Articles

Predicting Consumer Choices Through Analysis of Interactions in Social Networks

Todor Krastevich*

Summary:

Analysis of interactions in social networks has emerged as a new research paradigm in modern marketing. It focuses not on modeling behavior of the individual but rather than on the measurement and analysis of its relationships and interactions with other users within the network. Measurement and analysis of these interactions can help understand the structure and dynamics of social networks and their impact on consumer choice.

In this paper we present a data mining approach to measure and analyze the interactions in social networks between clients of mobile telecommunication networks. Our goal is to demonstrate how to use Call Data Records (CDR) to build predictive choice models (e.g. to predict customer churn). The approach and methodology can be applied to analyze the customer choice behavior in other markets where customer interactions are tracked automatically and saved electronically.

Keywords: social network analysis, churn management, direct marketing, predictive analytics and modeling, data mining, knowledge discovery.

JEL: M31, C52, C53

1. Introduction

Over the past two decades, mobile phone services have become a dominant means of communication. In most countries, the market size has reached a level of saturation. In late 2011, according to the International Telecommunication Union (ITU), market penetration of mobile services goes beyond 85%. (ITU, 2012, p. 3). In developed countries it even exceeds 120% (see Figure 1). This means that countries have more mobile subscribers than citizens. According to some analysts, as well as press releases from the management of some major mobile operators, the penetration of mobile services in Bulgaria exceeds 150% (i.e. there are 1.5 activated SIM cards for every inhabitant of the country), which makes it one of the most saturated markets in Europe. Under these conditions, attracting

*Associate Professor, PhD, Department of Marketing, Tsenov Academy of Economics, Svishtov; krst@uni-svishtov.bg

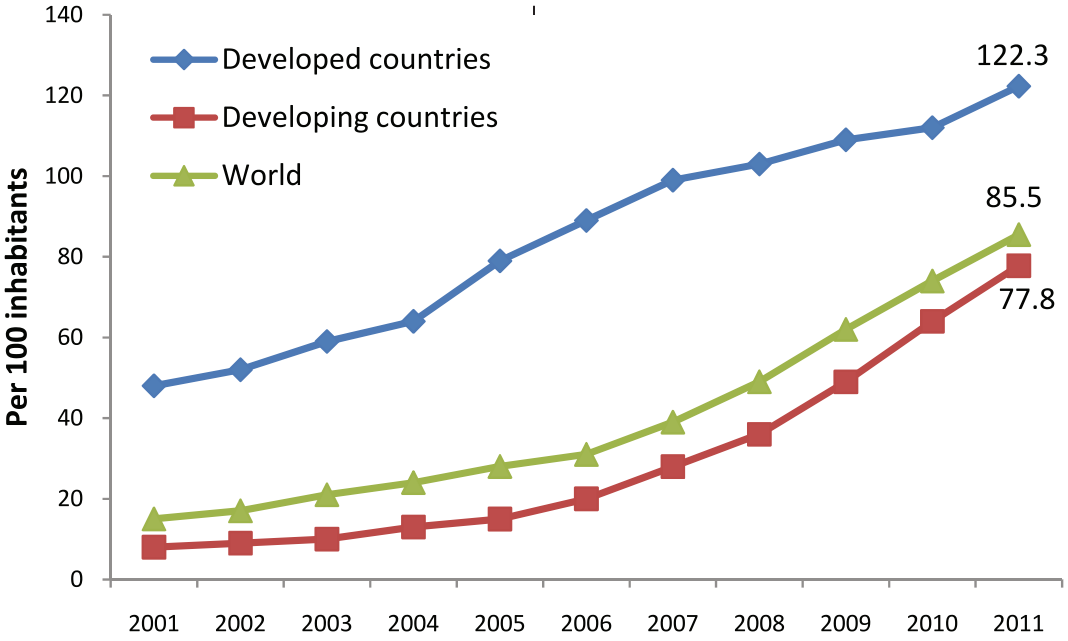


Fig 1. Penetration of mobile-broadband subscriptions, world and level development

Source: ITU World Telecommunication/ICT Indicators database.

new customers is possible only at the expense of competitors.

At the same time, standardization and clear procedures for mobile number portability allow customers to change their mobile operator relatively easily. This further exacerbates competition and leads to instability in the market shares of mobile providers. Under these conditions, marketing activities (and costs) to attract new subscribers are much less effective than efforts to retain existing customers. That is why analytical customer relationship management and, in particular, analysis of customer behaviour and predictions of customer decisions to change their mobile operator and /or switch to another tariff plan become a major challenge to the marketing of mobile telecommunication services.

2. Network interactions in customer databases

Traditional approaches for modeling and predicting consumer behavior based on customer database usually use analytical data mining methods, as each customer is treated as a separate, distinct entity. The goal is to: (1) select a sufficiently representative number of significant predictors (independent variables) that describe the behaviour of existing customers (subscribers of mobile operator), whose choice response (dependent variable) concerning switching or resigning a new contract is already observed, (2) to build and evaluate competitive predictive models (based on statistical principles or machine learning), that include the entire set of predictors and target (dependent) variable

and (3) to select the model with the highest predictive power and use it to predict the future behaviour of customers whose choice responses have not been observed yet (e.g. who consider to renew their contract). Most used predictors are: intensity of use (actual and monetary) of different types of services, payment methods and different sociodemographic and geolocation characteristics. The research is focused on the characteristics of the individual customer. Some assumptions about similar behaviour patterns are formulated based on the identified similarities between customers.

Analysis of interactions in social networks is emerging as a new research paradigm in modern marketing. It is not focused on monitoring the characteristics of the individual consumer, but rather on the measurement and analysis of its relationships and interactions with other users within the network (Easley & Kleinberg, 2010). Measurement and analysis of these interactions can help understand the structure and dynamics of social networks and their impact on consumer choice. The specifics of the research on the behaviour of the telecommunication services consumers is that their service usage behaviour depends on and is a result of technology-mediated interpersonal communication, i.e. interactions in a social network. A social network encompasses all customers of the mobile operator and the interactions between them, expressed by the frequency, duration and type (incoming, outgoing) of the calls. Interactions in the network may be regarded as directional (e.g. subscriber A initiates a conversation with the recipient subscriber B) and non-directional (A and B subscribers make calls without

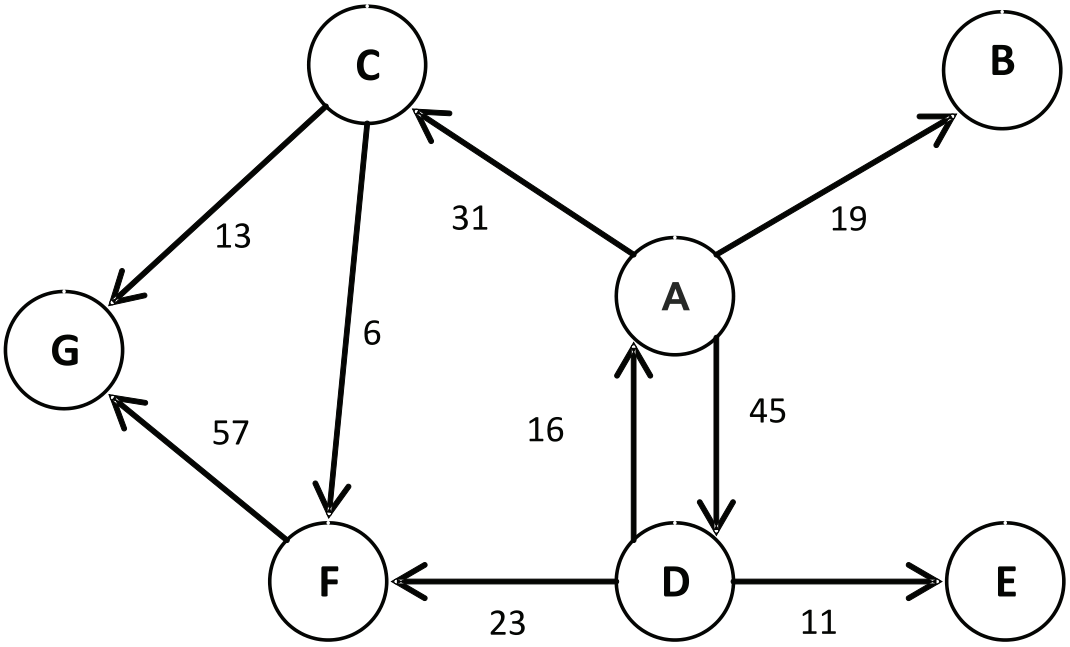


Fig 2. Social network represented as a sociogram

initiator/recipient identification). Instrumentally, interactions in the network may be coded as a dichotomous variable (e.g. subscriber A has been conducted a conversation with a subscriber) and as continuous variables (e.g. subscriber A has been conducted 20-minute conversation with the subscriber). Figure 2 presents a visual model, called sociogram (Moreno, 1954, p. 95) of a sample social network consisting of seven clients (A, B, C ..., G), that conducted a number of conversations. Directional interactions between users are marked with vectors and the intensity of these interactions – with numbers (e.g. number of calls made or duration of all calls for a specific period of time).

The real-world communication networks contain millions of subscribers and more complex interaction structures, as illustrated in the figure, but the concept of their forming is the same. Information about the structure, groups and individuals in each social network may be extracted and described by various key measures that give insight and knowledge about the behavior of the whole network (or part of it) and the individual actors (subscribers).

A key measure to describe each telecommunication network is the network density (Newman, 2010, pp. 134-135). Density (ρ) is a percentage of all possible actually recorded interactions within the network (e.g. recorded calls) and is calculated as follows:

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)},$$

where n is the number of subscribers in the customer base, m is the number of subscribers, having at least one conversation with each other. This coefficient varies from 0 to 1. The closer its value to 1, the denser is the network. Conversely, a value of 0 indicates lack of any interactions between subscribers. In high-density networks information flows between participants (including sent targeted marketing messages) are spread much more easily and quickly than in low-density networks.

The role and importance of each subscriber on the network of mobile operators is determined not only by its customer value, but by its potential to influence the retention or repulsion of other customers as well. This influence can be inferred from the number of his/her connections (respectively calls) with other potential actors. If an individual both receives calls from many people and initiates calls to many other persons, it has an influence and key role in the dissemination of information. The more isolated is the person (i.e. he/she maintains limited, unilateral interactions with few people), the lower and more insignificant is his/her influence on the social network. The total number of unique interactions (e.g. calls to other subscribers) per subscriber in the network is called the degree of centrality (Newman, 2010, pp. 68-72). Subscribers with a high degree of centrality are more active and play a significant role in the network. For directional interactions (such as phone calls) the incoming and outgoing degrees of centrality differ. Individuals who initiate more outgoing than incoming calls have higher output centrality. Individuals who receive more incoming calls (respectively initiate

less outgoing) have higher input centrality. Input centrality can be interpreted as a "prestige" of the subscriber in the network, i.e. the customer is contacted by a number of other subscribers. Outgoing centrality, in turn, is an indicator of influence. More calls a subscriber initiate to other people, the stronger is its impact on the network. Figure 2 shows that the subscriber A has the highest output centrality (i.e. the greatest impact on the network), while subscribers F and G have the same highest input centrality (prestige).

3. Customer retention management through analysis of interactions in telecommunication customer databases

Many researchers try to improve the traditional data mining approach for predicting customer churn in mobile networks through analysis of social networks (Jacob & Kerremans, 2010) (Richter, Yom-Tov, & Slonim, 2010) (Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012) (Pushpa & Shobha, 2012). Almost all of them hypothesize that interpersonal interactions have a significant and often a stronger influence on the decision to abandon the use of service compared to the traditional predictors based on descriptive data for each customer. In this study we present some instrumental aspects of social network analysis and their application in the customer retention management.

The main problem in customer retention management of mobile telecommunication services can be defined as follows: to identify customers which are likely to stop using the service (wholly, switching to another tariff plan or competitive provider) for a particular time horizon. Methodologically, the solution of this problem could be represented as a

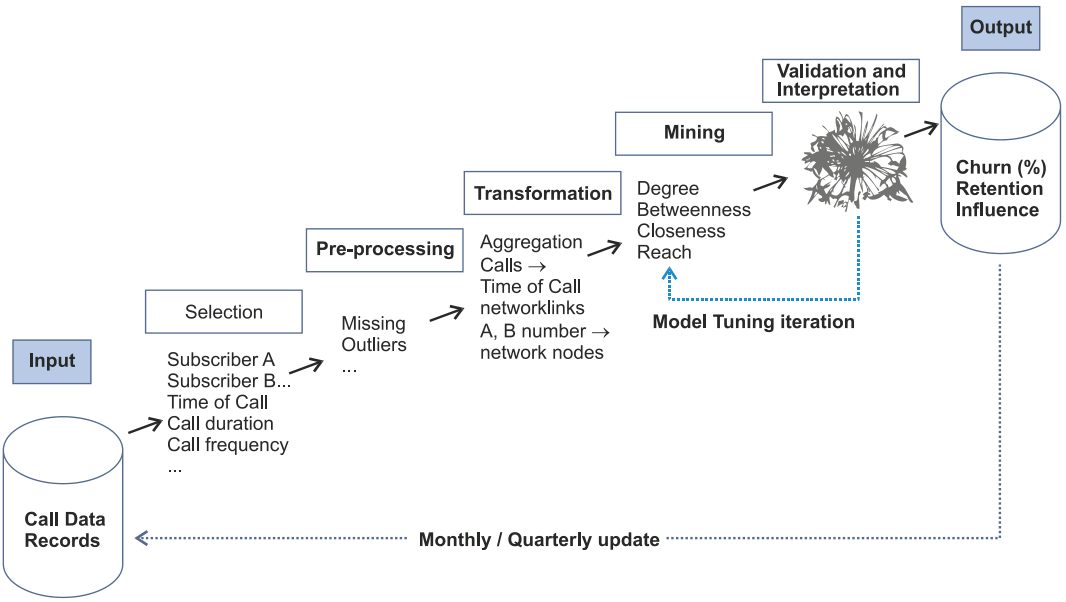


Fig 3. Data mining methodology using social network analysis in telecommunications (Jacob & Kerremans, Social Network Analysis: Decrease Churn Rate at Telecom Operators, 2010, p. 5)

multi-step process, where the focus is on data mining of the interactions between phone subscribers in the mobile network (see Figure 3) and using this knowledge to predict their future behaviour.

A key feature in the analysis of the future customer behaviour is the prediction of time of abandonment the use of service. This is critical to any mobile operator and results in customer churn. If we assume that the customer base of a specific mobile operator is 2 million subscribers and each of them pays an average monthly fee of 10 BGN, then if only 0.5% of subscribers tend not to renew their contracts during the next period (month), the company's revenues will fall by 200,000 BGN in each subsequent month. It is widely believed that the annual dropout rate varies between 50 and 70 percent for the prepaid mobile segment. Even a modest reduction in this rate would have a significant monetary impact on the company, because the cost of retaining an existing customer is approximately five times lower than the cost of attracting a new one.

In general, the input data of this analysis include records of incoming and outgoing calls within a particular time horizon in the past and their corresponding personal and /or business data of the subscribers. It is also possible to include specific indicators for each subscriber who has abandoned the use of service within the observed period and/or changed his/her tariff plan and/or service package. These data are useful for training and testing predictive models.

4. Research approaches for analyzing customer interactions in database

There are two competitive approaches for social network analysis applied on the mobile operators' customer base: group-based analysis (Richter, Yom-Tov, & Slonim, 2010) and diffusion-based analysis (Dasgupta, et al., 2008).

Group-based analysis is focused on identifying homogeneous groups of mobile network customers through studying the structure of interactions between them. The assumption is that the characteristics of the identified homogeneous customer groups influence individual behavior of each group member. For example, if within a relatively small group of customers with a high density of interpersonal communications could be identified distinct group leaders (i.e. individuals with high centrality with whom all the members of the network communicate intensively and who have a strong influence on others), targeted marketing effects on these subscribers could reduce dropout rates and /or influence other subscribers more effectively.

Table 1. Database structure of call data records

Initiator (e.g. telephone number or other identifier of the outgoing call)	Recipient (e.g. telephone number or other identifier of the incoming call)	Weight (i.e. duration or monetary value of the conversation)
A	B	21
A	C	2
C	A	19
B	C	4
...	...	...

Diffusion-based analysis is focused on identifying those subscribers who are most influenced by other subscribers in the mobile network. The aim is to quantify the strength of influence of interpersonal influences.

Both approaches require call data records to be organized as follows:

The database could contain additional variables such as time of the recorded call, specific sociodemographic or geolocation characteristics of the initiator and recipient, their previous responses (e.g. how the person have been responded to previous promotional activities, whether he/she is a current or former customer, what type of device he/she uses, etc.).

Group-based analysis

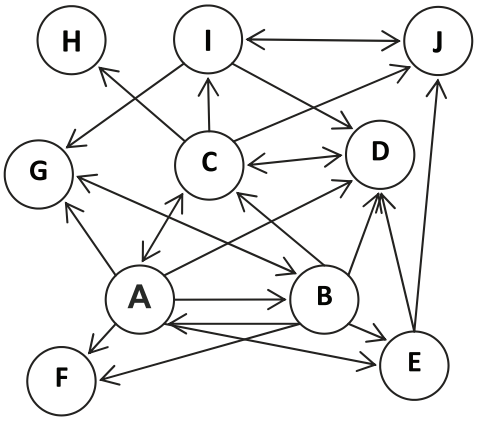
Conducting group-based analysis involves three steps (IBM Corporation, 2012, p. 11):

(1) Determining the similarity between subscribers based on the nature and strength of the interactions between them.

(2) Subdividing the network into groups based on the strength of interactions, taking into account a predefined threshold value of the group size.

(3) Profiling groups and subscribers and identifying group leaders.

First step. Unlike traditional clustering approaches, where the degree of similarity (respectively dissimilarity) between the objects is derived based on the differences in their descriptive characteristics (i.e. compositional variables), in the group-based algorithm two objects could be qualified as similar through analysis of structural variables. Structural variables contain data (measurements) for all possible pair of subscribers having at least one interaction with each other (i.e. at least one outgoing call). Two objects (subscribers) are qualified as similar if they interact with the same group of other subscribers. In other words, the degree of similarity between subscribers is measured by the degree of overlap in their interactions with other subscribers. Figure 4 illustrates a graphical and tabular representation of this measurement concept.



Initiator	Recipient
A	B, C, D, E, F, G
B	A, C, D, E, F, G
C	A, D, H, I, J
D	C
E	D, J
F	-
G	B
H	-
I	D, J
J	I

Fig. 4. Illustration of a network with ten subscribers (adapted from IBM Corporation, 2012, p. 12)

There is high degree of similarity in outgoing calls of subscribers A and B – in five of the six outgoing calls the recipients are the same (5/6 = 83.333%). For subscribers E and I this matching coefficient is 100%, but the informational contribution to its calculation is smaller (only two recipients). The magnitude of the "mutual information" is used to calculate all possible similarity (dissimilarity) relations. Mutual information I is a symmetric function of two random variables (X and Y) and describes the amount of information contained simultaneously in both variables. In this case it is used as a measure of dependence between two random variables X and Y (a concept, similar to the statistical correlation). It is symmetrical to the X and Y, always takes nonnegative values and is equal to zero only if X and Y are independent. It can be displayed as unconditional (for non-directional interactions) and conditional (for directional interactions). In the case of nominal variables (such is the present case) mutual information is expressed as follows:

$$(X;Y) = \sum\sum p\left(x;y) \log\left(\frac{p(x,y}{p(x)p(y)} \right) \right),$$

where p (x, y) is joint probability distribution function of X and Y, p (x) and p (y) are marginal probability distribution functions of X and Y respectively. Abstractly speaking, mutual information represents the likelihood of two subscribers to be connected with the same third subscriber. In calculating the mutual information between two subscribers additional data (e.g. length of calls, intensity of contacts within a particular time horizon) could also be used (as weights).

Second step. Subdividing the network into groups is based on the identified similarities between the customers. The aim is to identify homogeneous groups which include customers who interact intensively with each other (i.e. they have "strong" intragroup interactions). For this purpose, two thresholds are defined simultaneously. First, the objects (subscribers) who have "weak" interactions with the network as a whole are removed from the database. This concerns coverage threshold that distinguishes between "strong" from "weak" interactions. Coverage threshold varies between 0 and 1. Value of 0.25 means, for example, that for the group identification will be used only the strongest 25% of all interactions between the subscribers (i.e. the remaining 75% will be isolated). Second, as it is possible to find a lot of small or large groups with low predictive power within the remaining set of interactions, it is reasonably to predefine a threshold value of the minimum and maximum group size. Therefore groups that are too small are isolated completely from the analysis, while large groups are divided into smaller, according to the predetermined thresholds. Following such a restrictive approach (coverage and group size thresholds) core groups are identified. It is possible some of the subscribers to be omitted from any of the identified core group. However, if they have interactions (conducted incoming or outgoing calls) with members of the core groups (core members) they are further joined to these groups. Formal join criteria are relatively strong relationships with members of the core groups and

covering threshold value of the minimum and maximum group size. The final result is groups of subscribers consisting of core members and additionally joined members based on "strong" interactions.

Third step. Group profiling and identification of group leaders is the third important aspect of group-based analysis. Among the main measures such as density (for each identified group) and centrality (incoming and outgoing) for each subscriber could be assessed the status and influence of individuals within each group (i.e. group dynamics). As the role of each subscriber in the group is critical (and could be used to predict the behaviour of the entire group and its members), for its quantitative measurement could be used two measures – authority and dissemination. The authority of each subscriber is estimated through the tendency of other group members to interact with him/her, i.e. how many of the calls initiated within the group are focused on him/her. A useful algorithm for assessing the authority is the John Kleinberg's approach (Kleinberg, 1999), known as HITS algorithm (Hyperlink-Induced Topic Search), resulting in authority value for each subscriber from 0 to 1. It is assumed that the subscriber with the highest score in the group is an authoritative leader. The strength of dissemination of the leader on the group could be expressed as a ratio of the highest and lowest score of authority among the members of the group. Vice versa, a subscriber that initiates outgoing calls to other group subscribers could be used as an indicator of dissemination on the entire group opinion. Formally, this tendency could be expressed for each subscriber by the number of his/her outgoing calls to other group members. Evaluation and interpretation of this measure is similar to the authority – if its value tends to 1, the ability of a subscriber to disseminate information within the group is higher. Subscriber with the highest score is a dissemination leader of the group. The ratio of highest and lowest score among the group members indicates the overall influence of the dissemination leader.

In addition to the key measures of group dynamics analysis, within the social group may be derived a number of ancillary statistics such as: total number of subscribers in each identified group, total number of interactions in each identified group, total number of identified groups, average number of subscribers in a group, density of identified groups, average share of the core members of the identified group, density of core groups (average share of direct interactions between the core subscribers), average input and output centrality, etc.

In addition to group dynamics analysis at aggregated level (customer groups) an analysis of the individual subscribers could be also conducted. Among the key measures for individual level analysis are: role of the individual customer (whether he/she is a "core" member of the group or not), customer authority, his/her position within the group based on his/her authority score, his/her ability to disseminate information (dissemination), its position within the group based on his/her dissemination score, number of calls in which the particular subscribers is a recipient, number of calls in which the particular subscriber is an initiator, whether the particular subscriber is an authoritative leader or a dissemination leader, etc.

Diffusion-based analysis

The purpose of the diffusion (or spreading activation)-based analysis is to identify those subscribers who are most strongly influenced by others within the social network. The strength of this effect is referred as diffusion energy. Algorithm to identify customers with the highest diffusion energy is as follows (Dasgupta, et al., 2008, pp. 671-672):

(1) Identifying the subscribers who have been responded, within a specified time horizon, in a manner affecting the mobile provider (e.g. who have terminated their contracts or responded to a message) and monitoring their outgoing and incoming calls within the preceding period.

(2) Using the spreading activation technique originally proposed by cognitive psychology (Collins & Loftus, 1975) and later described as a computer algorithm (Ziegler & Lausen, 2005). The logic is that every subscriber who has responded within the observed time horizon is qualified as "activated" and receives a certain weight (initial energy), which is iteratively spread to the associated subscribers and "activated" them by transferring some of the energy. The share of the transmitted energy of each subscriber is called spreading factor and is considered to be a constant parameter. The higher is the spreading factor value, the more distant subscribers could be activated. The energy that receives each subsequent activated subscriber depends on the strength of his/her relationships with already activated subscribers. Part of the diffusion activation energy that a particular subscriber-recipient receives is equal to the ratio of weight (strength) of his/her relationship with the subscriber-initiator and weight (strength) of all outgoing relationships of the initiator. Therefore, if a given subscriber has more intensive interactions with an initiator (e.g. talking with a given initiator more often and longer), he/she will receive a greater influence (i.e. more of the diffusion energy) by the initiator, in comparison to anyone who supports weaker interactions. Activation can be spread through different marked routes. This process is terminated when a given activated subscriber is not an initiator of outgoing calls, when all the initial energy within the diffusion process reaches the minimum threshold value and/or when it reaches a predefined number of iterations.

(3) After the termination of the diffusion process the subscribers with the largest share of transferred diffusion energy are identified. These subscribers are the most sensitive and susceptible to influence and they should be of a special interest to the researcher.

This approach could be illustrated using data from Table 1, associated with a fragment of the network of interactions between seven subscribers of a hypothetical mobile provider (see Figure 5).

Let us assume that customer A has terminated his/her subscription contract at

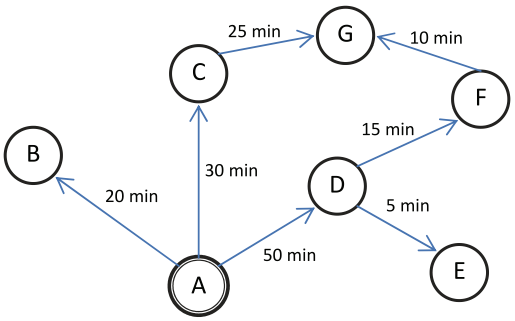


Fig. 5. Example of interactions between mobile subscribers

a given moment. Following the logic of the diffusion-based analysis, in him/her should be concentrated all the initial diffusion energy that ultimately they should influence the subsequent decisions of entities related with him/her. We assume that the magnitude of this energy is equal to 1.00 and that the spreading factor is 0.8 (the value of this factor may vary and it is a subject to simulation, respectively, what-if analysis). As already mentioned, the diffusion-based algorithm uses an iterative approach. At the first iteration the "activated" subscriber "A" spreads 80% of the initial diffusion energy between the subscribers B, C and D, whom he has conducted phone calls with. The distribution of these 80% of the initial energy between the three subscribers is based on the "weight" of the recorded interactions. In the current example, the length of the calls in minutes is used as a weight (see Figure 5), but it is possible to use other bases (e.g. frequency of the recorded outgoing calls). The remaining 20% of the energy is "retained" for the initiator. Following this rule, subscriber D receives half of the distributed energy (i.e. 0.40), as the length of the outgoing calls between A and D is 50 minutes, the subscribers C and B receive 30% (0.24) and 20% (0.16) respectively.

At the second iteration the activated subscribers B, C and D should retain 20% of the received diffusion energy each and to spread the remaining 80% among the recipients related with them. Since the subscriber B has not any registered outgoing calls, he/her retains all of the received diffusion energy (0.16). On the other hand, subscriber C spreads 80% (0.19) to subscriber G, and subscriber D spreads 80% of his/her energy between subscribers E and F in ratio 1:3 (respectively, 0.08 and 0.24).

At the third iteration, the activated subscribers E, F and G would have to spread their diffusion energy again to the recipients associated with them. Since the subscribers F and G have not registered outgoing calls (see Figure 5), they retain all the accumulated diffusion energy (respectively, 0.8 and 0.19). However, subscriber F spreads 80% of his/her energy to subscriber G (i.e. 80% of 0.24 = 0.19). The total of the diffusion energy accumulated in subscriber G (distributed from subscribers C and F) is 0.39. If the subscriber G has not any other registered outgoing calls, the spreading activation process is terminated. Thus the initial diffusion energy of the subscriber A is spread on the network and the subscriber G is most influenced by it. The interpretation is that if this energy reflects a recorded response (e.g. contract termination or responding to a particular

Table 2. Example of spreading activation technique used for diffusion-based analysis of social networks

Iteration	Subscriber A	Subscriber B	Subscriber C	Subscriber D	Subscriber E	Subscriber F	Subscriber G
0	1.00	0	0	0	0	0	0
1	0.20	0.16	0.24	0.40	0	0	0
2	0.20	0.16	0.05	0.08	0.08	0.24	0.19
3	0.20	0.16	0.05	0.08	0.08	0.08	0.38
...	...	...	...	...	...	...	...

Source: (IBM Corporation, 2012, p. 20)

advertising message), then the subscriber G should be influenced most by the subscriber's A response.

While implementing the above algorithm on a real database it is possible to record two types of descriptive measures – measures that describe diffusion at an aggregate level (i.e. customer database as a whole) and measures of the diffusion process at an individual level. As an aggregate level measures can be assigned: number of customers, number of interactions within the network, number of subscribers, which are used as starting elements of the diffusion process, average diffusion energy of a subscriber, average number of incoming and outgoing calls to other subscribers. More significant for marketing decisions are the measures at an individual level. At an individual level measures can be assigned: total diffusion energy accumulated by a given subscriber, number of incoming and outgoing messages for each subscriber. The most important measure is the accumulated diffusion energy. It is an indicator of the influence that an activated subscriber can have on any other subscribers that are indirectly connected with him/her. For example, this is a key indicator in formulating models for predicting the subscribers with high propensity to terminate their contracts.

5. Some guidelines for implementation

In technical terms, social networks analysis may be applies as: (1) an independent method for studying and predicting the behaviour of the customers of telecommunications services (e.g. based on existing records of

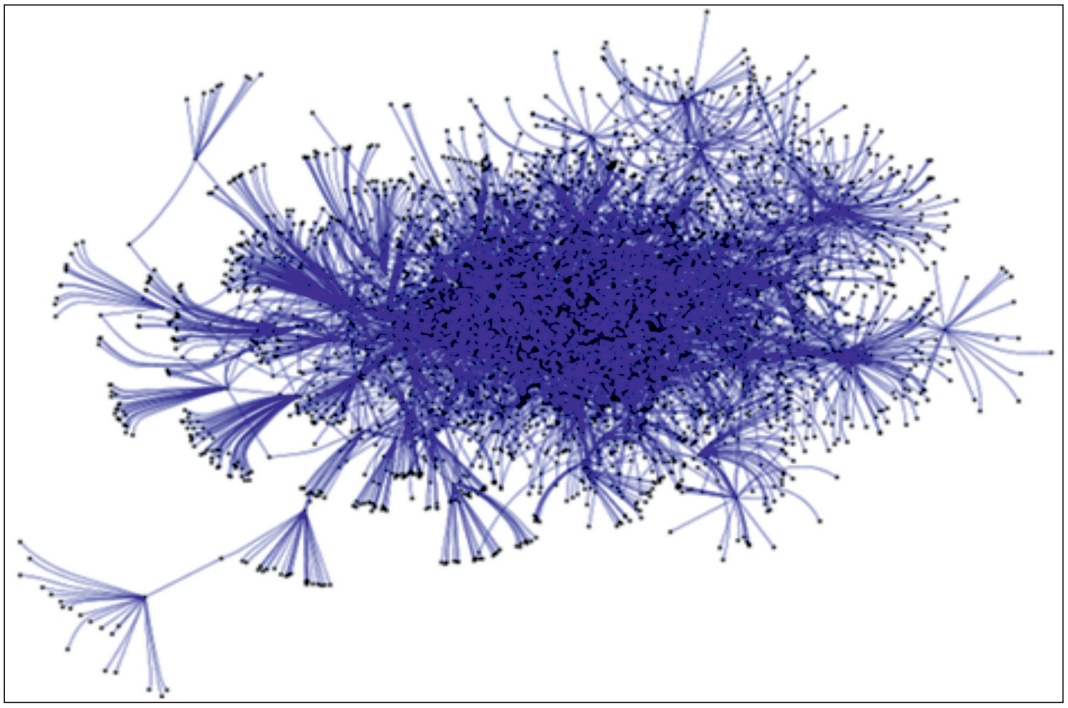
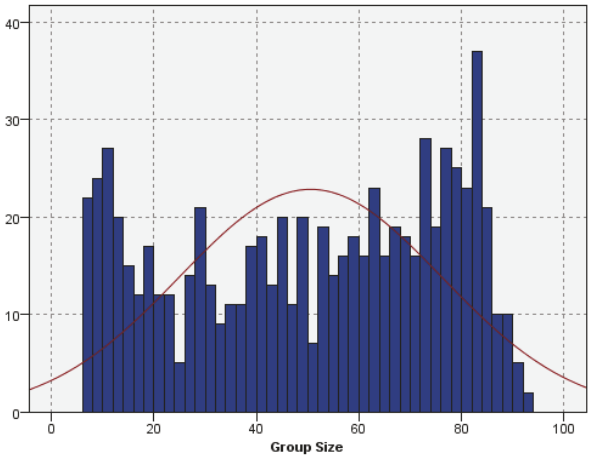


Figure 6. Visualization of the interaction between mobile subscribers

Table 3. Indicators to describe the group structure

Description of identified groups

Number of subscribers in groups	36 666
Number of conversations within groups	1 659 529
Number of identified groups	755
Average group size	50,02
Average density of group	0,75
Average share of core subscribers in group	0,56
Average density of the core group	0,43
Average input centrality	45,26
Average output centrality	45,26



the outgoing and incoming calls to derive the conditional probability that if a given subscriber terminates the contract with his/her mobile provider, the other subscribers that are related with him/her also is possible to switch). The method could be applied as (2) a priori tool to derive meaningful predictors that could be used for formulating more accurate predictive models. In both cases, the aim of the analysis is to increase the effectiveness of direct marketing campaigns through more precise targeting.

To illustrate the application of group-based analysis a fraction of a real customer database of an anonymous mobile telecommunications services provider is used. The database is structured as shown in Table 1. It contains 50,964 subscribers, among which were detected 2,701,699 conversations. In Figure 6 is illustrated a sociogram. It is constructed using the raw data.

During the group-based analysis two types of key measures are calculated – measures of group dynamics and measures for each subscriber. As stopping criteria for the group formation are assigned the following values: coverage threshold of 10%, minimum group size of 7 and maximum group size of 100 subscribers. Following these stopping criteria, 733 groups are identified containing 36,666 subscribers (conducted together 1,659,529 conversations). The description of the group structure and a histogram of the identified groups according to their size are presented in Table 3.

The detailed analysis of group dynamics and the description of characteristics of individual subscribers are possible using two types of measures. The values of the first ten customers from the first identified group (column 1) that contains 59 subscribers (column 2) is displayed in Table 4. Each of

Table 4. Key indicators to describe the group dynamics and the individual subscribers (fraction of the first ten records from the first identified group)

	Key Indicators for group dynamic										Individual level key indicators											
Group number	Number of individuals in a group	Group density	Fraction of direct connections between core individuals in a group	Fraction of individuals in a group that are core individuals for the group	Maximum authority score of any group member.	Minimum authority score of any group member	Ratio of the largest authority score to the smallest. (authority strength)	Maximum dissemination score of any group member	Minimum dissemination score of any group member	Ratio of the largest dissemination score to the smallest.	Unique identifier for an individual (e.g. Phone number)	Indicator of whether the individual is a core individual for a group or not	Authority score for the individual	Rank order in the group based on the authority scores	Dissemination score for the individual.	Rank order in the group based on the dissemination scores	Number of relationships in which the individual is the target of the relationship	Number of relationships in which the individual is the source of the relationship	Whether the node is an authority leader, whose leadership score is derived from incoming links	The confidence that the node is an authority leader	Whether the node is a dissemination leader, whose leadership score is derived from outgoing links	The confidence that the node is a dissemination leader
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(2	(22)	(23)
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0631	1	0.01	11	0.01	29	50	48	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0631	1	0.01	5	0.01	48	52	46	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0631	1	0.01	27	0.01	4	48	52	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	6	0.01	54	52	45	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	48	0.01	33	46	48	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	32	0.01	59	48	41	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	7	0.01	32	52	48	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	42	0.01	56	47	44	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	14	0.01	25	50	48	0	0.0	0	0.00
1	59	0.82	0.48	1.00	0.01	0.01	1.36	0.01	0.01	1.26	0632	1	0.01	35	0.01	44	48	47	0	0.0	0	0.00
…	…	…	…	…	…	…	…	…	…	…	…											

these indicators may be used as a predictor in a model for predicting the behavior of the subscribers. However, it is necessary to monitor the past choice behavior (e.g.

who of the observed subscribers have been abandoned the use of certain services). These data are summarized in a variable that is usually dichotomously coded (e.g. 1 = terminated the contract; renew the contract = 0). An interesting predictor is the group membership variable. The hypothesis is that if a given group loses a larger proportion of customers than the average group proportion, it can be defined as a group "at risk" of dropping out its customers. Identifying the groups "at risk" and those with no risk to drop out its customers (e.g. binary or polynomial dependent variable), this variable could be used as a target variable in a full predictive model (e.g. using CHAID, naive Bayesian classifier or logistic regression). Thus a group membership could be predicted (proneness to terminate /renew the contract) for all other subscribers within the database, whose choice decisions are not known.

As a next step of the analysis, it is possible to calculate the strength of influence of each subscriber of the network at "a risk" of dropping out customers (or in general, to respond in a manner affecting the operator) using diffusion-based analysis. These list of subscribers need to be combined with the initial database of the incoming and outgoing calls (subscriber's

Table 5. Key indicators for the evaluation of diffusion processes at an individual level (fraction of the first ten subscribers)

Unique identifier for an individual (e.g. Phone number)	**Churn (1 = yes, 0 = no)**	**Amount of diffusion energy associated with the individual. Higher values indicate a greater propensity to churn than lower values**	**Number of relationships in which the individual is the target of the relationship**	**Number of relationships in which the individual is a source of the relationship**
(1)	(2)	(3)	(4)	(5)
42583	1	0.411	16	14
42280	1	0.390	35	36
51272	1	0.388	65	50
37413	1	0.376	38	38
44337	1	0.375	54	52
47908	1	0.374	57	57
13848	1	0.373	19	19
44352	1	0.371	59	53
5157	1	0.370	41	40
…	…	…	…	…
16279	1	0.306	68	68
43465	1	0.305	67	72
2885	1	0.305	65	59
41652	0	0.134	13	9
32982	0	0.125	8	5
1013	0	0.119	11	12
41668	0	0.119	14	7
35316	0	0.109	38	38
42591	0	0.109	13	12
13853	0	0.106	17	16
…	…	…	…	…

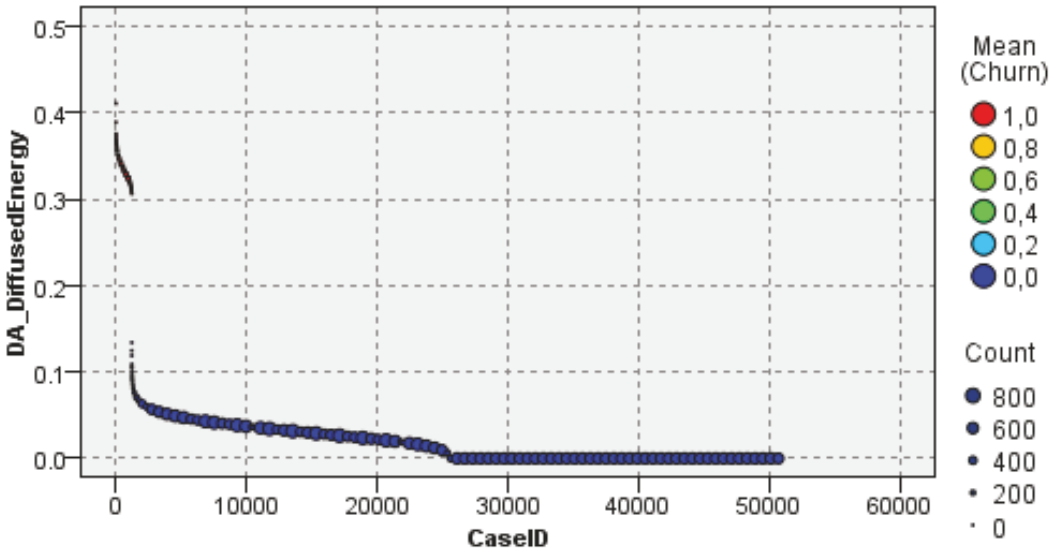


Figure 7. Distribution of subscribers according to the diffusion energy (small fraction of them have high diffusion energy and all of them have terminated their contracts)

ID is used as a key variable). Key indicators for some of the subscribers are calculated at a spreading factor of 0.7 (i.e. 70%). These indicators are presented in Table 5. The data are sorted in descending order according to their diffusion energy (column 3). It is clear that the magnitude of the diffusion energy corresponds strongly to the subscriber's choice – customers that have terminated their contracts have high diffusion energy. The threshold at which is situated the demarcation line (line in gray in Table 5) of the decision to continue or terminate the contract is between the values of 0.305 and 1.225. The full distribution of the subscribers is presented in the Figure 7.

Again it is important to note that the results of group- and diffusion-based analyses could be used as useful inputs (predictors) of more sophisticated predictive models.

Besides the demonstrated applications group- and diffusion-based analyses could be used to increase the effectiveness of viral marketing strategies and campaigns. Viral marketing is a marketing technology based on the use of social networks to increase the brand and achieve other marketing goals (e.g. increasing sales). The group-based analysis techniques are suitable analytical procedures for planning and evaluating the impact of a successful viral marketing strategy. For example, by identifying the group leaders, dissemination leaders and authoritative leaders it is possible more precisely to target marketing communication messages and faster to spread the viral effects. Direct impact on these individuals with a product or pricing information increases network effects and the effectiveness of the marketing campaign. Moreover, using the diffusion-based analysis (applied to the identified opinion leaders) it is possible to quantify the effect of information dissemination and to predict the likelihood that other network customers could purchase a given product or service.

6. Conclusion

Through the analyses of the interactions between the subscribers in mobile communication networks it is possible to extend and complement the traditional solutions for studying and predicting their responses. The application of tools for social networks the analysis (group- and diffusion-based analyses) allow for operators to build more effective "early warning" systems, proactively to identify the potential churners and to increase the retention rate of their customer base.

Social network analysis offers insights for the mobile operators on how to reduce customer churn and costs and how to increase the effectiveness of direct and viral marketing campaigns through better targeting of the customer base.

References:

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407-428.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., & Joshi, A. (2008). Social Ties and their Relevance to Churn in Mobile Telecom Networks. EDBT '08 Proceedings of the 11th international conference on Extending database technology: Advances in database technology (pp. 668-677). NY: ACM.

Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets. Cambridge: Cambridge University Press.

IBM Corporation. (2012). IBM SPSS Modeler Social Network 15. IBM Corporation.

ITU. (2012). Measuring the Information Society. ITU, Telecommunication Development Bureau. Geneva: International Telecominication Union.

Jacob, R., & Kerremans, P. (2010, March). Social Network Analysis: Decrease Churn Rate at Telecom Operators. NimzoSTAT.

Kleinberg, J. (1999, Sept.). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.

Moreno, J. L. (1954). Who Shall Survive?: Foundations of Sociometry, Group Psychotherapy, and Sociodrama. N.Y.: Beacon House Inc.

Newman, M. J. (2010). Networks. An Introduction. N.Y.: Oxford University Press.

Pushpa, & Shobha, G. (2012, May). An Efficient Method of Building the Telecom Social Network for Churn Prediction. International *Journal of Data Mining & Knowledge Management Process* (IJDKP), 2(3), 31-39.

Richter, Y., Yom-Tov, E., & Slonim, N. (2010). Predicting customer churn in mobile networks through analysis of social groups. SIAM International Conference on Data Mining - SDM, (pp. 732-741).

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211-229.

Ziegler, C. -N., & Lausen, G. (2005). Spreading Activation Models for Trust Propagation. *Information Systems Frontiers*, 7(4/5), 337-358.