Identifying influencers in a social network: the value of real referral data

I. Roelens[a,b,1], P. Baecke[b], D.F. Benoit[a]

[a] *Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Ghent, Belgium*
[b] *Vlerick Business School, Reep 1, 9000 Ghent, Belgium*

Abstract

Individuals influence each other through social interactions and marketers aim to leverage this interpersonal influence to attract new customers. It still remains a challenge to identify those customers in a social network that have the most influence on their social connections. A common approach to the influence maximization problem is to simulate influence cascades through the network based on the existence of links in the network using diffusion models. Our study contributes to the literature by evaluating these principles using real-life referral behaviour data. A new ranking metric, called Referral Rank, is introduced that builds on the game theoretic concept of the Shapley value for assigning each individual in the network a value that reflects the likelihood of referring new customers. We also explore whether these methods can be further improved by looking beyond the one-hop neighbourhood of the influencers. Experiments on a large telecommunication data set and referral data set demonstrate that using traditional simulation based methods to identify influencers in a social network can lead to suboptimal decisions as the results overestimate actual referral cascades. We also find that looking at the influence of the two-hop neighbours of the customers improves the influence spread and product adoption. Our findings suggest that companies can

Email addresses: iris.roelens@ugent.be (I. Roelens), philippe.baecke@vlerick.com (P. Baecke), dries.benoit@ugent.be (D.F. Benoit)

[1] ICM-FWO Fellow of the Research Foundation - Flanders

take two actions to improve their decision support system for identifying influential customers : (1) improve the data by incorporating data that reflects the actual referral behaviour of the customers or (2) extend the method by looking at the influence of the connections in the two-hop neighbourhood of the customers.

Keywords: Influence maximization, social network, customer referral, Shapley value

1. Introduction

Customers are crucial assets for a firm but they can be costly to acquire. The focus of this study is on customer acquisition, which is of utmost importance to any organisation. Ensuring the inflow of customers to be larger than the outflow so that the customer base increases is not at all straightforward. As a result, marketers are in a continuous battle for attracting potential customers' attention and getting into their consideration set. Many have shifted part of their marketing efforts portfolio from directly communicating with potential customers to incentivizing existing customers to do so [16]. This is driven by the growing acceptance of the fact that people are highly influenced by information received from others [17] and that word-of-mouth (WOM) is the most influential source of information to a customer [21]. Empirical research confirmed that consumers rely heavily on the advice of others in their personal network when making purchase decisions [19, 36, 32, 20, 33] and that positive WOM has a positive effect on business outcomes, i.e. sales [31, 3]. Referral marketing has become an important marketing technique to stimulate WOM in a controlled way for acquiring new customers [5]. A good example of referral marketing success is Dropbox. They managed to expand their customer base from 100,000 to 4 million users in a 15-months period by leveraging the power of referrals. Prior to using referral marketing, Dropbox was using Google's AdWords and affiliate marketing, with a cost of acquisition between $288 and $388 per individual [1]. Dropbox's CEO, Drew Houston, calculated the cost of acquiring this large customer base at $10 billion if traditional marketing programs had been used [4]. As a consequence, leveraging social influence can greatly decrease the costs of acquiring new customers.

Suppose we have data on the social network of our customers, in which the interactions give an indication of how influence flows between the individuals. If we want to attract as many new customers as possible by relying on the power of social influence, we want to initially target only a few individuals whom we expect to trigger a cascade of influence in which friends recommend the product to other friends. The key question is how to select those initial influencers who will seed this process. In order to do that, managers need to have an intelligent system that supports them in finding the optimal group of influential customers. In literature, selecting a group of individuals who are most likely to generate the largest cascade of influence through WOM is also known as the influence maximization problem [22]. Multiple approaches to solve the influence maximization problem have been developed. However, these algorithms typically are not based on data that represent influence flow as it is not straightforward to gather such data set. Rather, they simulate influence spread in a social network at random based on the links that exist in the network according to diffusion models [28, 8, 7, 9].

This work focuses on how the most influential customers can be identified and how well such methods perform compared to actual referral behaviour. In this paper, we combine social network data and actual referral behaviour data. Social network data is used in the form of a telecommunication network that represents call and text interactions. The referral data set comprises the same set of customers which can be represented as a referral network. In this work we will answer the following questions: (1) What is the value of simulation-based methods to select the top influencers in a customer network? (2) To what extent does a weighted approach, taking into account interpersonal connection strength, improve on a non-weighted approach? (3) What is the value of using actual referral data for selecting top influencers in a customer network? (4) Can the method be improved by not only considering the direct influence potential of the seed customers, but also the second-hop connections of the customers?

We propose a novel method based on the game-theoretic concept of the Shapley value to compute an influence score for every existing customer. This list of scores then allows managers to select the top *k* most influential customers to be involved in a customer referral program. We contribute by assessing the quality of the often-used simulation-based methods that find the top influencers by simulating influence spread through a network based on the links in the network. The unique data set used in this study enables evaluating the performance of this type of method by comparing the estimated influence spread of the top *k* influencers with their actual referral behaviour. As such, this study bridges the more design-oriented research from the field of information systems research and the more empirical papers from the field of marketing. In that way this paper contributes to what Probst et al. [29] propose in their comprehensive literature study as a necessary addition to influence maximization literature. Additionally, we add to the literature by responding to the need for simulations that can quantitatively be compared with specific social phenomena as pointed out by Conte et al. [12].

The remainder of the paper is organized as follows. We introduce the referral marketing model in Section 2. Section 3 gives an overview of related work. In section 4, the proposed methodology is discussed and Section 5 summarizes the results. Section 6 concludes this paper.

2. The referral marketing model

Customer referral programs encourage existing customers to recommend a firm's services or products to their social network. They aim to provoke marketer-directed cascades of word-of-mouth (WOM). In that way, referral programs leverage on the powerful impact of WOM and the influence of social connections [17, 21]. Figure 1 illustrates how referral marketing programs create value for firms.

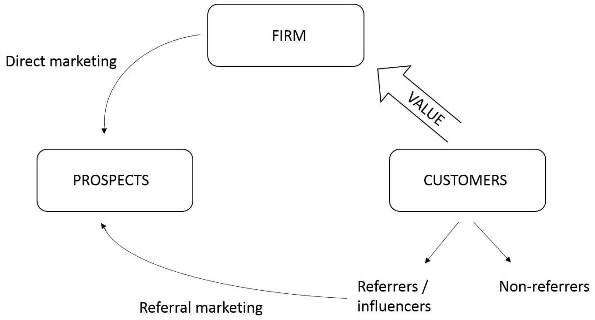

Figure 1: Referral marketing leverages the power of word-of-mouth to attract new customers

2.1. Background

The acquisition efforts used to attract a customer have an important effect on the long-term value of the customer to the firm. Villanueva et al. [37] show that customers who joined the firm as a result of WOM recommendations of social connections add almost twice as much long-term value to the firm than customers who did not join as a result of WOM. The same was found by Schmitt et al. [33]. They concluded that the difference in customer lifetime value between referred (attracted as a result of WOM recommendations) and non-referred customers is at least 16%. Kumar et al. [24] show that the most valuable customers (with high customer lifetime value) are not always those who buy most, but those whose WOM attracts the most profitable new customers. Next to a difference in value between referred and non-referred customers, there is also a difference in costs. Reichheld [30] argues that referred customers have a lower cost to serve than non-referred customers because another customer may provide help with understanding various offerings and navigating certain procedures without having to rely on the firm's customer support. Hence, the customer acquisition process has a significant impact on customer value. Many companies have already understood this and referral programs now exist in many industries such as telecommunication, retail, energy providers and restaurants [16]. Referral programs are often used by service companies since personal referrals work particularly well for experience goods like telecommunication services or gym club membership [4]. Moreover, the number of referral marketing programs is expected to increase significantly as a result of the rise in social media usage, the heightened use of customer databases by firms and the growing number of platforms to outsource referral programs [4].

2.2. The referral marketing process

As with any marketing program, a process needs to be executed in order to set up and launch a customer referral program. Berman [4] identifies an eight-step process for planning, implementing and evaluating referral programs, shown in Figure 2 on the left side.

Our study focuses on the third step "identifying a group of customers as referrers". The aim of this step is to find a group of customers that is able to influence as many other potential customers as possible through WOM and social influence. This problem is also called the influence maximization problem [22] of which the selected group of customers is called the seed set. Marketers need to have an intelligent system that supports them in solving the influence maximization problem and coming up with a list of customers that are most suited to target with a customer referral program. Figure 2 provides an overview of the three different decision support methods for seed set selection based on different data sources. Selecting the seed set for a customer referral program can be done on a random basis, by using previously proposed algorithms relying on influence simulation through the customer network or by using actual referral behaviour data.

In the first part of this study, we explore the value of using call detail record (CDR) data in the form of a communication network to select influencers for the seed set. In a first step we use the unweighted communication network for identifying the top influencers, while in a second step we use the weighted network taking into account

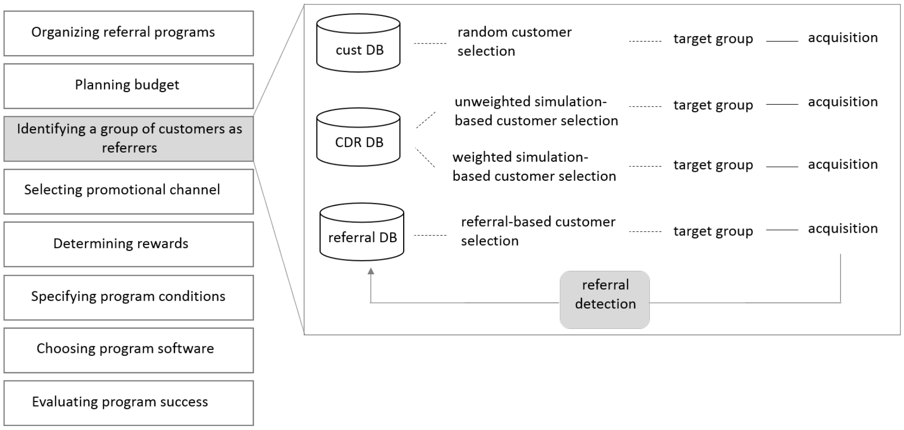
Organizing referral programs
Planning budget
Identifying a group of customers as referrers
Selecting promotional channel
Determining rewards
Specifying program conditions
Choosing program software
Evaluating program success
cust DB
random customer selection
target group
acquisition
CDR DB
unweighted simulation-based customer selection
target group
acquisition
weighted simulation-based customer selection
target group
acquisition
referral DB
referral-based customer selection
target group
acquisition
referral detection

Figure 2: Three strategies for selecting a group of customers as referrers

the connection strength between individuals. The connection strength has an impact on how likely it is that individuals influence each other: a stronger relationship implies a higher chance of recommending products to each other. We examine whether a weighted approach leads to simulations of influence flow that are a better representation of real life.

In the second part of this paper, we analyze the added value of using referral behaviour data for selecting top influencers. In order to be able to use referral data as input to the decision support system, this data needs to be captured every time a new customer is referred. As a result, this requires a referral detection process. This study examines whether implementing such process is beneficial and leads to an optimized selection of influencers thanks to the use of referral behaviour data.

In the third part of this paper, we investigate how much the selection of the seed set improves when taking into account the two-hop neighbourhood of the customers during the selection process. As Li and Shiu [25] argue, there is no use in targeting a campaign to an influential customer if his connections do not spread the message on their turn. Thus, we incorporate a measure for the influence of a customer's connections and verify whether this results in a better seed set selection.

3. Related work

There is a wide literature from sociology, psychology, economics and computer science studying social influence. In this section, we focus on influence diffusion models and the influence maximization problem. Customers and prospects influence each other through WOM and in that sense can be thought to form a network *G(V,E)* in which the individuals (customers and/or prospects) represent the nodes V and the relationships between the individuals form the edges *E*. This representation allows for graph theoretic analysis of customer activities. The influence maximization problem identifies a group of customers that leads to the largest influence spread in the social network under a given diffusion model.

3.1. Diffusion models

Diffusion models are models that simulate a diffusion process in a complex network. Multiple types of diffusion models exist that take different approaches to the features of the spreading process. In general, the object of the diffusion travels from node to node over the links in the network. Two classical diffusion models based on mathematical sociology are the linear threshold model and the independent diffusion model.

The linear threshold model (LT model) starts with an initial seed set of active customers (in our setting an active customer is a customer who has adopted the product) who have already adopted the product and are selected to start the diffusion process (all other nodes are inactive). Every node *i* in the network has its own uniformly distributed threshold value $\theta_i \in [0, 1]$. This threshold value determines how much influence from the direct neighbours is necessary for this node to also become active. Let us consider node *i* and represent its neighbours by N_i. Node *i* is influenced by its neighbour *j* according to a weight w_{ij}, reflecting the strength of the relationship. These weights are normalized for every node such that $\sum w_{ij} \leq 1$. The decision of node *i* to become active depends on the total influence of *i*'s neighbours scaled by weight. If this total weight exceeds the personal threshold θ_i, such that $\sum_{j \in N_i} w_{ij} \geq \theta_i$ then node *i* will decide to also adopt the product.

The *independent cascade model* (IC model) considers individual and independent interactions and influence among connections in a network. Initially, a set of active nodes $S \in V$ is fixed that constitutes the start of the diffusion process. Every edge in the network is assigned a probability p_{ij} illustrating the chance of node *i* successfully influencing neighbour *j*. This probability p is assigned uniformly over all edges in the network before simulating the diffusion process. Once a node is active, it will try to activate all its neighbours. However, every node only has one chance per connection to attempt to individually influence it. Whether node *i* can influence node *j* directly depends on the probability p_{ij}. In case none of *i*'s neighbours can be activated, this branch of the diffusion process is terminated. Once a node is activated it remains activated during the rest of the process. This process progresses iteratively until no more nodes can be activated.

These two diffusion models provide a way to model spreading in a network. Influence maximization methods build on these models.

3.2. The influence maximization problem

Domingos and Richardson [15] were among the first to study influence maximization in a social network as an algorithmic problem. They proposed the idea of assigning each customer a value that reflects the influence of this person on other individuals in the network. In contrast to Domingos and Richardson's algorithm that relied on probabilistic methods, Kempe et al. [22] were the first to formulate the same problem as a discrete optimization problem. They proposed a Greedy approximation algorithm to find the *k* most influential nodes in a network. The Greedy approach starts with an empty seed set *S* = *O*. On every iteration of the algorithm, the node *u* with the largest increase in the expected influence spread $\sigma(S)$ is added to the seed set *S*. When using this estimation method, the chosen seed set *S* activates at least $(1 - \frac{1}{e}) \approx 63\%$ of the nodes in the network compared to the activated nodes by any set S^* of k chosen seed nodes. Despite the fact that this problem is NP-hard under both the LT and IC model, this approximation can be reached thanks to certain characteristics of the function $\sigma(S)$ (submodularity and monotonicity, see [22] for more details).

Since Kempe et al. [22] published the Greedy algorithm, many studies have proposed other methods to solve the influence maximization problem. Leskovec et al. [25] propose a 'Lazy-Forward' optimization as an improvement to the Greedy algorithm. Their CELF algorithm eliminates the need to evaluate all nodes at every iteration thanks to the submodularity property. By ranking the nodes in order of decreasing influence, it suffices to evaluate the influence of only the top few nodes in the ranked list. A similar method for the IC model was proposed by Chen et al. [7] in which they reduce the size of the graph G by only taking into consideration those edges that have a minimum propagation probability p. Wang et al. [38] developed a heuristic that first identifies communities in a social network and then discovers influential users within these communities. They argue that influence mainly flows within communities rather than across communities and that for this reason focussing within communities is a reasonable approach. Chen et al. [10] create a local directed acyclic graph (DAG) for every node in the network and consider influence to this node only within its local DAG. A Greedy approach is then used to find the nodes with the largest marginal influence increase within their local DAG. Chen et al. [8] introduce the degree discount heuristic that accounts for the fact that potential seed nodes might have links to each other. The maximum degree heuristic implies that the seed set should be composed of the *k* nodes that have the highest degree. The degree discount heuristic then adjusts the number of links of every node by accounting for the number of links this node has to other seed nodes. This avoids a situation in which the seed set consists of nodes that have overlapping links. Goyal et al. [18] developed the credit distribution model which directly estimates influence propagation by assigning an influenceability score to all nodes based on historical data (action logs). Dasgupta et al. [14] provide evidence that social relations have a large impact on customer churn. They discuss a diffusion-based approach to identify potential churners by taking into account the influence spread and its impact on churn behaviour.

A different angle was taken by [28] who developed the SPIN algorithm based on the game-theoretic concept Shapley value to identify influencers in a social network. The SPIN algorithm models the information diffusion process as a cooperative game in which the Shapley value of the individuals in the network reflects their influence. Shapley value can be used to distribute the total influence in the network among all customers based on their individual influence. After ranking the nodes based on influential value — Shapley value — the seed set is chosen by iteratively taking the most influential node from the ranked list that is not adjacent to any node already in the seed set. They found that the SPIN algorithm is more efficient, requiring less computational resources than the previously proposed Greedy algorithm and achieves comparable influence spread.

In addition to the advanced approaches described in this section, many studies use other more simple heuristics based on node characteristics such as maximum degree, maximum betweenness centrality or maximum closeness centrality [7, 10].

Previous research on influence maximization uses networks such as co-authorship in physical publications [22, 28, 8, 7], political books sold on Amazon.com that are often bought by the same buyers [28], trust relationships in the online social network Epinions.com [7, 23], social connections in Flickr [6], Second Life friendships [2] and synthetic networks [7, 28]. These networks do not constitute actual referral networks, rather they are social networks linking items and/or people. These networks are then used to simulate spreading often based on the LT and/or the IC model that assign randomly generated adoption probabilities respectively to the nodes and the edges in a network. Hence, these networks do not incorporate explicit data on influence spread and resulting product adoption, rather this is simulated by chance based on the LT or IC diffusion models. The underlying method of these simulation-based studies is to simulate which customers in the network will become active based on the ones that already are active and their links to others. This is different to our method since we have real-life data about referrals made by existing customers and resulting product adoption. By using this data set, it is not necessary to simulate which customers become active since this is readily available in the data set.

Table 1: Main differences between simulation-based decision methods and our real referral-based method

	Simulation-based methods	Real referral-based method
Data available	Nodes, links	Nodes, links, real influence cascades
Influence cascade identification	Simulate cascades based on existence of links	Cascades readily available in data
General method	Based on active nodes determine which connections become active	Based on real activation history assign credit to active nodes

4. Methodology

As discussed in the previous section, the literature on influence maximization is highly diverse with many different methods and variations on the methods proposed by different authors. In this section we propose a general way of simulating influence spread through a network based on game theory. We use this method for selecting six different seed sets of influencers: a first based on an unweighted communication network, a second based on a weighted communication network, a third based on a real referral network for which no simulations are needed as true referral behaviour is known and three others using the same input data as in the first three but accounting for two-hop influence spread during the seed set selection.

4.1. Game-theoretic preliminaries

Every time a new customer joins the telecommunication provider, the total value of the network increases. As such, a customer who has influenced many new customers to join the network has created significant value for the telecom provider. In this regard, the influence of every node in the network can be denoted by the number of referrals that were initiated by the influence of this node. Formally modeling such a situation in which participants contribute to a shared total value can be done by using the concept of a *cooperative game*, which has its roots in game theory.

A cooperative game is defined as the pair *(N,v)* where $N = 1, 2, ...n$ is the set of players and *v* represents the value of the game by a real-valued mapping $v\colon 2^N \rightarrow \mathbb{R}$ of a set of players $S \subseteq N$ to their value *v(S)*. Note that 2^N is the set of all possible subsets of *N* and that $v(\emptyset) = 0$.

Every node $i \in N$ contributes to the overall utility of the game with a value *v(i)*. Analogous, a set of nodes $S \subseteq N$ reaches a total utility of *v(S)*, excluding any contribution of the players in $N \backslash S$. The value *v(i)* that every node $i \in N$ contributes to the total utility of the network is typically described as the marginal contribution $mc(i, N)$.

In this study, a cooperative game that captures referral behaviour in a social network is defined. The customers of the telecom provider are the players in the game and the marginal contribution of each player is defined as the number of new customers who joined the provider thanks to the influence of this individual player.

A cooperative game can be analyzed using a *solution concept*, which provides a method for distributing the total value of the game among the participants. The Shapley value is a solution concept that formulates an efficient approach to the fair allocation of the total utility *v(i)* in the game among the players [34]. Crucial here is the fairness of the allocation. Fair implies that players who contribute more to the total value should be allocated a larger fraction thereof than players who contribute less to the game. Hence, it provides a way of computing the average marginal contribution mc(i,N) of each player *i*. The Shapley value of the cooperative game *(N,v)* is denoted by

$$\phi(N,v) = \phi_1(N,v), \phi_2(N,v), ..., \phi_n(N,v) \tag{1}$$

The Shapley value $\phi_i(N,v)$ of player $i \in N$ is given by

$$\begin{aligned} \phi_i(N,v) &= \sum_{S \subseteq N \setminus i} \frac{|S|!(n-S-1)!}{n!}(v(S \cup i) - v(S)) \\ &= \sum_{S \subseteq N \setminus i} \frac{|S|!(n-S-1)!}{n!} mc(i,S) \end{aligned} \tag{2}$$

The communication and referral network can be analyzed as a cooperative game, using the Shapley value to indicate the marginal contribution of every individual player to the overall value of the game. Every node's Shapley value then signalizes the number of referrals triggered by the influence of this customer. From this it follows that these values can be used to identify the most influential customers in the network.

4.2. The Estimated Referral Rank for finding top influencers

4.2.1. An unweighted approach

Previously proposed methods for finding the top influencers in a network use simulations to mimic influence flow in the network. This is necessary because of the lack of data on actual influence flow. The IC or LT model are used to simulate product adoption when nodes in the network have been subject to a minimum level of influence from social connections. As explained in Section 3.1, these methods start by randomly initializing the group of first adopters and randomly assigning an influence threshold to every node or link (in respectively the LT and IC model) that determines how much influence is needed for this node to also adopt the product or how much influence flows to this node.

Contrary to previous research, the unique data set used in this study also contains data on the product adoption timing of the customers in the network. These dates of subscription are used to replicate the adoption sequence. The communication network can then be leveraged for identifying the top influencers using simulations that are similar to previous research. To do so the communication links are used as a proxy for influence flows. When a new customer *j* joins the telecom operator, it is unknown which existing customer *i* has influenced this person. Therefore, we are not able to assign the value *v(j)* of this new customer to the marginal contribution $mc(i, N)$ of the customer who influenced this person to join. We can only make use of the existing communication links in the network and assume that a new customer who joins the operator has been influenced by one of his existing customer connections. Consequently, we need to simulate this by distributing the value of this new customer *j* over all his connections that are existing customers. The resulting measure is called Estimated Referral Rank as it facilitates constructing a ranking of all nodes in the network based on their estimated potential of referring new customers. Customer *i*'s marginal contribution, or Estimated Referral Rank (ERR), to the network is defined as

$$ERR_i = \sum_{j \in N_1(i)} \frac{\phi_j(N, v)}{n_j} \tag{3}$$

in which $\phi_j(N, v)$ is the Shapley value of new customer *j*, which is distributed over n connections of customer *j* that are existing customers.

Figure 3 presents an example communication network. Let's say that node *a* is a new customer who recently joined the network and communicates with customers *b*, *c*, *d* and *e*. As we only have communication data, we do not explicitly know which existing customer influenced *a* to join. Therefore, we assign to each connection of *a* an equal portion of the value of *a*, which in this case is 0.25. In this small example, customers *b*, *c*, *d* and *e* are equally influential with an ERR of 0.25.

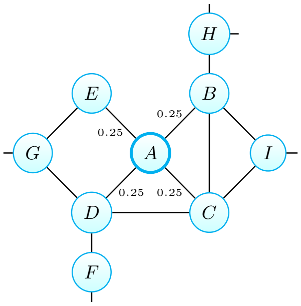

Figure 3: Unweighted example communication network

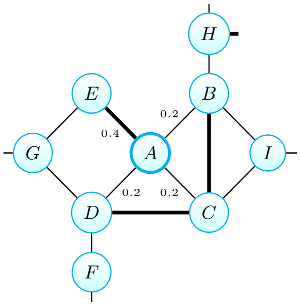

Figure 4: Weighted example communication network

4.2.2. A weighted approach

In the ERR, the weight of an edge between nodes *i* and *j* is one if there is an edge and zero if there is none. However, when insights on the edge strengths are available, this can be incorporated in the method for selecting top influencers to render the model more realistic. Connection strength can be used to weight the allocation of the value of a new customer over the existing customer connections. Instead of equally distributing the value, it can be done proportionally to the connection strength. If new customer *j* has a stronger connection with existing customer *i* than with existing customer *k*, it is likely that *i* has a larger influence on *j* because they communicate more often which results in more social influence. As a result, existing customer *i* should be assigned a larger portion of the value of new customer *j* than existing customer *k*. Thus, customer *i*'s Weighted Estimated Referral Rank (WERR) can be computed as

$$WERR_i = \sum_{j \in N_1(i)} \phi_j(N, v) \cdot \frac{a_{ij}}{a_{.j}} \tag{4}$$

where a_{ij} is the weight of the edge between node *i* and *j* defined by weight matrix a and $a_{.j}$ is the sum of the weights of all connections of node *j*. We define the weight of an edge between two customers in our communication network based on the total duration of phone calls and the number of single SMS's (1 SMS can consist of multiple SMS's) sent between these two customers. Figure 4 provides an example of how the distribution of the value of node a over the neighbours is established based on the weights of the connections. Based on the communication frequency and duration between the existing customers with new customer a, we can see that existing customer c has a stronger connection with new customer a. This implies that it is more likely that customer c has influenced the new customer to join the network and therefore should be assigned more value than the other existing customers.

4.3. The Real Referral Rank for finding top influencers

Although it is not that common yet, management can also gather information on the referrals made by their customers by for example tracking this online. When such information is available, managers could use data on the actual referral behaviour of their customers when selecting the top influencers. In this section we define the Real Referral Rank which uses actual referral behaviour for quantifying individuals' influential value. A customer's Shapley value based on his actual referrals reflects this customer's true contribution to the network, defined as the Real Referral Rank (RRR). A limitation of the data is that every new customer can only be referred by strictly one existing customer. As a result thereof, for every new customer *j* the number of neighbours n_j who are existing customers is equal to 1. This also implies that there is no need to account for edge strength. A customer's RRR is denoted by

$$RRR_i = \sum_{j \in N_1(i)} \phi_j(N, v) \tag{5}$$

which reflects the total value of the new customers referred by existing customer *i*.

4.4. Extending the influence area: Two-hop selection

It is important to realize that influence does not suddenly vanish after the one-hop neighbours of a node. According to Christakis and Fowler [11], noticeable interpersonal influence propagates as far as the two-hop neighbourhood of the influencing

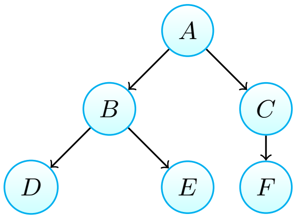

Figure 5: Example referral network

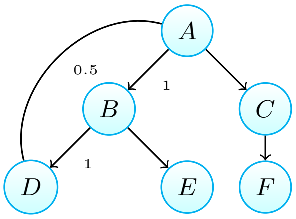

Figure 6: Example referral network with a transformed two-hop edge

node. Moreover, Li and Shiu [26] also compute influence propagation in a recursive way in their diffusion model. They argue that there is no use in targeting a campaign to an influential customer if his connections do not spread the message on their turn. Thus, the influence of node a in Figure 5 propagates as far as the second hop neighbourhood which is the level of nodes *d*, e and *f*. Customer a has thus referred new customers who are also relatively influential and can further propagate a's influence. For that reason, the marginal contribution of customer a should take into account the marginal contributions of customers *b* and *c*. This seems reasonable as nodes that have a large influence cascade contribute more to the total value of the network than nodes that induce only a short influence cascade.

As the influence of the neighbours of the seed set has a significant impact on the resulting influence spread, we investigate whether the selection of top influencers can be improved by incorporating a measure of second hop influence. In order to do that, we transform two-hop neighbours to one-hop neighbours by using a formula proposed by Verbeke et al. [35]. They propose a novel way to reduce a two-hop link into a one-hop link by computing a weight for the two-hop link that is based on the weights of the one-hop links. They define the weight matrix λ for transforming two-hop edges to one-hop edges as

$$\lambda_{ik} = \max_j \frac{a_{ij} \cdot b_{jk}}{a_{ij} + b_{jk}} \tag{6}$$

in which λ_{ik} is the new weight between the node *i* and its second hop neighbour node *k* and in which a_{ij} and b_{jk} represent the weights between nodes *i* and *j* and nodes *j* and *k* respectively.

This is illustrated in Figure 6, where the two-hop connection between node *a* and *d* is created based on the one-hop connections between nodes *a* and *b* and nodes *b* and *d*.

When extending the influence area considered from one-hop to two-hops, the ERR and RRR of customer *i* become

$$ERR_i^2 = \sum_{j \in N_1(i)} \frac{\phi_j(N, v)}{n_j + \lambda_{.j}} + \sum_{k \in N_2(i)} \frac{\phi_k(N, v) \cdot \lambda_{ik}}{n_k + \lambda_{.k}} \tag{7}$$

and

$$RRR_i^2 = \sum_{j \in N_1(i)} \frac{\phi_j(N, v)}{\lambda_{.j}} + \sum_{k \in N_2(i)} \frac{\phi_k(N, v) \cdot \lambda_{ik}}{1 + \lambda_{.k}} \tag{8}$$

where λ is the factor that transforms the two-hop neighbours to one-hop neighbours, as described in Equation 6, and the existing customer *i* has new customer *j* as direct connection who in turn has new customer *k* as neighbour. $N_1(i)$ and $N_2(i)$ respectively represent the one-hop and two-hop neighbourhoods of existing customer *i*. The first terms give the total direct contribution of customer *i* received because of *i*'s one-hop neighbours who are new customers. The second terms represent the total indirect contribution of customer *i* because of the influence of *i*'s direct neighbours resulting in *i*'s two-hop neighbours who are new customers. Note that λ will always be 0.5 for the ERR and RRR as edge strengths are either one or zero.

It will be different for the WERR as the edges can have any weight $\in [0.1]$. Extending the WERR to the second hop neighbourhood is done as follows

$$WERR_i^2 = \sum_{j \in N_1(i)} \phi_j(N, v) \cdot \frac{a_{ij}}{a_{.j} + \lambda_{.j}} + \sum_{k \in N_2(i)} \phi_k(N, v) \cdot \frac{\lambda_{ik}}{a_{.k} + \lambda_{.k}} \tag{9}$$

Here too, λ is the factor from Equation 6 to convert two-hop links to one-hop links. The difference with the ERR formula is that the assigned value is scaled for the weight of the connections by weight matrix *a*.

5. Results

5.1. Data

For this study, we use a call detail records (CDR) data set of a European telecom provider, consisting of 1,6 billion communication records of 211,075 customers. The telecom operator continually runs the same referral program using the same marketing message and incentives. As a result, the impact of the characteristics of the referral program on the referral behaviour of the customers is limited, which benefits the generalisability of the findings. The communication behaviour of the customers defines a network in which the individuals are the nodes and the communication interactions are the links. In addition to data on the communication behaviour of the customers, also data on the referral behaviour of the same group of customers is available. The total number of recorded referrals is 65,078. The ID of the referring customer is recorded together with the assigned ID of the new customer, as well as the method of referral and the date and time and whether the referral was successful or not. Using referral behaviour, we can construct a referral network in which the individuals are the nodes and the referral-referred customer relations are the links. The communication network used in this study was selected as follows. First, we selected all existing customers who ever communicated with a new customer. A new customer is defined as a customer who joined the telecom operator in the 6 months before the evaluation period. There are 57,551 new customers and 478,109,344 records of communication. Second, we defined a minimum communication threshold of a total of 10 minutes of communication over 6 months to ensure that only intended calls representing a true social connection are considered. As Ma et al. [27] note in their study, setting such threshold is subjective and can only be based on exploration of the data. It is important that the threshold is not too low because this will lead to including individuals who are not part of the caller's social network, nor too high since it will reduce the power of the analyses [27]. Our selection of 10 minutes of communication aims to achieve a balance between the two.

Exploration of the data showed that results based on different cut-off levels delivered similar results. Applying this threshold results in 130,537,963 CDR records. A total of 65,078 referrals were made by 48,798 customers of this group of 57,551 customers. This means that the communication network has 57,551 nodes and 156,020 edges and the referral network has 48,798 nodes and 65,078 edges.

5.2. The evaluation procedure

In order to be able to evaluate the increase in the number of customers, the data set was separated into two distinct sets. One year of mobile communication and referral data was split in six months training set and six months test set. The training set is used for selecting the seed set of influential individuals and the test set is used for generating the resulting product adoption in the network.

Many organisations currently do not hold data about the referral behaviour of their customers. Communication behaviour on the other hand is more common to have access to. That is why in a first part of the experiments, we only use the communication network for selecting influencers and evaluating the resulting product adoption as companies that have only this data available would do. This is done by simulating influence propagation based on the communication edges in the network (as is traditionally done in influence maximization literature). We use the same ERR and WERR methods on the test data set for evaluating the influence spread of the customers as we did on the training data set for assigning the customers an influencer score. This is referred to as the simulation-based methods in the remainder of the paper. In order to determine the value for companies to gather referral behaviour data, in a second part of the experiments we evaluate the real referral behaviour of the customers based on the referral network. This implies that the influence spread of the customers is determined by the actual number of referrals made by every customer.

To statistically test the results, we analyze the results using Friedmans Chi-square test, which is the non-parametric equivalent of the repeated-measures ANOVA. The results indicate that all differences are statistically significant (p-value < 0.001). As the Friedman test indicates significance, the Nemenyi post-hoc test is employed to analyze the differences between the methods [13]. The results show that the differences between all pairs of methods are significant (p < 0.001).

5.3. The value of real referral data for seed set selection

In the first step of the simulations we compare the simulation-based methods, both the unweighted and the weighted approach (the ERR and WERR respectively), for selecting influencers with the referral data-based method. It is important to realize that in both the training period — when selecting the top influencers — as well as in the test period — when evaluating the resulting product adoption — simulated influence spread and real referral data can be used.

5.3.1. Evaluating on simulation-based influence spread

First, we use the communication network to simulate the influence spread and resulting product adoption in the test period. Figure 7 and Table 2 illustrate that the RRR method outperforms the ERR method. This implies that even when evaluation is done based on simulated influence spread, using referral behaviour data for selecting the top influencers leads to larger influence spread. Thus, selecting influencers based on referral behaviour data results in more product adoption through a larger influence spread than selecting influencers based on simulations. The lower-bound on performance is a random approach. In order to ensure unbiased results, the process of picking random nodes and simulating network growth is performed 100 times. Per number of k seed nodes, the average of these 100 observation then represents the random spread.

In the second step of the simulations we use the edge strengths in the communication network to compute the WERR. Again, we use the communication network to simulate the influence spread and resulting product adoption in the test period. The results, as visualized in Figure 9 and Table 4, demonstrate that the difference between the ERR and WERR method is rather small. If the manager would be restricted, because of a lack of real referral data, to evaluate the results on simulated data, only little differences between ERR and WERR would be observed.

5.3.2. Evaluating on real referral behaviour

Second, we use the real referral network to evaluate the resulting product adoption in the test period. In Figure 8, the difference in performance between the ERR and RRR method is very large. The RRR method for selecting influencers performs very well when evaluation is done on real referral behaviour data. It is clear that in case referral data is available it should definitely be utilized when searching for influencers in a customer base. Further, comparing Table 2 and 3 indicates that evaluating on simulated data underestimates the influence spread when selecting influencers using the RRR method, but overestimates it when selecting influencers using the ERR method.

The fact that the RRR method performs better than the ERR method signalizes that better results are attained when the data approximate real life behaviour. However, most organisations do not have access to referral behaviour data of their customers, so we propose to use the WERR to render the ERR method a better representation of real life. Hence, we use the edge strengths in the communication network to compute the WERR. Figure 10 shows that the WERR outperforms the ERR, which implies that the weighted approach in fact outperforms the unweighted approach. This was not visible when evaluating on simulations, visualized in Figure 9. As a result, this method performs better when evaluation is based on the referral data. This implies that incorporating information on edge strength optimizes the identification of influential customers, especially when a large seed set is selected.

An important aspect of the results is the over- and underestimation of these methods. By comparing Table 4 and Table 5, it can be stated that the influence spread of the influential customers selected with the ERR evaluated on real referral data leads to a lower number of activated new customers than evaluating on simulated data. This

Table 2: Nr of new customers evaluated using simulation after selecting the seed set based on the two different methods

Nr of seed nodes	ERR	RRR
50	15	14
100	28	43
300	84	118
500	130	181
1000	265	307

Table 3: Nr of new customers evaluated using real referral data after selecting the seed set based on the two different methods

Nr of seed nodes	ERR	RRR
50	14	165
100	24	235
300	69	363
500	115	445
1000	224	631

Table 4: Nr of new customers evaluated using simulation after selecting the seed set based on the two different methods

Nr of seed nodes	ERR	WERR
50	15	17
100	28	27
300	84	84
500	130	133
1000	265	265

Table 5: Nr of new customers evaluated using real referral data after selecting the seed set based on the two different methods

Nr of seed nodes	ERR	WERR
50	14	15
100	24	25
300	69	82
500	115	117
1000	224	244

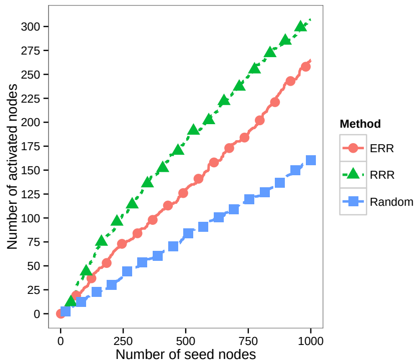

Figure 7: The RRR method outperforms the ERR method when evaluating on simulated data

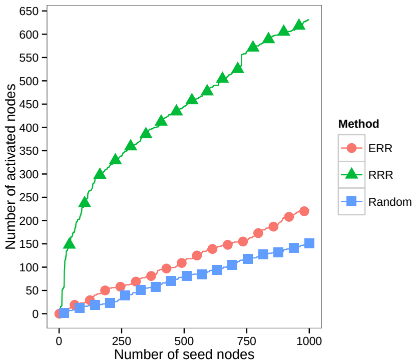

Figure 8: The RRR method far outperforms the ERR method when evaluated on real referral data

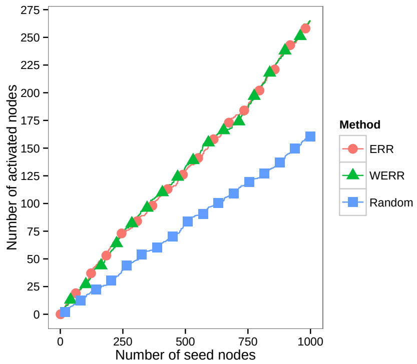

Figure 9: The WERR and ERR method perform similarly when evaluated on simulated data

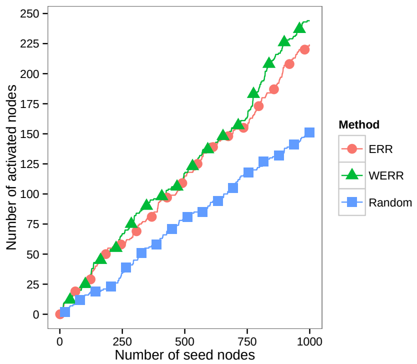

Figure 10: The WERR method outperforms the ERR method when evaluated on real referral data

implies that the product adoption realized by influencers selected based on the links in the network might give an overestimated view of the product adoption in reality.

In conclusion, we find that the RRR method leads to the best selection of influencers when both evaluating on simulated influence spread and on referral behaviour. It is thus beneficial to invest in a referral detection process. If no referral behaviour data is available as input to the decision support system for finding influencers, the WERR method should be preferred over the ERR method as it performs slightly better in identifying the top influencers.

5.4. The value of two-hop selection

So far the experiments only incorporated the one-hop neighbourhood influence spread of every node. In the following, we describe the results of the experiments when taking into account the influence of the two-hop neighbours of the node considered.

5.4.1. Evaluating on simulation-based influence spread

First, we examine the results when evaluating the product adoption in the second period by simulating the influence spread over the communication network links. We again include a lower bound based on random selection. The experiments show that, for all three methods, considering the two-hop neighbourhood of the customers results in a better selection of influential customers and higher influence spread. This is shown when comparing Table 2 and Table 4 with Table 6. It can be seen from Figure 11, Figure 12 and Figure 13 that the difference even increases as the seed set grows. Consequently, for large seed sets it pays off for managers to take the effort of considering the influence of the customers' two-hop neighbourhood connections.

5.4.2. Evaluating on real referral behaviour

Second, we investigate the results when evaluating the product adoption in the evaluation period by looking at real referral behaviour. The results show that, for all three methods, the performance is better when considering the two-hop neighbourhood. This can be seen by comparing Table 3 and Table 5 with Table 7. However, the improvement is smaller than when evaluation is done based on simulations. Thus, the previous evaluation based on simulated influence spread overestimates the actual product adoption in the test period. Figure 14, Figure 15 and Figure 16 demonstrate that taking into account the influence of the two-hop neighbours indeed results in a larger influence spread and more product adoption.

In conclusion, we can state that the difference in performance between the one-hop and two-hop methods is smaller when evaluating on real referral data than when

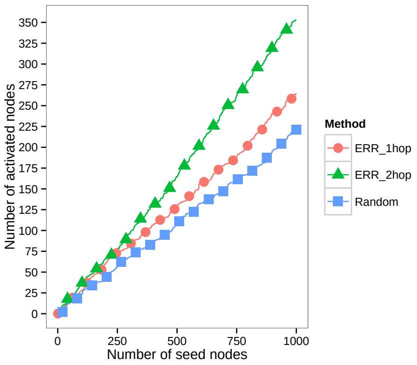

Figure 11: The ERR two-hop method evaluated on simulated influence spread performs better than the one-hop method

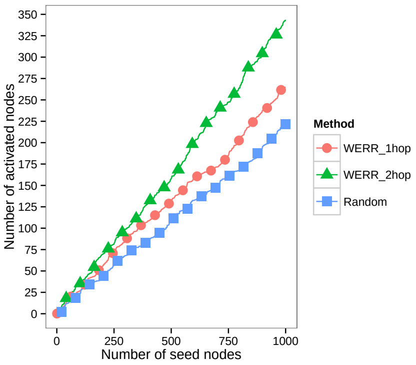

Figure 12: The WERR two-hop method evaluated on simulated influence spread performs better than the one-hop method

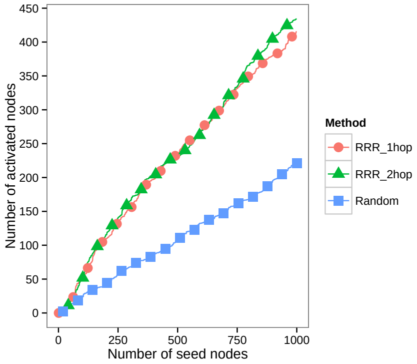

Figure 13: The RRR two-hop method evaluated on simulated influence spread performs better than the one-hop method

evaluating on simulated influence spread. This indicates that evaluating on simulated influence leads to an overestimation of the improvement in influence spread when taking into account the two-hop neighbourhood influence rather than just one-hop. The results evaluated on real referral data show that there is indeed an improvement when using the two-hop method instead of the one-hop method, but this improvement is rather limited.

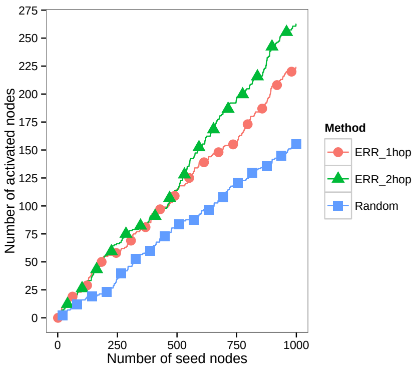

Figure 14: The ERR two-hop method evaluated on real referral data performs slightly better than the one-hop method

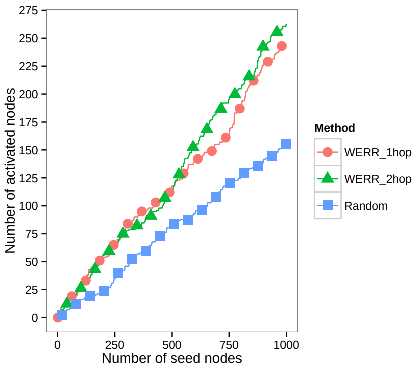

Figure 15: The WERR two-hop method evaluated on real referral data performs better than the one-hop method for large seed sets

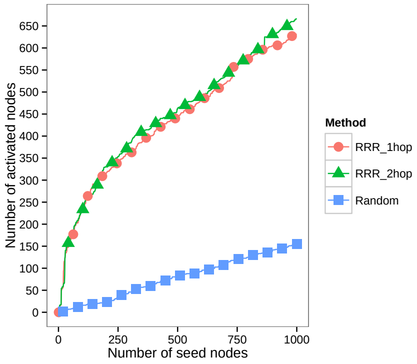

Figure 16: The RRR two-hop method evaluated on real referral data performs slightly better than the one-hop method

6. Conclusion and discussion

This study investigates an issue critical to the success of referral marketing programs: how can a group of customers be identified who are most influential and can affect the largest number of potential customers through word-of-mouth. Previous research is generally design- and technology-oriented and use simulation-based methods to simulate influence spread over networks. Using a unique data set composed of both communication data and referral behaviour data, this study investigates whether the

Table 6: Nr of new customers evaluated using simulated data after selecting the seed set based on the three different 2-hop methods

Nr of seed nodes	ERR	WERR	RRR
50	19	22	16
100	37	35	51
300	96	99	165
500	165	158	208
1000	353	344	387

Table 7: Nr of new customers evaluated using real referral data after selecting the seed set based on the three different 2-hop methods

Nr of seed nodes	ERR	WERR	RRR
50	14	14	173
100	26	26	230
300	78	78	384
500	114	114	463
1000	263	263	666

algorithms based on influence propagation simulations perform well in terms of identifying the most influential individuals in the network and estimating their resulting influence spread. The results show that limiting the decision support method for finding top influencers to simulations leads to overestimations of the actual influence spread and resulting product adoption. The best results are attained when referral data is used for selecting top influencers. Unfortunately, it is not that common yet for organisations to capture data about their customers' referral behaviour. In that case, a measure of tie strength between individuals should be incorporated in the selection method as this leads to a larger influence spread. Next to that, the results also prove that it is important to not just look at the influence of the targeted customers, but also at the influence of their connections. If the connections of the most influential customers are not willing to spread word-of-mouth, there is no use in targeting these customers with a marketing campaign since the influence will not spread very far. Overall, this study shows the value of a referral behaviour detection process. A decision support system for selecting the most influential customers based on referral data allows companies to identify their most influential customers of whom the influence spread will trigger the largest cascade in product adoption. Fortunately this kind of data is becoming easier to obtain thanks to the widespread use of social media. Hence, also other organisations that possess any kind of data related to referrals or recommendations and wanting to reach a large audience can benefit from the approach suggested in this paper. In case no referral behaviour data or proxy data thereof is available, the simulation methods based on network data are already valuable and succeed in identifying influencers in a social network, although less so than those based on referral data.

7. Limitations and future research directions

The referral data set used in this study has a one-to-one relation between the referred customer and the referrer. This is because it is inherent to the business model of the telecom provider. In reality however, multiple customers can have influenced a new customer to adopt a product. This subtlety is not visible from the data set but could potentially improve the model of the influence flow.

8. Acknowledgements

The authors gratefully acknowledge financial support from the Flemish Research Council (for the FWO-ICM Fellowship of Iris Roelens) and the National Bank of Belgium (NBB).

9. References

[1] (). Referral program examples an epic list of 47 referral programs. http://www.referralcandy.com/blog/47-referral-programs/. Accessed: 2016-04-13.

[2] Bakshy, E., Karrer, B., & Adamic, L. (2006). Patterns of influence in a recommendation network. In *Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining* (pp. 380–389).

[3] Bao, T., & Chang, T. (2014). Finding disseminators via electronic word of mouth message for effective marketing communications. *Decision Support Systems*, *67*, 21–29.

[4] Berman, B. (2016). Referral marketing: Harnessing the power of your customers. *Business Horizons*, *59*, 19–28.

[5] Van den Bulte, C., Bayer, E., Skiera, B., & Schmitt, P. (). How customer referral programs turn social capital into economic capital. *Working paper*, .

[6] Cha, M., Mislove, A., & Gummadi, K. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 721–730).

[7] Chen, W., Wang, C., & Y.Wang (2010). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1029–1038).

[8] Chen, W., Wang, Y., & Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 199–208).

[9] Chen, W., Yuan, Y., & Zhang, L. (2010). Scalable influence maximization in social networks under the linear threshold model. In *IEEE International Conference on Data Mining* (pp. 88–97).

[10] Chen, W., Yuan, Y., & Zhang, L. (2010). Scalable influence maximization in social networks under the linear threshold model. In *IEEE International Conference on Data Mining* (pp. 88–97).

[11] Christakis, N., & Fowler, J. (2013). Social contagion theory: examining dynamic social networks and human behaviour. *Statistics in Medicine*, *32*, 556–577.

[12] Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, K., Loreto, V., Moat, S., Nadal, J.-P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., & Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal*, *214*, 325–346.

[13] Coussement, K., Benoit, D. F., & Antioco, M. (2015). A bayesian approach for incorporating expert opinions into decision support systems: a case study of online consumer-satisfaction detection. *Decision Support Systems*, *79*, 24–32.

[14] 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. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology* (pp. 668–677).

[15] Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 57–66).

[16] Garnefeld, I., Eggert, A., Helm, S., & Tax, S. (2013). Growing existing customers' revenue streams through customer referral programs. *Management Science*, *77*, 17–32.

[17] Godes, G., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*, 545–560.

[18] Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2011). A data-based approach to social influence maximization. In *Proceedings of the VLDB Endowment* (pp. 73–84).

[19] Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, *21*, 256–276.

[20] Iyengar, R., Van den Bulte, C., & Valente, T. (2013). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, *30*, 195–212.

[21] Keller, E. (2007). Unleashing the power of word of mouth: creating brand advocacy to drive growth. *Journal of Advertising Research*, *47*, 448.

[22] Kempe, D., Kleinbert, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 134–146).

[23] Kim, J., Kim, S.-K., & Yu, H. (2013). Scalable and parallelizable processing of influence maximization for large-scale social networks. In *IEEE 29th International Conference on Data Engineering* (pp. 266–277).

[24] Kumar, V., Petersen, J. A., & Leone, R. P. (2010). Driving profitability by encouraging customer referrals: who, when, and how. *Journal of Marketing*, *74*, 1–17.

[25] Leskovec, J., Krause, A., Guestrin, C., & Faloutsos, C. (2007). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 420–429).

[26] Li, Y.-M., & Shiu., Y.-L. (2012). A diffusion mechanism for social advertising over microblogs. *Decision Support Systems*, *54*, 9–22.

[27] Ma, L., Krishnan, R., & Montgomery, A. (2014). Latent homophily or social influence? an empirical analysis of purchase within a social network. *Management Science*, *61*, 454–473.

[28] Narayanam, R., & Narahari, Y. (2010). A shapley value based approach to discover influential nodes in social networks. *Automation Science and Engineering*, *8*, 1–18.

[29] Probst, F., Grosswiele, D. K. L., & Pfleger, D. K. R. (2013). Who will lead and who will follow: Identifying influential users in online social networks. *Business and Information Systems Engineering*, *5*, 179–193.

[30] Reichheld, F. (2006). *The ultimate question: Driving good profits and true growth*. Harvard Business School Press.

[31] Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems*, *55*, 863–870.

[32] Sadovykh, V., Sundaram, D., & Piramuthu, S. (2015). Do online social networks support decision-making? *Decision Support Systems*, *70*, 15–30.

[33] Schmitt, P., Skiera, B., & Van den Bulte, C. (2011). Referral programs and customer value. *Journal of marketing*, *75*, 46–59.

[34] Shapley, L. (1953). A value for n-person games. contributions to the theory of games volume ii. (pp. 307–317).

[35] Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing*, *14*, 431–446.

[36] Verbraken, T., Goethals, F., Verbeke, W., & Baesens, B. (2014). Predicting online channel acceptance with social network data. *Decision Support Systems*, *63*, 104–114.

[37] Villanueva, J., Yoo, S., & Hassens, D. (2008). The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. *Journal of Marketing Research*, *45*, 48–59.

[38] Wang, Y., Cong, G., Song, G., & Xie, K. (2010). Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1039–1048).