Verification of User Behaviar Model in P2P Storage Distributed System Simulations

Goran Skondric*, Indira Hamulic* Eugen Mudnic**
*Faculty of information technologies, Mostar, Bosnia and Herzegovina/
** Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia
goran@fit.ba, indira@fit.ba, emudnic@fesb.hr

Abstract – In previous years many researchers, for their proof of concepts, used some kind of computer simulation. Simulation tools have predefined models for different aspects of their simulations. Correctness of these models has a big impact on simulation results. As time elapses, technologies and user behavior change, so models should follow these changes. This is especially important when we simulate p2p systems since they rely on user presence and their behavior. In previous research we noticed that in a group of users, administrative workers have similar pattern of behavior, and we developed model for this type of user. Our goal in this research is to try to verify this model using much larger dataset then we used in the previous research.

Keywords: modeling users, digital file collection, file size, file generation

I. INTRODUCTION

Simulators are main tool for creating systems blueprint and for proof of concepts for many researchers. Some simulators are designed and built heaving in mind specific use, like parallel systems, distributed systems, computer networks and so on. Sulistio et al. [1] analyzed 75 simulation tools and their characteristics and they proposed classification of simulation tools. They classified simulators based on modeling framework to entity-based and event-based. We agree with Ciprian et al. [2] since they concluded that distributed system simulation uses both modeling frameworks and that distributed system simulation should calibrate models using data collected from real environment [3]. Some simulators like Query Cycle, developed by Stanford University, allow us to configure model peer behavior [4] and others like Sim Grid allow import of real-world data using dumps (captured traffic from real environment). Many researchers [5,6] analyzed dynamic aspect of peers in order to recognize peer behavior and to properly configure models. In large complex distributed system, researchers are trying to simplify model, using generic peer models that often do not reflect real behavior of nodes.

In our previous research we proposed including user behavior that is different from churn behavior of a node. Churn reflects node behavior pattern of availability for distributed system, and user behavior reflects usage of node by user. We can have node that is constantly active, but user generates traffic just during the work hours.[7].

II. METHODOLOGY

In order to produce simulation that corresponds to real life environment, researchers monitor existing systems and collect data. Collecting data is a challenging task, researchers usually collect data from home networks or test bed [8] or gather from some network repositories [9]. These datasets that are available on Internet allow researcher to repeat experiments and verify someone's research but very often datasets are obsolete and don't reflect current characteristics of a certain type of traffic. Second option gives us much more realistic data, but there is a huge concern about privacy and security as noticed by Jia at al. [10]. After a few months we managed to get dataset from one enterprise where majority of users correspond to our user profile that we trying to verify. Dataset with censored data was used to conduct this research. Censored data are data from which administrator removed sensitive information that could be misused, data like folder and subfolder structure, names of users (real usernames are replaced with code UserX), some filenames are crypted and so on. This showed us that gathering data from real world is difficult to achieve, and censored data lack of some useful variables that could reveal some interesting facts about the observed phenomenon.

Collected data are from a microcredit organization in Bosnia and Herzegovina, from its headquarter and several branch offices. We believe that this is good for a dataset since in this way we can say that our dataset is not homogenous and affected by locality of place from where we collected data. Dataset contained data from 128 users / workstations, about file system and user data.

Our goal was to confirm that data of a certain type of users (administrative workers) follow negative binomial distributions in file generation, comparing file size and file structures. We were aware that personal digital collection (user files) can be affected by company culture, since some enterprise policy could forbid music, video or picture collections. During research while we were seeking for an enterprise that was willing to share data, we discovered that some companies do not allow users to store any personal files, or to visit social networks and so on. The goal of using statistical tools like RStudio, MS Excel is to explore dataset, to try to fit empirical data to probability distribution models, and to find one that best matches the collected data in order to compare it with our previous research.

III. RESEARCH

Total number of collected datasets we discarded is 53 since these datasets contained data from period that was shorter than two years. Reason could be related to the fact that user either got a new PC or was employed recently. And since we wanted results comparable with our previous research, we took the same period of time for our datasets, so we reduced all records to same time period.

A. File generation pattern

For the remaining datasets (75) we conducted statistical analysis and tried to fit empirical data to probability distributions suitable for this type of data. We tested five distributions: Poisson distribution, Negative binomial distribution, Zero inflated poison distribution, Zero inflated negative binomial distribution and Hurdle distribution. During statistical modeling phase, we included Hudrle and Zero inflated distributions since, based on the histogram, we concluded that datasets have high percentage of zeros and do not have typical histogram shape for the Poisson distribution.

Zero inflated negative binomial distribution treat data as they come from a combination of two distributions (zero part – Bernaulli distribution and second could be Poisson, or some other distribution). We will not elaborate theoretical aspects of the distribution since there are many papers related to this [11,12,13]. ZINB distribution relies on a predictor variable to model zero part.

Hurdle model treats data as they come from two separated processes, first those that generate zeros - binary outcome of whether a count variate has a zero or positive realization - Bernaulli distribution and if outcome is positive, and other that generate count data. [14]

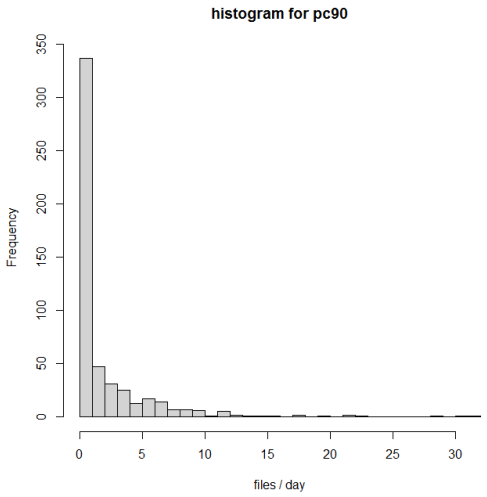

Figure 1. Histogram for PC90

Histogram clearly showed that these data do not fit to Poisson distribution but since many researchers use Poisson for discrete count data we also included this distribution. Using goodness of fit (GOF), we tested all datasets with the proposed model. We compared all models using Akaike information criterion - AIC parameters and summarized results.

TABLE I. AIC CLASSIFICATION FOR OBSERVED DATASETS

AIC	Total
NB < ZINB	49
NB > ZINB	18
NB > ZINB (0 – max 2)	8

According to Burnham and Andersons, if a difference between two models is less or equal to 2 then the difference between models is minimal and it is recommended to take a simpler model [15]. They also mention that AIC values from 5 to 10 constitute certain differences between models, and AIC value higher than 10 is a clear evidence that model with a lower AIC should be the preferred model. Since the majority of data sets follow negative binomial we chose to use negative binomial distribution for modeling datasets.

In our datasets we found a dataset of PC90 with ZINB and NB model differences of AIC values close to 2, so we used a rootagram to visually compare those two models (NB – figure 2 and ZINB - figure 3). Diagrams show minimum difference between the two models. For models that AIC designated better fit to ZINB distribution, we increased time of observation for the whole dataset. We discovered that some of datasets converged to NB after some additional time (more than our sample of 2 years).

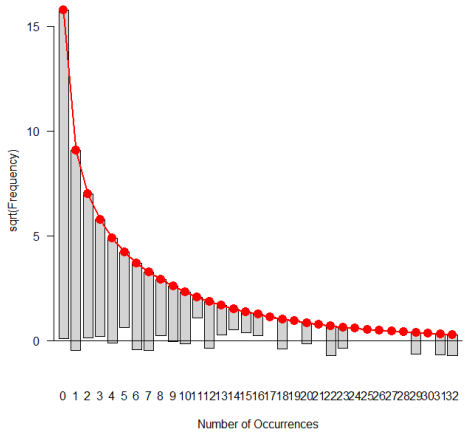

Figure 2. Rootagram for NB model of PC90

For model development we used statistical tool RStudio with R – programing language combined with libraries fitdistrplus (MASS package), zeroinfl (PSCL package) and library hurdl (package hurdler and VGAM). Developed model provided us with two parameters for pseudo generator (rnbinom). Parameters derived from NB model for PC90 dataset are *size* = 0.38916697 (0.03473193) and *mu* = 2.25049908 (0.17071406). Using this parameter, we generated our dataset. Frequency table of generated dataset was used to calculate expected frequencies. Using chisq.test we compared frequencies from the empirical dataset and expected frequencies from the generated dataset and based on the P parameter we concluded that there is no statistically significant difference between these datasets.

After we analyzed all datasets and generated models, we got 75 different parameter *size* and *mu* for our models. We wanted to find average parameters for the "mean model". As we mentioned earlier, our dataset was limited, and we could not conduct statistical merging of models but simple arithmetic mean of parameters. Merging all data points failed to produce model that follows negative binomial distribution.

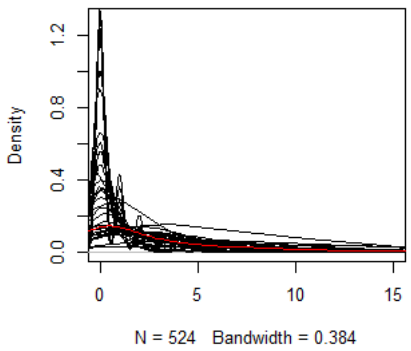

Figure 4. Average (mean) model

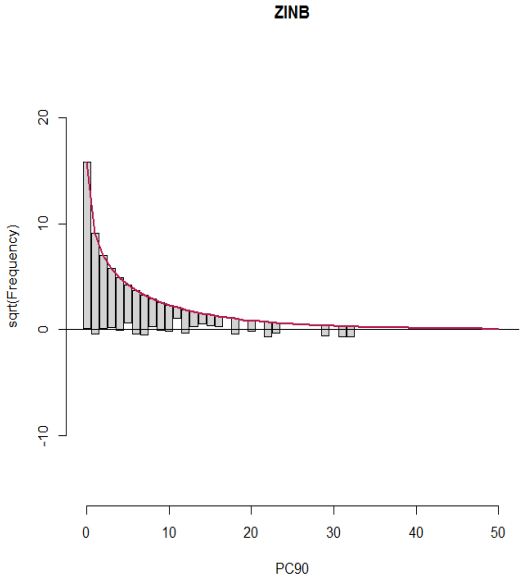

Figure 3. Rootogram for ZINB of PC90

Arithmetic mean of parameters from all models resulted with the following values *mu* = 5.397567 and *size* = 0.4061394. We generated the dataset using these parameters and produced density graph and histogram.

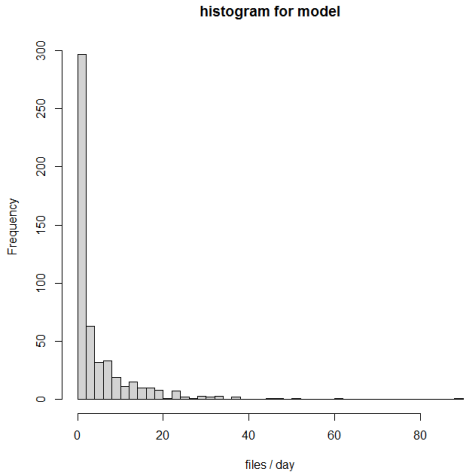

Figure 5. Histogram of data generated by average model

B. Structure of files

Second part of research is related to file structure of collected data. We analyzed file extension from all users that were used in modeling. Collected data are presented in Table 5. (using percentage of total files).

TABLE II. STRUCTURE OF USER FILES

Ext	Pdf	Xls	Doc	Zip	Jpg	Ppt	Txt	Mp3
%	45,20	27,74	12,2	0,66	12,27	0,08	0,02	1,8

Comparing to our previous research [7] it's clear that dominant file extension is different. Structure is definitely lead by company profile, since it s microcredit organization, employees use different reports (pdf) to make some calculations (xls) and to document that (doc), so we can say that this file structure is expected.

C. *File creation time*

We analyzed time stamps from collected datasets, we were interested in file generation time patterns.

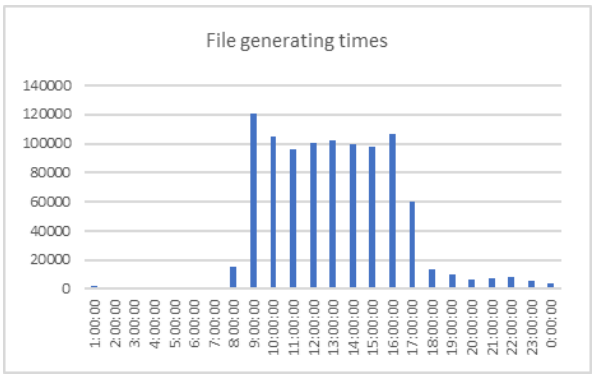

Figure 6 Times of file generation

Results corresponded to working hours and based on the histogram we could conclude that this data generation pattern followed normal distribution. Using statistical analysis KS test, we confirmed that this time generation followed normal distribution. This confirmed results from our previous research [7].

D. *file size*

Total number of files collected showed that 75% of files fall in a size range from 64 KB – 4 MB, what is slightly different then what we found out from our previous research [13]. Graph shows PDF of files

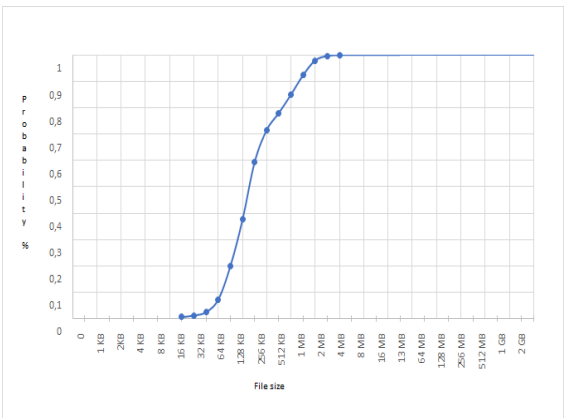

Figure 7. Probability distribution function for file size of collected datasets

size. Average file size from collected dataset is 240,59 KB and it is different than the forecasts found in similar research [16, 17] even comparing to our research. Our previous research included limited number of workstations, so that is not comparable with this dataset.

IV. CONCLUSION

Goal of our research was to verify our previous research result. In order to do that we had to find a representative dataset. We got dataset from real enterprise environment where majority of users belong to target group of users (administrative workers). Datasets were analyzed and from available datasets we chose 75 datasets for further processing.

First goal was to determine if dataset from this type of users really follows negative binomial distribution. Datasets were compared with several distributions (Poisson, Negative binomial, Zero inflated poisson, Zero inflated negative binomial and Hurdle distributions). Results of model comparison confirmed that this type of user can be modeled using Negative binomial distribution, with small percentage of users that follow Zero inflated negative binomial distribution. Lack of additional information in datasets disabled us from doing deeper analytics for this group of users, but we noticed that some of them converge to Negative binomial if we increase observation time. Final result is producing of the "default" model for modeling this type of users.

Second part of research was related to filetype structure, file creation time and file size. We noticed some difference in file structure since dominant file extension is pdf. while in our previous research that was .doc extension. Conclusion is that file structure is related to company portfolio of activities, and also to company culture since we noticed that only additional category of file was mp3 with 1,8 %, and in our previous research "other file extension" was around 8%.

File creation time confirmed results from the previous research but this was expected since company working hours is from 08:00 – 17:00 and majority of files were created during this period and our research confirmed that file creation follows normal distribution as we concluded in our previous research.

Last part of our research was related to file size. We noticed that there is increase in average file size, but not as expected from other researches. Thus, we can conclude that average file size increases with a time but with different rate than expected. This can be related to certain type of users and observer file types. Other researchers analyzed file system as a

whole and not just user files, so we can't confirm that their findings are incorrect.

REFERENCES

[1] A. Sulistio, C. S. Yeo, R. Buyya, "A taxonomy of computer-based simulations and itsmapping to parallel and distributed systems simulation tools", Software – Practiceand Experience, 34(7), June 2004, pp. 653–673.

[2] C. Dobre, F. Pop and V. Cristea, "New Trends in Large Scale Distributed Systems Simulation," *2009 International Conference on Parallel Processing Workshops*, Vienna, 2009, pp. 182-189.

[3] Fernandez, V. Gramoli, E. Jimenez, A. Kermarrec and M. Raynal, "Distributed Slicing in Dynamic Systems," *27th International Conference on Distributed Computing Systems (ICDCS '07)*, Toronto, ON, 2007, pp. 66-66,

[4] Schlosser, M., Condie, T., & Kamvar, S. (2003). „Simulating A File-Sharing P2P Network.",Proceedings of the First Workshop on Semantics in P2P and Grid Computing (December 2002).

[5] Daniel Stutzbach, Reza Rajaie, "Understanding churn in peer-to-peer networks", IMC '06: Proceedings of the 6th ACM SIGCOMM conference on Internet measurementOctober 2006 Pages 189–202.

[6] O. Herrera and T. Znati, "Modeling Churn in P2P Networks," *40th Annual Simulation Symposium (ANSS'07)*, Norfolk, VA, 2007, pp. 33-40

[7] G. Skondric, I. Hamulic and E. Mudnic, "Optimization of availability and resource utilisation in LAN based P2P storage distributed systems," 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH), East Sarajevo, Bosnia and Herzegovina, 2020, pp. 1-6

[8] Rolf Stadler, Rafael Pasquini, Viktoria Fodor, "Learning from network device statistics", Journal of Network and Systems Management, September 2017.

[9] Shen-Tat Goh, Panos Kalnis, Spiridon Bakiras, Lee T Tan, "Real Datasets fot File-shareing Peer-to-Peer Systems", Database Systems for Advanced Applications, 10th International Conference,Beijing, China, April 17 – 20, April 2005

[10] Q. Jia, L. Guo, Z. Jin and Y. Fang, "Privacy-Preserving Data Classification and Similarity Evaluation for Distributed Systems," 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), Nara, 2016, pp. 690-699

[11] 1. Loeys T, Moerkerke B, Smet OD, Buysse A (2012) The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. Br J Math Stat Psychol 65: 163- 180.

[12] Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, et al. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8: 1235-1246.

[13] Zeileis A, Kleiber C, Jackman S (2008) Regression models for count data in R. J Stat Softw 27: 1-25

[14] Berry C. Arnold, N.Balakrishnan, Jose Maria Sarabia, Roberto Minguez, "Advances in Mathematical and Statistics Modeling", Birkhäuser; 2008th Edition (September 2, 2008)

[15] Burnham, Kenneth P., David Raymond Anderson, and Kenneth P. Burnham. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

[16] N. Agrawalet al., "A five-year study of file-system metadata,"Trans. Storage, vol. 3, no. 3, Oct. 2007.

[17] K. M. Evans and G. H. Kuenning, "A study of irregularities in file-size distributions," inIn InternationalSymposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS '02, 2002.