Paper

# Application of Social Network Inferred Data to Churn Modeling in Telecoms

Witold Gruszczyński[1] and Piotr Arabas[2,3]

[1] *Codeflip, Warsaw, Poland*
[2] *Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland*
[3] *Research and Academic Computer Network (NASK), Warsaw, Poland*

**Abstract—The subject of this work is the use of social network analysis to increase the effectiveness of methods used to predict churn of telephony network subscribers. The social network is created on the basis of operational data (CDR records). The result of the analysis is customer segmentation and additional predictor variables. Proposed hybrid predictor employs set of regression models tuned to specific customer segments. The verification was performed on data obtained from one of the Polish operators.**

***Keywords*—churn reduction, classification, social networks.**

## 1. Introduction

One of the main challenges faced by telecommunications companies is to stop the migration of customers. This process, called *churn*, is in a great extent caused by the operators themselves, who offer attractive conditions for new users to attain dominance in highly competitive market. The issue is so important that, as shown by statistics, client acquisition may be several times more expensive [1], [2] than churn prevention, and nearly 80% of users surveyed admitted they have changed provider at least once [2].

A typical method to reduce churn are advertising campaigns encouraging the purchase of new products, or participation in the loyalty program. To achieve positive effect such an offer should be addressed to carefully selected group of customers to reduce campaign cost, and to avoid bothering loyal clients with too numerous contacts.

The telcos possess abundant data related to the customers and their use of services. The usage information is available mainly in the form of Call Detail Records (CDR) and is believed to be valuable input for users behavior modeling. The classic approach is to use the methods defined broadly as data mining [3] to predict subscribers decisions based on changes in the use of their services. This allows to select the group of customers prone to churn, which then can be addressed adequately prepared offer.

The subject of the work described in this paper is to enrich the standard techniques with data describing social links between subscribers. This information may be obtained by building social network graph basing on CDR data.

The remainder of the article is organized as follows: Section 2 presents a description of the problem and the proposed solution, taking into account previous work on similar issues. In Section 3 the concept of predictive model and its augmentation with social network analysis is outlined. Section 4 briefly presents implemented programs and software packages used to carry out the research. Next Section discusses results of the model validation and finally Section 6 concludes on the research.

## 2. Problem Description

### 2.1. Purpose of Work

The aim of the presented work was to create set of models, which could predict if a particular customer is prone to churn. The only input for these models were operating data available in the form of CDRs. In the absence of any other customer data (e.g. information on the place of residence, age or type of work) it was especially important to take full advantage of all its aspects. Authors' previous work [4] demonstrated that applying regression modeling to information derived from CDRs have given promising results. The idea presented here is to introduce segmentation basing on the structure of the social network modeling customers relations to construct a specific and thus more efficient predictive models in each of segments. Additional data obtained during the analysis of the social network and not used for segmentation is included as an input of models. Information inferred through analysis of social graph should augment usage statistics with details describing social linkage between users and may be considered as an attempt to reproduce (or replace) unavailable user related data. Authors believe that so derived variables will increase resolving power of models as they should be unrelated to the base set of variables.

### 2.2. Similar Solutions and Related Work

A typical approach to the problem of predicting customers behavior is to use different kinds of classifiers, including regression [5], decision trees [6] or genetic algorithms [6]. These techniques have long been known and used in many fields, however, as indicated by a number of authors, the use of one of them does not guarantee to sufficiently exploit the information hidden in the telecoms operation data [6]–[8]. Application of social networks analysis [9], [10] is proposed by many authors [11], [12], but usually it is employed to analyze the relationship between subscribers only [13]–[15].

In the method proposed in this paper both kinds of data are used: common statistics of traffic data, as well as those resulting from social network analysis. This approach is similar to those presented in [16]. A major difference, however, is that customers are segmented prior to modeling making it possible to identify a number of specific models and thus to increase their accuracy.

### 2.3. CDR Data

The data were provided by one of the Polish wired telephony operators and contained usage records within three months collected out of a selected part of the operator network. The set consisted of 130 million records generated by 315,000 users. Due to the confidentiality restrictions data have been anonymized by deleting some fields and encoding subscriber and recipient numbers. As a result only date, time, call duration, tariff ID and customer ID were available. The latter information allowed to associate the subscriber with all phone numbers owned.

Although the anonymization is deterministic (i.e. the IDs are coded each time the same) there is some loss of information. Especially it is impossible to distinguish some well known numbers[1], local and long distance calls or interpret tariffs. For similar reasons it was not possible to obtain any additional customer data. The most important problem was lack of information describing user status – a user was considered churner when no CDRs were generated for longer than three weeks. The available information is much more constrained than in the typical analysis carried out by the network operator, who has full access to the data. However, it resembles case of prepaid services, where customer information is generally not available.

## 3. Construction of the Model

The main idea was to derive new information through the analysis of the social network constructed on the basis of CDR data. This information can provide additional input for classifiers, it can also be used for customer segmentation.

### 3.1. Social Network Reconstruction

The main problem was the incompleteness of the data, resulting in a significant portion of connections leading outside the network. It may be attributed to the limited customer base, but also to the fact that they were collected only from a part of network. As a result the social network built (internal network) covered less than 281,000 of the approx. 299,000 individual users. The corporate users, i.e. those having more than one phone number were deliberately omitted. It is supposed that the churn mechanism may be different in their case, so other, perhaps individual churn reduction methods should be used.

To increase the number of analyzed users a bipartite network was constructed – a connection between two users was added if they call the same number outside the provider network (for extended discussion of social network reconstruction see e.g. [17]). So constructed network consisted of 13711 subscribers, the remaining 4178 users could not be included in any network.

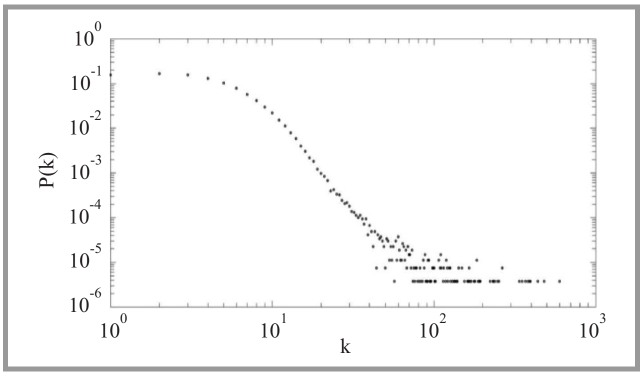


*Fig. 1.* Inner network node degree distribution.

Analysis of degree distribution in the internal network (see Fig. 1) suggests power-law with characteristic exponent $\alpha = 2.75$ – a value slightly above reported for wired telephone networks [18]. Scale-free nature of the network, allows to expect typical social networks phenomena, especially the nodes with an extremely high degree among majority of low degree nodes. The existence of as many as seven separate connected components (see Table 1) is easily explained by the limited range of the network.

Table 1
Connected components of the inner network

| Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Users | 80007 | 67853 | 38410 | 26289 | 25240 | 22525 | 20419 |
| % | 28 | 24 | 14 | 9 | 9 | 8 | 7 |

It is probable that in a bigger set of data much more calls can be terminated internally reducing the number of connected components, a more likely outcome would be, however, a greater coverage of the internal network. Connected components can be used for a natural segmentation of users. In the contrast to typical segmentation methods this one does not use any additional user information to explicitly include them into predetermined classes (e.g. division by age, gender, place of residence, etc.). Instead of this unavailable data it uses social contacts characteristic.

[1] An example may be customer information office typically considered as one of numbers routinely consulted before deciding to leave the network.

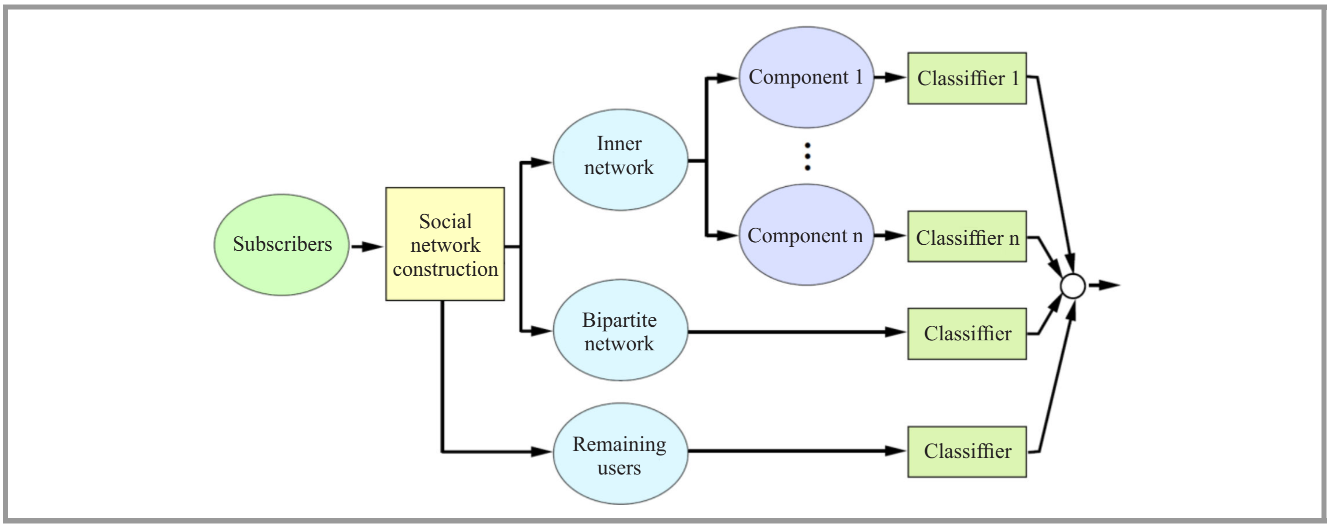


*Fig. 2.* Hybrid model.

### *3.2. Regressive Model*

Regression provides a reference point for the hybrid model described in the Section 4. The hybrid model is in fact a combination of several specific regression classifiers working for separate segments of customers, hence the principles of their construction and use are the same. In both cases logistic regression was used to create classifier dividing all customers into two classes, with these being prone to churn marked as "positive". A convenient feature of using regression is possibility to evaluate the selection of input variables by examining their significance levels. It should be remembered that as a non-linear classifier logistic regression allows a slightly better distinction of classes than the linear methods, however it still remains quite sensitive to the selection of predictor variables.

The construction of the regression model consists of two phases. First the predictor variables must be selected, then model parameters are identified and classifier quality is assessed. Data used for model identification covered three weeks. To detect changes in the customer behavior three time windows with a width of the week are defined. Predictor variables calculated in these windows, were the statistics corresponding to different types of user activity – e.g. number of failed connections, number of connections of the specified type, cost of calls (i.e. pulses counted). Most of the variables were differences between successive weeks, which allowed to eliminate dependence between subsequent variables. It is crucial for the effectiveness of a classifier to select variables of adequate predictive power – apart from avoiding correlation, it could be done through experiments and evaluation of their significance level. Insignificant variables were removed, and replaced with another variables (if available).

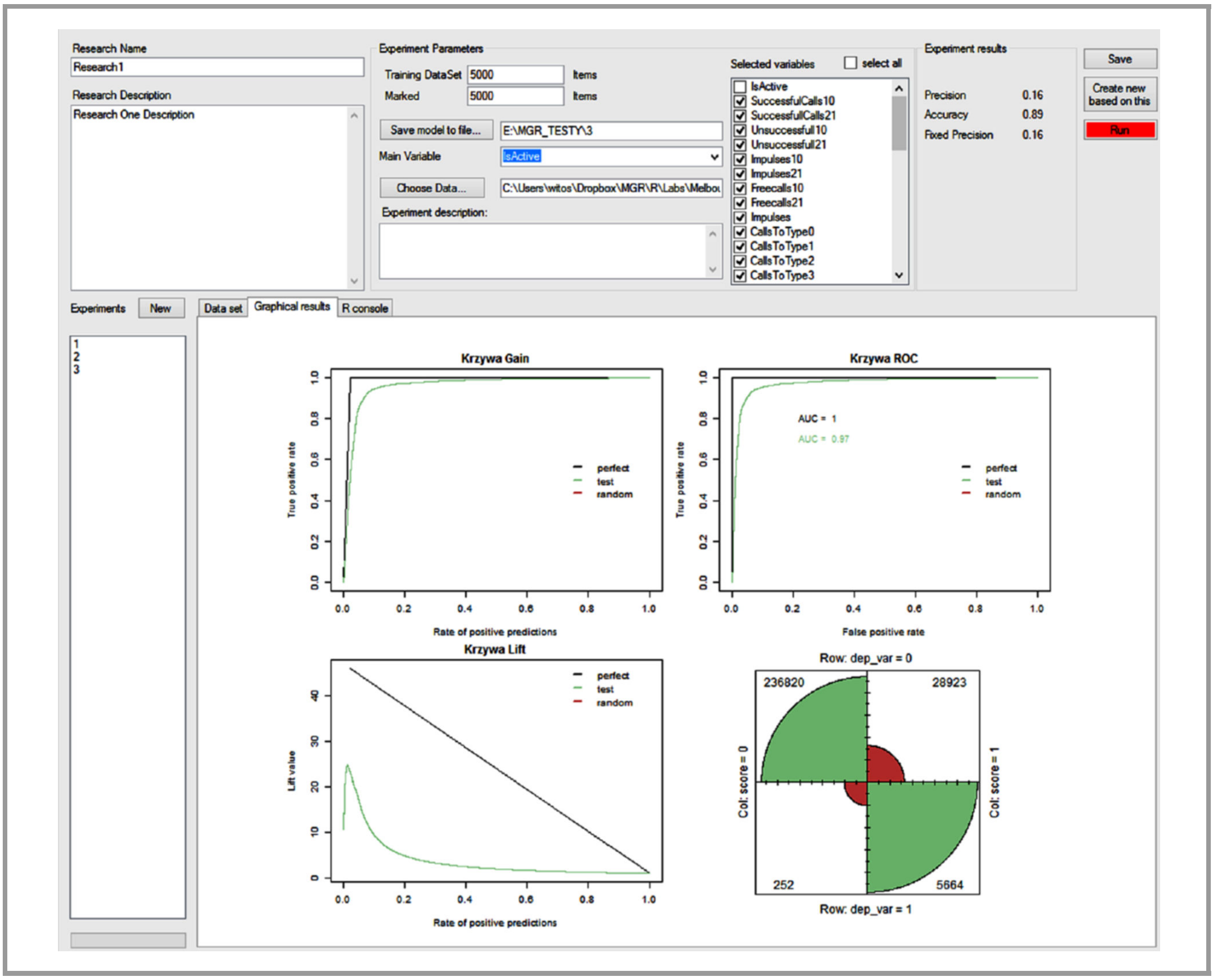
Another problem is the proper choice of samples constituting the training set. It should be noted that the number of users leaving the network was relatively small and amounted to slightly more than 3% in the analyzed period. So uneven distribution of the classes in the training data usually leads to classifiers that work correctly only for the more numerous class. For this reason, the training set was created by sampling of both classes independently, so that the data contained 20–25% of churners.

### *3.3. Hybrid Modeling*

Hybrid model uses seven connected components of the internal network to segment most (namely 280,743) users. In addition social ties of 13711 users may be modeled by bipartite network. In this way, it is possible to build eight customized regression classifiers, which hopefully should better reflect specificity of each segment, and be more effective than basic classifier described in Subsection 3.2. The improvement may be attained not only by preselecting customers and so making input data more homogeneous, but also by using new predictor variables derived by the social network analysis. The variables considered were node degree, closeness or page-rank, as well as the number of node neighbors, who left the network recently. The last variable is a natural, and widely used churn indicator, which calculation has little algorithmic complexity. Relatively small number (4178) of subscribers cannot be connected into any social network and so remains outside 8 segments. In their case it is only possible to build a classifier using basic predictor variables like the one described in Subsection 3.2. The complete diagram of the hybrid model is presented in Fig. 2.

## 4. Software

To facilitate the prepossessing of data, and the identification and verification of classifiers custom programs were prepared. Visualization of graphs was performed with help of LaNet-vi package [19]. MS SQL Server was used to store initial CDR data and results of classification. Pre-



Research Name: Research1

Research Description: Research One Description

Experiment Parameters

Training DataSet 5000 Items

Marked 5000 Items

Save model to file... E:\MGR_TESTY\3

Main Variable: IsActive

Choose Data... C:\Users\witos\Dropbox\MGR\R\Labs\Melbou

Experiment description:

Selected variables □ select all

□ IsActive
☑ SuccessfulCalls10
☑ SuccessfulCalls21
☑ Unsuccessfull10
☑ Unsuccessfull21
☑ Impulses10
☑ Impulses21
☑ Freecalls10
☑ Freecalls21
☑ Impulses
☑ CallsToType0
☑ CallsToType1
☑ CallsToType2
☑ CallsToType3

Experiment results

| | |
|---|---|
| Precision | 0.16 |
| Accuracy | 0.89 |
| Fixed Precision | 0.16 |

*Fig. 3.* Model identification and validation GUI.

processing, done mostly using capabilities of the MS SQL Server included CDR compression, separation of corporate users and removal of anomalous entries. Specifically, nodes generating thousands of calls which are supposed to be kind of automatic network testers. Compression of records may be achieved mainly by removing fields that were cleared during the data anonymization. Removal of automatically generated calls was necessary as they could distort user activity statistics. To speed up creation and analysis of social networks appropriate indexes were added. The most important part of the developed software was designed to carry out the identification and validation of models. Its main functionalities are importing data from the database and assisting user in the selection of predictor variables. Identification of classifier (logistic regression) parameters was implemented with help of R package [20] and was performed on a training data set prepared through sampling. After identifying the model it is possible to verify it on a separate data set. Program implements simple to use graphical interface – e.g. Fig. 3 presents verification of classifiers tab.

# 5. Model Identification

## *5.1. Verification Methods*

In order to verify the results of identification a number of classification experiments on a separate datasets was carried out. The results can be illustrated in the form of a confusion matrix, being a table containing the numbers of correctly and incorrectly classified samples in both classes – see Table 2. The true negative (TN) indicates the number of samples correctly classified as having no investigated characteristic – in this case the users who chose to remain subscribers of the telecoms. False positive (FP) refers to samples incorrectly classified as leaving the network (type

Table 2
Confusion matrix

| Classification result | | N | P |
|---|---|---|---|
| Original classes | N | TN | FP |
| | P | FN | TP |

1 error). Samples labeled false negative (FN) are users incorrectly classified as non churners (type 2 error), and finally, true positive (TP) are correctly detected users that decided to leave the network.

Confusion matrix contains a full description of results, however, it is not convenient when several models need to be compared. To make it easier F1 measure, defined by the following formula, was used:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FN}+\mathrm{TP}}, \tag{1}$$

where symbols TP, FN i FP are defined according to the Table 2. It may be visible, that F1 does not consider true negatives allowing to minimize the overhelming influence of dominant class [21].

### 5.2. Regressive Model

As mentioned in the Subsection 3.2 basic regressive model is a reference point for the hybrid modeling. The following predictor variables were used:

- difference of the number of successful connections,
- difference of the number of failed connections,
- difference of the number of calls of specific types,
- difference of the number of calls calls on weekends,
- difference of the number of pulses,
- difference of the number of free calls.

Differences are calculated in the three consecutive weekly periods. The call was considered unsuccessful if it lasted less than 2 s. Connection types were read from the appropriate column in CDRs, however unfortunately any interpretation could be associated to them. Using logistic regression allowed to identify relatively efficient classifier – see the verification results in Table 3 and Fig. 4 which is a graphical representation of the confusion matrix (cf. Table 2) – parts of pie chart correspond to the cells of confusion matrix, note that FP radius is scaled by TN and FN is scaled by TP.

Table 3
Identification of basic regression classifier

| | |
|---|---|
| Training set cardinality | 12000 |
| Number of churners in training set | 3000 |
| Validation set cardinality | 285017 |
| F1 value | 0.558 |

The results of the verification show that the model performs better for the more numerous class, while the most visible shortcoming is the high number of errors of the first type – in fact, there are more false positives then properly detected churners. This means that taking preventive measures – e.g. advertising campaign would cost twice as much as in the case of perfect information, however would be far more effective than untargeted action. The number of undetected churners – 2846 (which amounts to 28%) seems to be less alarming. However, it also shows that better modeling is necessary.

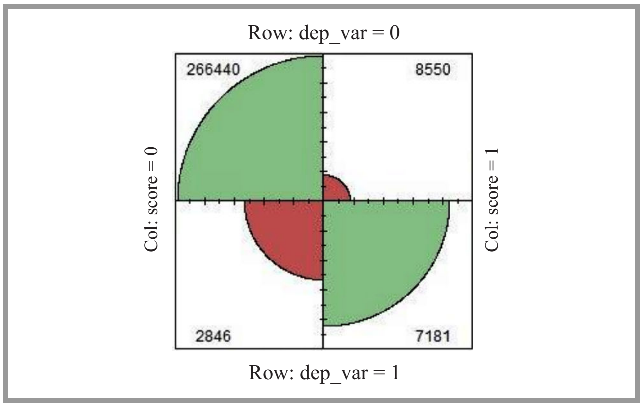


*Fig. 4.* Graphical representation of basic regression classifier confusion matrix.
(See color pictures online at www.nit.eu/publications/journal-jtit)

### 5.3. Hybrid Model

Hybrid model consisted of nine logistic regression classifiers operating on separate segments of customers built of internal network components (7 segments, see Table 1), bipartite network and set of remaining users. Predictor variables of basic regression model (see Subsections 3.2 and 5.2) have been supplemented with the data resulting from the social network analysis:

- node degree,
- closeness,
- page-rank,
- number of node neighbors who left the network.

### 5.4. Verification of Hybrid Model Components

The results of the identification and verification of the classifier for the components 1–7 are presented in Table 4 and Fig. 5a-g respectively. Comparing the value of the F1 measure and analyzing confusion matrices for the first component it can be noticed that the classifier is much more effective than simple regression (see Subsection 5.2). This results mainly from limiting the scope to the group of users sharing common features. Augmenting the set of predictor variables is another source of improvement – it allows to introduce new, important information derived through analysis of social ties between customers.

The results of verification for the component 2 are shown in Table 4 and Fig. 5b. Unfortunately, they are not as good as for the component 1, and are even slightly worse than the results obtained using the basic regression classi-

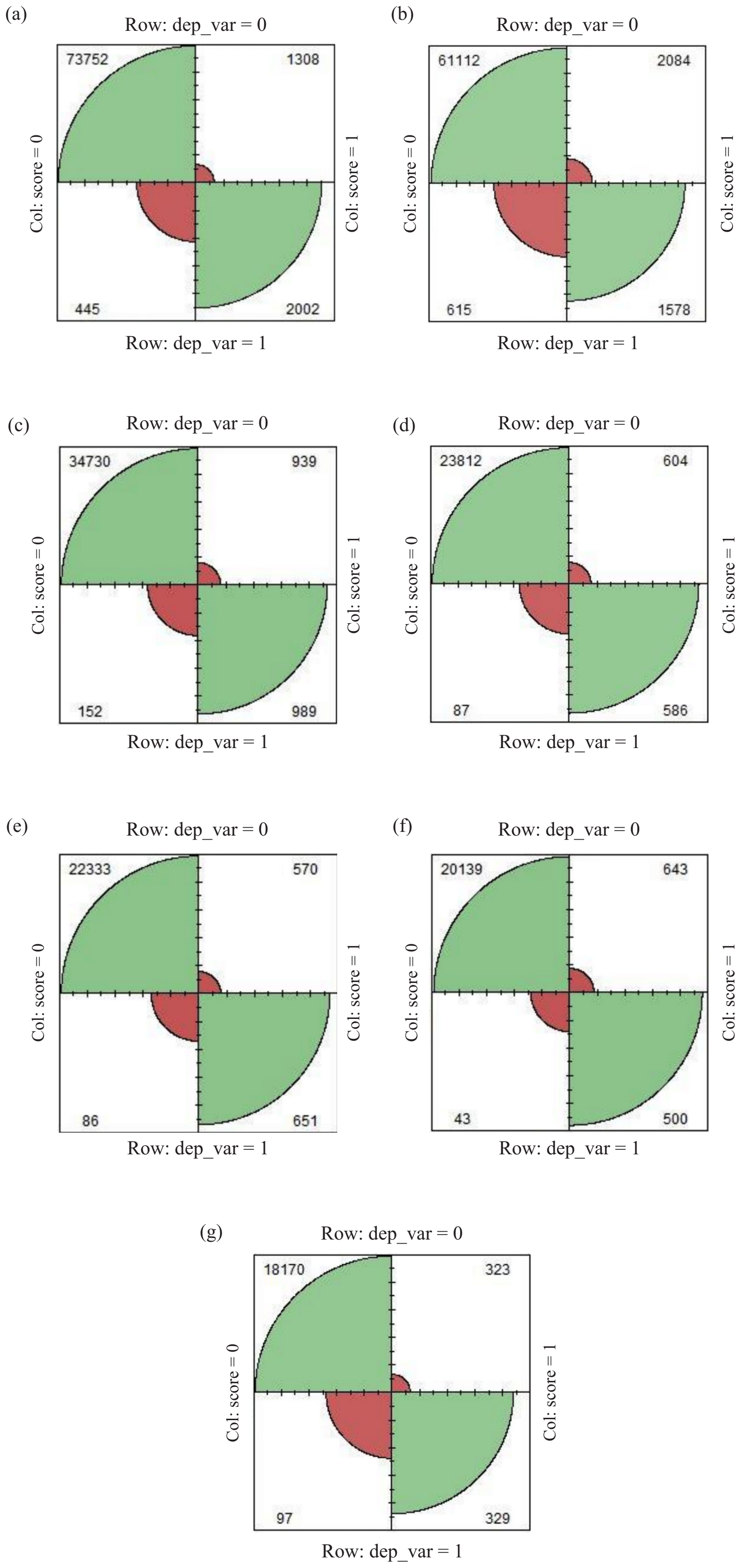


***Fig. 5.*** Graphical representation of the confusion matrix for the components: (a) 1, (b) 2, (c) 3, (d) 4, (e) 5, (f) 6, (g) 7.

Table 4
Identification of the classifier for the components 1–7

| Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Training set cardinality | 2500 | 2500 | 1600 | 1200 | 1600 | 1600 | 1500 |
| Number of churners in training set | 500 | 500 | 400 | 300 | 400 | 400 | 300 |
| Validation set cardinality | 77507 | 65353 | 36810 | 25809 | 23640 | 36810 | 18919 |
| F1 value | 0.696 | 0.539 | 0.645 | 0.629 | 0.665 | 0.593 | 0.610 |

fier. Presumably the segment consists of different kinds of users and requires further division or the use of more elastic classifier. Figure 5c-g shows the results for the other components. They are worse than for the component 1, however, smaller component size should be taken into consideration, it must be also noticed that the effectiveness of prediction is greater than the reference method.

Table 5
Identification of the classifier for the bipartite network

| | |
|---|---|
| Training set cardinality | 1500 |
| Number of churners in training set | 500 |
| Validation set cardinality | 12211 |
| F1 value | 0.507 |

Seven classifiers described above cover 89% of users. For the remaining 17889 users two additional classifiers were built. The first one was based on bipartite network containing 13711 users. Verification results for this model are shown in Table 5 and Fig. 6. They are much worse than in the previous cases. It may be explained by the fact of weaker ties between customers building it and the provider as their calls are terminated outside. For this reason, they are more likely to leave the network. In fact churn rate is significantly higher in this segment – 15.48% compared to 3.68% in the entire network. It should therefore be expected that special methods are required for churn prediction as well as for its limitation in this segment.

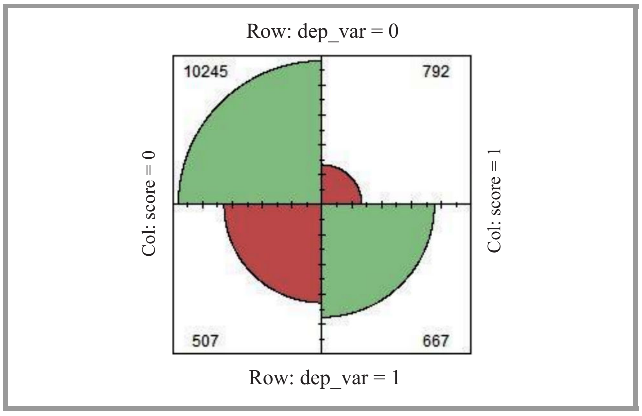


*Fig. 6.* Graphical representation of the confusion matrix for the bipartite network.

The last segment is formed of users that cannot be connected into any network. It can be assumed that they are also not similar in any other way, explaining relatively week results of prediction – see Table 6 and Fig. 7. However, it is the smallest of the segments and its impact on the effectiveness of the hybrid model is negligible.

Table 6
Identification of the classifier for the remaining users

| | |
|---|---|
| Training set cardinality | 1200 |
| Number of churners in training set | 300 |
| Validation set cardinality | 2978 |
| F1 value | 0.466 |

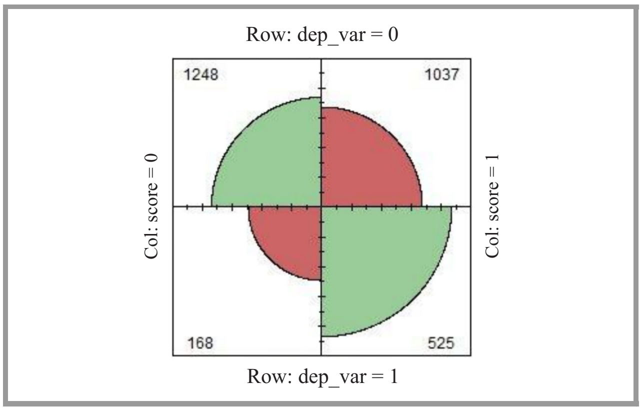


*Fig. 7.* Graphical representation of the confusion matrix for the remaining users.

### *5.5. Hybrid Model Verification*

Classification by the hybrid model is equivalent to proper use of the nine described classifiers, so verification outcome is in fact the sum of the results for individual models (see Tables 7 and 8) giving little, but visible improvement

Table 7
Identification summary for hybrid model

| | |
|---|---|
| Training set cardinality | 14800 |
| Number of churners in training set | 3500 |
| Validation set cardinality | 283868 |
| F1 value | 0.599 |

Table 8
Hybrid model confusion matrix

| Classification result | | N | P |
|---|---|---|---|
| Original classes | N | 265541 | 8300 |
| | P | 2200 | 7827 |

Table 9
Effectiveness of hybrid model components, "B" stands for bipartite network and "R" for users outside any network; bold numbers are predictors performing below average

| Components | 1 | 2 | 3 | 4 | 5 | 6 | 7 | B | R |
|---|---|---|---|---|---|---|---|---|---|
| F1 | 0.696 | **0.539** | 0.645 | 0.629 | 0.665 | **0.593** | 0.610 | **0.507** | **0.466** |

over basic regressive model. The biggest problem is still the excessive number of false positives. It is worth mentioning however, that classifiers for the components 1 and 3–7 perform better than basic classifier.

While investigating sources of imperfections the worst operating classifiers should be examined first (see Table 9). The lowest value of the F1 measure is achieved by the last classifier, acting for users who do not belong to any social network. As noted earlier this segment is related to a small group of users (a little more than 1%) so its impact on the overall classification effectiveness is also low. A slightly larger group are the subscribers forming bipartite network. They are weakly bound to the service provider and therefore it can be difficult to detect their intention of leaving. The case of component 2 seems to be different. It represents more than 20% of all users, so it cannot be easily neglected. A hint can be components visualization by LaNet-vi software using $k$-cores decomposition – see Fig. 8 for component 1, and Fig. 9 for component 2.

The $k$-core is a set of vertices of the network, each of which has at least $k$ connections with its neighbors. It is important to note that this is not equivalent to a subset of nodes that had degree $k$ in the original graph. Instead a $k$-core can be found by recursive rejection of vertices with degree lower than $k$. The colors in Figs. 8 and 9 represent the $k$-core number and circle diameter initial degree of the node. The difference between the component 1 and 2 is visible – in the first one a majority of low degree nodes encircles the center consisting little number of higher degree nodes. In component 2 much more nodes belong to the higher $k$-cores while lower degree ones are concentrated in several groups. It may be suspected that topological differences may indicate distinct characteristics of users. The analysis of node distribution among $k$-cores (see Fig. 10) shows, that while most of them follows similar, growing line, components 2 raises rapidly for $k = 5$. For the component 6 most of nodes are concentrated in the first three cores. Both components 2 and 6 are most problematic ones with F1 measure of 0.539 and 0.593 suggesting that differences in the degree structure may have important effect on churn.

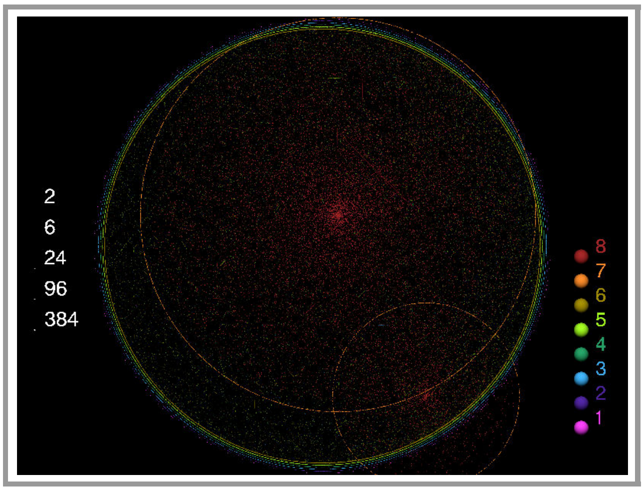


*Fig. 8.* Component 1 $k$-cores.

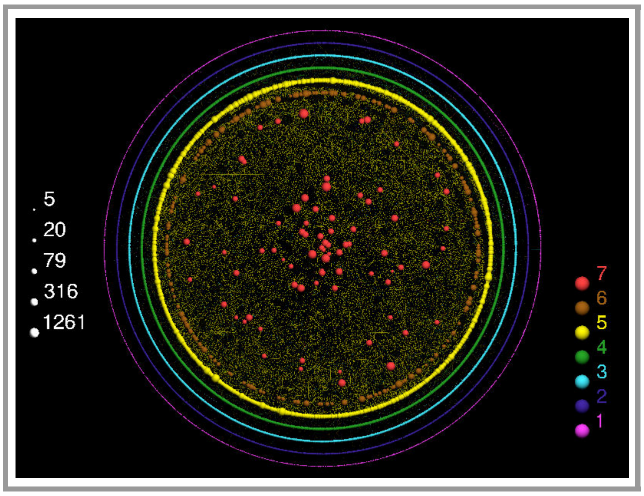


*Fig. 9.* Component 2 $k$-cores.

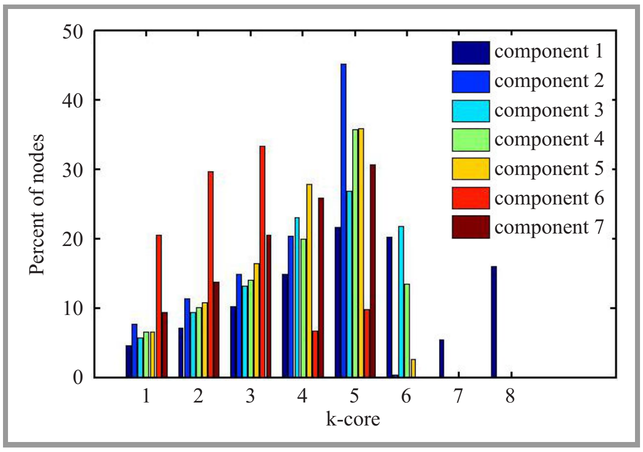


*Fig. 10.* Cardinality of $k$-cores for the inner network components.

To investigate it further triadic census[2] and clustering coefficient were calculated for all segments – see Table 10. It may be seen that component 2 differs from the remaining

[2] As the graph analyzed is undirected 4 types of triads are possible.

Table 10
Triadic census of model components

| Component | Full triads | Forbidden triads | Clustering coefficient |
|---|---|---|---|
| 1 | 75 550 | 6 477 654 | 0.0751 |
| 2 | 34 972 | 6 239 851 | 0.0691 |
| 3 | 26 205 | 3 037 690 | 0.0769 |
| 4 | 16 484 | 1 371 366 | 0.0820 |
| 5 | 15 852 | 1 208 919 | 0.0910 |
| 6 | 7 473 | 742 104 | 0.0913 |
| 7 | 9 972 | 1 191 967 | 0.0780 |
| Bipartite | 1 213 494 564 | 723 940 498 | 0.7384 |

inner network components as it has approximately twice as much forbidden triads, i.e. triads with connection between two nodes missing. Such a composition is reflected by the lowest value of clustering coefficient. Different structures of ties between users – lack of triangle closure may suggest that they are using a different kind of communication channel and therefore are more prone to churn. On the other hand the users may have no reason to communicate – e.g. their professional activity may require only contacts with a kind of a central office. In both cases it seems that local (i.e. user) value of the clustering coefficient may convey some information of user willingness to stay and could be considered the additional predictor variable.

## 6. Conclusion

The results of the presented work show that the social network analysis can be a valuable tool for customer segmentation. The approach is particularly convenient when available information is limited mainly to CDRs making it applicable for prepaid services, or a situation when the access to customer data is restricted due to the confidentiality. An additional benefit from social network analysis is the possibility to augment the set of predictor variables, with ones carrying new information, usually uncorrelated with simple statistics, and thus improving the classification efficiency.

There is no doubt that these results, although promising, are far from ideal. Particularly worrying is the high number of errors of type 1 (FP). It is possible to propose two ways to solve this problem. One of them may be fine tuning classifiers, especially by choosing predictor variables in each segment separately. Confirmation of the validity of this approach is good performance of the classifier for the component 1. The second way is to use more sophisticated classifiers – it must be stressed that the logistic regression is used because of its simplicity and ease of analysis. In this sense, presented work is a demonstration of the applicability of social network analysis to user segmentation and new predictor variables definition rather than a final assessment of its effectiveness.

Finally, to consider the analysis of the social network itself – as it was mentioned the work uses relatively simple indicators. It is possible that a deeper analysis of the characteristics of the network, in particular topological, or related to the connection dynamics [22] may result in new significant predictor variables as well as valuable observations on user behavior.

## References

[1] H. Kim and C. Yoon, "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market", *Telecommun. Policy*, vol. 28, no. 9–10, pp. 751–765, 2004.

[2] C. Borna, "Combating customer churn", *Telecommun. – Americas Edit.*, vol. 34, no. 3, pp. 83–85, 2000.

[3] G. M. Weiss, "Data mining in telecommunications", in *Data Mining and Knowledge Discovery Handbook*. Kluwer Academic, 2005.

[4] W. Gruszczyński and P. Arabas, "Application of social network to improve effectiveness of classifiers in churn modelling", in *Proc. Int. Conf. Computat. Aspects of Soc. Netw. CASoN 2011*, Salamanca, Spain, 2011, pp. 217–222.

[5] T. Mutanen, "Customer churn analysis – a case study", VTT Research Report no. VTT-R-01184-06, 2006 [Online]. Available: http://www.vtt.fi/inf/julkaisut/ muut/2006/ customer_churn_case_study.pdf

[6] V. Yeshwanth, V. Vimal Raj, and M. Saravanam, "Evolutionary churn prediction in mobile networks using hybrid learning", in *Proc. 24th Florida Artif. Intell. Res. Soc. Conf. FLAIRS-24*, Palm Beach, FL, USA, 2011, pp. 471–476.

[7] T. Sato, B. Q. Huang, Y. Huang, and M. T. Kechadi, "Local PCA regression for missing data estimation in telecommunication dataset", in *11th Pacific Rim Int. Conf. Artif. Intell. PRICAI 2010*, Daegu, Korea, 2010, pp. 668–673.

[8] J. Haden, A. Tiwari, R. Roy, and D. Ruta, "Churn prediction using complaints data", in *Proc. World Academy of Science, Engineering and Technology*, vol. 19, 2006.

[9] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks", *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", *Physical Review E*, 69, 026113, 2002.

[11] Y. Richter, E. Yom-Tov, and N. Slonim, "Predicting customer churn in mobile networks through analysis of social groups", in *Proc. SIAM Int. Conf. on Data Mining SDM 2010*, Columbus, OH, USA, 2010, pp. 732–741.

[12] K. Dasgupta, R. Singh, and B. Viswanathan, "Social ties and their relevance to churn in mobile database technoltelecom networks", in *Proc. 11th Int. Conf. on Extending Database Technology: Advances in Database Technology EDBT'08*, Nantes, France, 2008, pp. 668–677.

[13] M. Karnstedt, M. Rowe, J. Chan, H. Alani, and C. Hayes, "The effect of user features on churn in social networks", in *Proc. 3rd Int. Web Science Conf. WebSci'11*, Koblenz, Germany, 2011, pp. 1–8.

[14] M. Zawisza, P. Wojewnik, B. Kamiński, and M. Antosiewicz, "Social-network influence on telecommunication customer attrition", in *Agent and Multi-Agent Systems: Technologies and Applications LNCS*, vol. 6682, pp. 64–73. Springer, 2011.

[15] M. N. Abd-Allah, A. Salah, and S. R. El-Beltagy, "Enhanced customer churn prediction using social network analysis", in *Proc. 3rd Worksh. Data-Driven User Behav. Model. & Mining from Social Media DUBMOD'14*, Shanghai, China, 2014, pp. 11–12.

[16] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction", *Appl. Soft Comput.*, vol. 14, pp. 431–446, 2014.

[17] M. Kamola, E. Niewiadomska-Szynkiewicz, and B. Piech, "Reconstruction of a social network graph from incomplete call detail records", in *Proc. Int. Conf. Computat. Aspects of Soc. Netw. CASoN 2011*, Salamanca, Spain, 2011, pp. 136–140.

[18] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs", *Proc. 32nd Annual ACM Symp. Theory of Comput. STOC '00*, Portland, OR, USA, 2000, pp. 171–180.

[19] LaNet-vi website [Online]. Available: http://lanet-vi.fi.uba.ar/

[20] Igraph: Network Analysis and Visualisation [Online]. Available: http://cran.r-project.org/web/packages/igraph/

[21] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation", *J. Machine Learn. Technol.*, vol. 2, no. 1, 37–63, 2011.

[22] R. Kasprzyk and Z. Tarapata, "Graph-based optimization method for information diffusion and attack durability in networks", in *Rough Sets and Current Trends in Computing*, M. Szczuka, M. Kryszkiewicz, S. Ramanna, R. Jensen, and Q. Hu, Eds., *LNCS*, vol. 6086, pp. 698–709. Springer, 2010.

**Witold Gruszczyński** received his M.Sc. in Computer Science from the Warsaw University of Technology, Poland, in 2015. Currently he is with Codeflip company. His research area focuses on applications of data mining and social networks.

E-mail: wgruszczynski@codeflip.pl
Codeflip
Belgradzka st 4/40
02-793 Warsaw

**Piotr Arabas** – for biography, see this issue, p. 12.