.76921340	0.93498123	0.92982513	0.99980520	0.98257187

8 Regression parameters

To merge regression line and scatter plot (from measured points), it is necessary to calculate regression parameters. Table 6 shows regression parameters for selected model regarding CPU time and table 7 shows them for selected model regarding Max Flow. Regarding parameters for regression of Max Flow, the type of algorithm has not any effects on the results of Max Flow values.

Table 6: Regression parameters for CPU time in 9 scenarios

CPU time: $\sqrt{y} = a + bx$	a	b
Dinic N1000 P70 Rdif	0.17636040	0.00017066
EK N1000 P70 Rdif	3.75898600	0.00014093
MPM N1000 P70 Rdif	0.57223600	0.00005790
Dinic N1000 Pdif R800	0.27421050	-0.00005526
EK N1000 Pdif R800	0.24858090	0.05224572
MPM N1000 Pdif R800	0.30471030	0.00364985
Dinic Ndif P70 R800	0.04689449	0.00021930
EK Ndif P70 R800	-0.89511330	0.00461471
MPM Ndif P70 R800	-0.12648550	0.00075809

Table 7: Regression parameters for CPU time in 9 scenarios

CPU time: $y = a + bx$	a	b
Dinic N1000 P70 Rdif	-261.6	350.631
EK N1000 P70 Rdif	-261.6	350.631
MPM N1000 P70 Rdif	-261.6	350.631
Dinic N1000 Pdif R800	-397	4000.44
EK N1000 Pdif R800	-397	4000.44
MPM N1000 Pdif R800	-397	4000.44
Dinic Ndif P70 R800	-6870.7	286.4265
EK Ndif P70 R800	-6870.7	286.4265
MPM Ndif P70 R800	-6870.7	286.4265

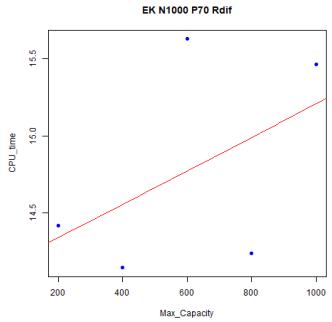
9 Regression Analysis

Considering graph (a) for figures 7-8-9 shows that in third scenario which number of vertices (N) are different, there is a strong correlation between N and CPU time; however, in scenario 2 a weak correlation between Flow capacity (R) and CPU time is shown. Regrading scenario 1, it is logical to accept that no correlation and relation is between probability (P) and CPU time. When it comes to consider MPM algorithm results of CPU time in three scenarios (figures 10-11-12 (a)) only different vertices would be accepted to have a correlation with CPU time and scenario 1 or 2 do not show any kind of correlation between CPU time and Flow Capacity or Probability. Results for regression CPU time in applying Dinic algorithms are the worst and making any conclusion regarding correlation between CPU time and all input variables (N, P, R) is hard (figures 13-14-15).

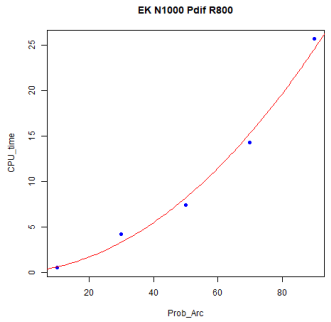
Figures 16 (a) shows regression line for all scenarios, considering the results of Max Flow are the same for all scenarios, so only one of them is drawn as representative of others. The same issue is defined for other algorithms regarding all scenarios. All of them shows strong correlation regarding linear regression and it is accepted that Max Capacity as input directly affect the Max Flow in all algorithms.

10 Q-Q Analysis

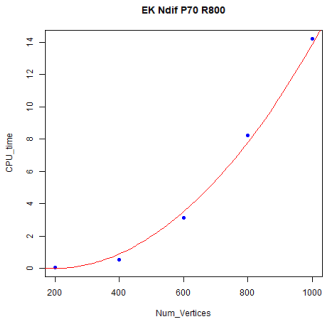
Figures 7~15 (c) shows Q-Q diagrams for all done experiences. Analyzing mentioned graphs can be considered as strong evidence to the analyzing results for Histograms and Density Curves in figures 3~6. In other words, results distributions do not follow any kind of normal distributions. As the same, there is not any evidence from figures 16-17-18 (c) that results for Max Flow follow any kind of normal distribution.



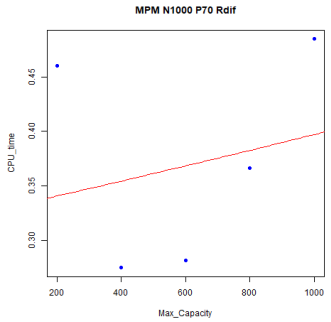
a



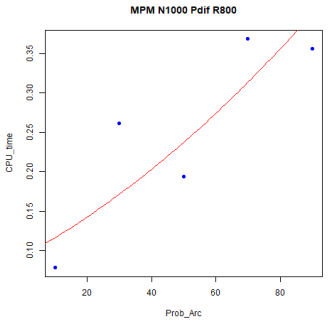
a



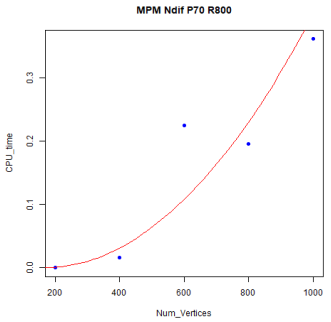
a



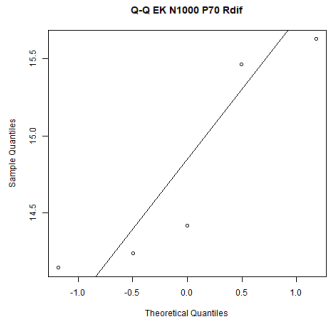
a



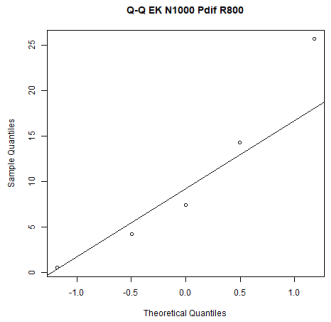
a



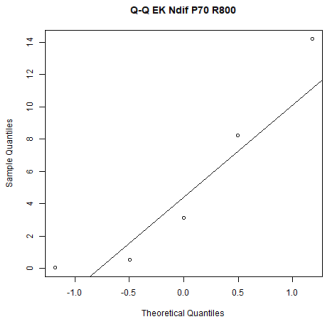
a



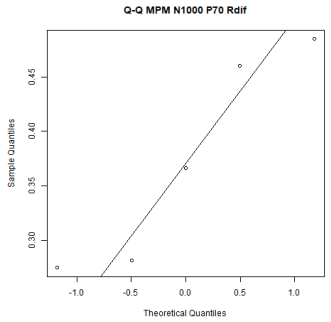
b



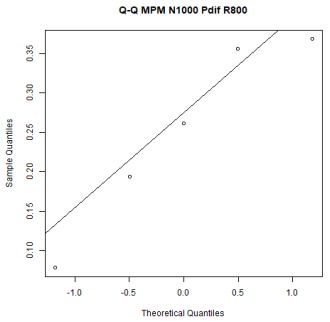
b



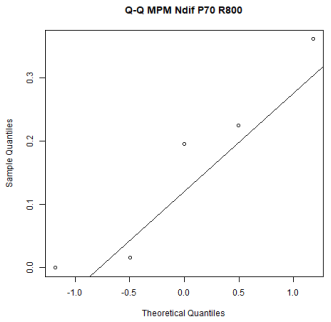
b



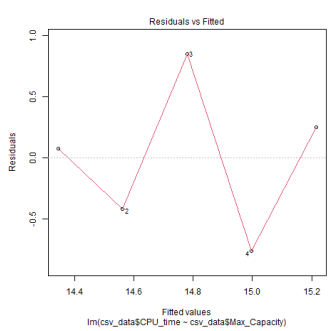
b



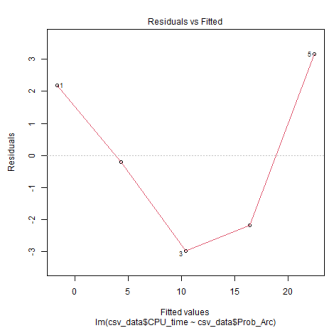
b



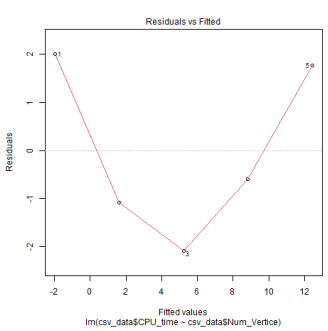
b



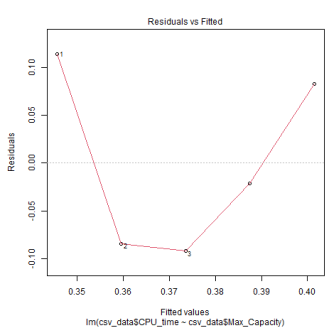
c



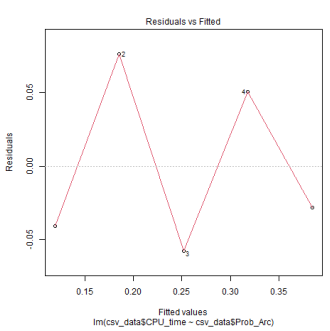
c



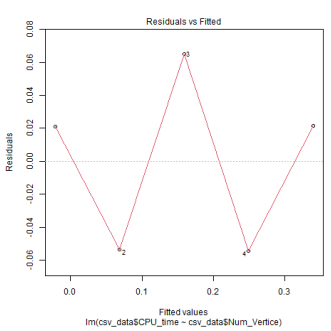
c



c



c



c

Figure 7: EK (Scenario 1, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted

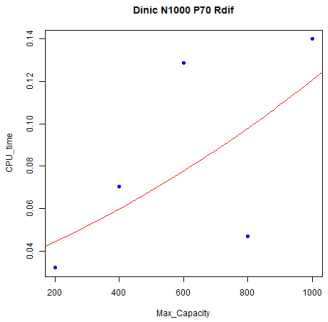
Figure 8: EK (Scenario 2, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted

Figure 9: EK (Scenario 3, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted

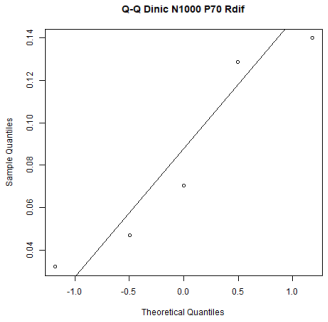
Figure 10: MPM (Scenario 1, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted

Figure 11: MPM (Scenario 2, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted

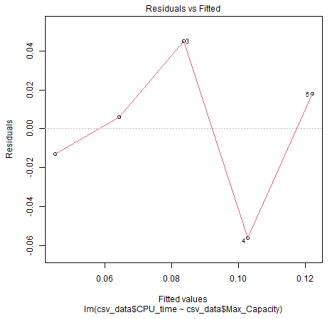
Figure 12: MPM (Scenario 3, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted



a

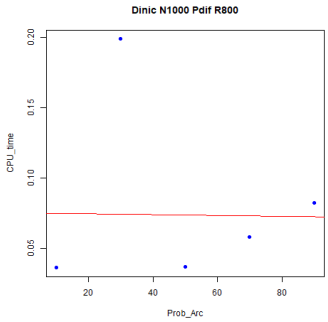


b

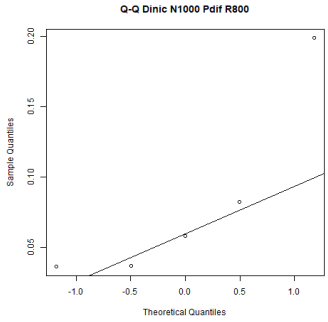


c

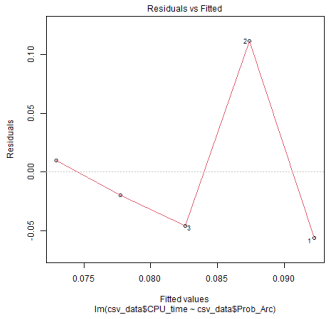
Figure 13: Dinic (Scenario 1, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted



a

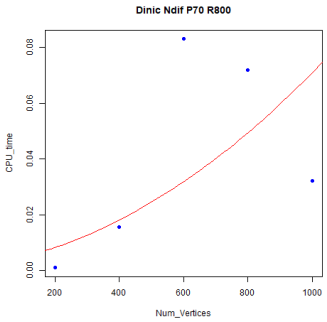


b

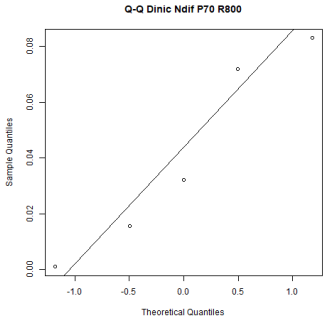


c

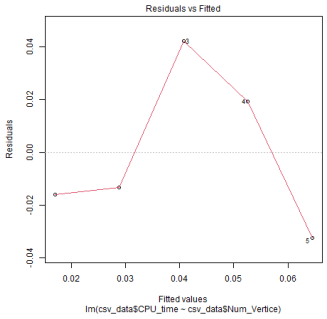
Figure 14: Dinic (Scenario 2, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted



a

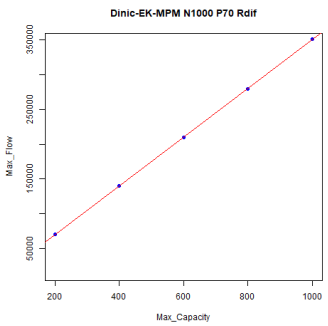


b

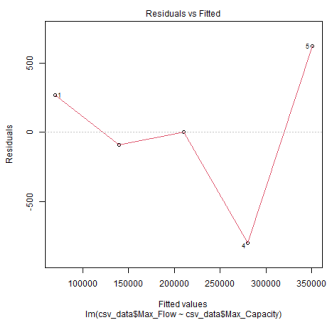


c

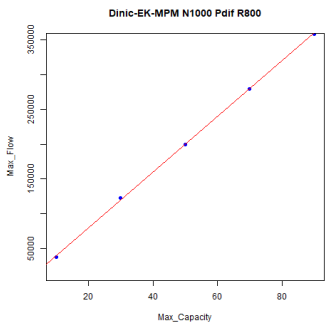
Figure 15: Dinic (Scenario 3, CPU time) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted



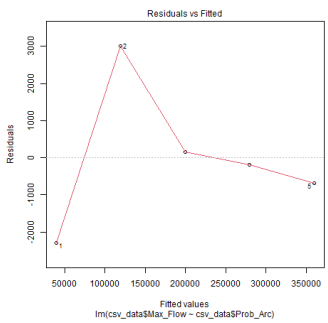
a



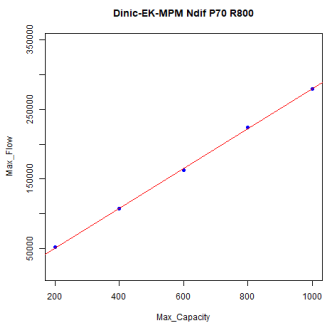
b



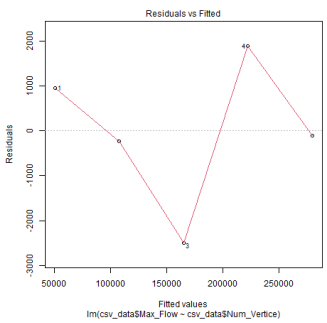
a



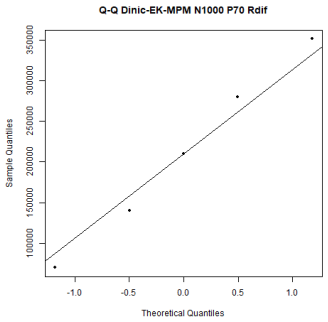
b



a

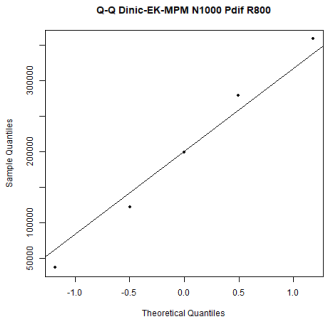


b



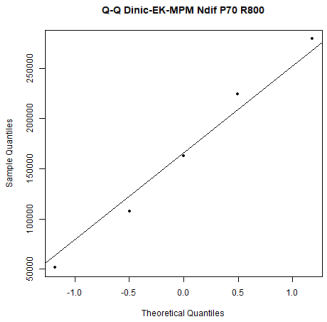
c

Figure 16: EK (Scenario 1-2-3, Max Flow) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted



c

Figure 17: MPM (Scenario 1-2-3, Max Flow) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted



c

Figure 18: Dinic (Scenario 1-2-3, Max Flow) a) Scatter & Regression, b) Q-Q Plot c) Residual vs Fitted

11 Residual vs Fitted Analysis

Even though, number of samples for all (b) diagrams from all figures are small but considering previous analysis can support the interpretation of results from Residual vs Fitted diagrams. Figures 7~15 (c) could not show any linearity as it is concluded from analyzing R2 for them. Moreover, no Homoskedasticity (or constant variance) could be seen in the all Residual vs Fitted graphs regarding CPU time and input variables (N, P, R) through all scenarios.

At first sight in analyzing Residual vs Fitted diagrams considering Max Flow for all inputs and scenarios from all algorithms, it is hard to draw that there is linearity, because the points are far away from supposed linear regression line. Although, if we consider the percentage of the values in y axis, they are very small portion which we can consider them to be so close to the middle line. As an example, 500 divided by 350000 is about 0.1%. Accordingly, it is possible to suppose that there is evidence regarding linearity.

12 Conclusion

Considering all above mentioned conclusion is summarized as the following:

- Regarding the performance time of algorithms; Dinic speed > MPM speed > EK speed
- Regarding strength of correlation, only Max Flow has strong correlation with input variables regardless of calculation algorithms.
- Regarding the type of regression, CPU time regression type is $1/y = a + bx$ regardless of calculation algorithms; however, Max Flow regression type is normal linear regression ($y = a + bx$)
- Regarding Max Flow, its values are neutral to the calculation algorithms.