Submitted to *Decision Support Systems*
manuscript DSS

Identification of Influencers - Measuring Influence in Customer Networks

Christine Kiss, Martin Bichler
Internet-based Information Systems, Dept. of Informatics, TU München, Germany, kissc@in.tum.de, bichler@in.tum.de

Viral marketing refers to marketing techniques that use social networks to produce increases in brand awareness through self-replicating viral diffusion of messages, analogous to the spread of pathological and computer viruses. The idea has successfully been used by marketers to reach a large number of customers rapidly. If data about the customer network is available, centrality measures provide a structural measure that can be used in decision support systems to select influencers and spread viral marketing campaigns in a customer network. Usage stimulation and churn management are examples of DSS applications, where centrality of customers does play a role. The literature on network theory describes a large number of such centrality measures. A critical question is which of these measures is best to select an initial set of customers for a marketing campaign, in order to achieve a maximum dissemination of messages. In this paper, we present the results of computational experiments based on call data from a telecom company to compare different centrality measures for the diffusion of marketing messages. We found a significant lift when using central customers in message diffusion, but also found differences in the various centrality measures depending on the underlying network topology and diffusion process. The simple out-degree centrality performed well in all treatments.

Key words: customer relationship management, viral marketing, centrality, network theory, word of mouth marketing

1. Motivation

Due to the wealth of data that is available in today's marketing departments data mining and econometrics, in particular classification techniques, have gained considerable importance for tasks such as churn prediction or campaign management to raise product or brand awareness [6]. These techniques are typically based on information about individual characteristics of customers, where decision tree learners or logistic regressions are used to estimate the probability of a person to respond to a campaign, to buy a product, or to switch to another provider [54, 38, 12, 37].

1.1. Customer Networks and Word-of-Mouth Marketing

Beyond individual customer characteristics, the structure of customer networks has attracted much attention in the marketing literature. Repeatedly, research has shown the importance of consumer Word-of-Mouth (WOM) communication in the formation of attitudes [15], in a purchasing decision-making context [7], and in the reduction of risk associated with consumer buying decisions [49, 31]. WOM is an informal communication behavior about the experiences with specific services, products or the characteristics of the providers that the consumers exchange among each other [60].

The distribution of the WOM message varies with the degree of satisfaction and dissatisfaction of the consumers [5, 48]. One of the few empirical analyses of this effect shows that customer satisfaction has a positive impact on word-of-mouth, which in turn has a positive impact on sales and market share [24]. Another empirical investigation of word-of-mouth [5] confirms popular expectations that dissatisfaction produces more negative word-of-mouth than satisfaction produces positive word-of-mouth.

One of the challenges in measuring word of mouth is that it is difficult to observe what is usually the content of private conversations. Therefore, much of this literature is based on surveys, but not on direct observations (see for example [17]). New online communities allow for direct

observation. Through monitoring online conversations, Godes et al. [32] demonstrate how word of mouth can be measured. In addition, they show a relationship between the overall dispersion of online conversations across online communities and the popularity of television shows. Recently, Van den Bulte et al. [56] have described a formal model of influentials in a network.

Since WOM has such an enormous impact on customer opinions and buying decisions, marketing departments try to focus more and more on influential customers. *"To succeed today, you need to connect with people who are at the center of the conversation ... Specifically, you should make sure you are reaching the decision makers who are influential in others' decisions. Influentials are well connected, they have ties to a significantly larger number of groups than the average American."* [39]

In order to promote and manage WOM communications, marketers use for example *viral marketing* methods to achieve the desired behavioral response. Viral marketing refers to marketing techniques that use social networks to produce increases in brand awareness by "viral" diffusion processes, analogous to the spread of pathological and computer viruses. It can be very useful in reaching a large number of people rapidly. The assumption is that if a campaign reaches a "susceptible" user, that user will become "infected" and can then go on to infect other susceptible users [41, 46]. Arguably, these forms of marketing work best when centered on influencers: influencers supply the authority that allows a message to be conveyed quickly and reliably through WOM techniques [41].

1.2. DSS Applications

The literature on WOM marketing describes multiple attributes of influencers. For example, Keller and Berry [39] argue that influencers have multiple interests, they tend to be early adopters in markets, they are trusted by others, and have a large social network. In this paper, we focus on the latter, in particular on centrality metrics that identify, how well different customers can serve as influencers in their social network.

Such metrics of customer centrality can be useful in a variety of DSS applications, including message spreading in viral marketing campaigns to raise brand or product awareness, usage stimulation, or churn management. Obviously the position of a customer in the customer network does have an important impact on their ability to spread marketing messages. As already mentioned, classification techniques are typically used in DSS to predict a customer's willingness to respond to or forward a message. These predictions ignore a customers capability of forwarding based on the topology of his customer network. Classification models will provide a list of customers with a high likelyhood of forwarding or responding to a message. In addition, a centrality measure can be used to select those customers, who not only have a high likelihood of forwarding a message, but whose position in the network allows them to reach a large number of other customers. We will focus on this type of decision support for viral marketing campaigns in this paper, which aim at increasing awareness of a product or service. Examples might be new broadband services or the possibility to switch from paper-based to online bills.

Usage stimulation is simililar, in the sense that the network operator actively stimulates usage of particular services such as the short message service (SMS) or the multimedia message service (MMS). Also in these cases the main barrier to usage is mostly lack of awareness and familiarity with a new service, while price is often less of an issue. Therefore, operators regularly approach customers with templates for SMS or MMS messages to be sent to friends and acquaintances, in order to stimulate usage and make customers familiar with a new service. Also here, scoring models predicting the likelihood of using such a service can be combined with the centrality measure to select customers to be included in a campaign.

Finally, centrality can be important in churn management applications, a topic of significant importance to service providers. A typical task in churn management is to identify those who are

likely to churn (often also using classification techniques) and then act on those who have a high "customer lifetime value" (CLTV). On the other hand, operators don't want to lose a customer, who interacts with many people, even if his CLTV is low, because he might infect others in his community.

1.3. Network Analysis

Applications of this sort require knowledge about the social network of customers and their interactions, which are typically difficult and expensive to elicit. Nowadays, more and more data is available on the Internet about customer networks and customer recommendations. For example, Leskovec et al. [46] have recently analyzed data from a person-to-person recommendation network. Besides these new online data sources, the telecommunication industry accumulates huge amounts of data about customer interactions in the form of call data records. Although information about the content of a communication (e.g., whether there was an explicit recommendation or not) is typically not available, the frequency of interactions and the resulting network of customer interactions contain valuable information for a marketer and can be used as an estimator for the influential power of a person. In this paper, we focus on structural metrics about the influential power of customers that can be estimated based on data about customer interactions, as they are available with telecommunication providers.

Network theory concerns itself with the study of graphs as a representation of relations between discrete objects. Within network theory, there are various measures of the centrality of a vertex within a graph that determine the relative importance of a vertex, for example, how important a person is within a social network. In other words, the centrality of a node in a network is a measure of the structural importance of the node. A central customer, presumably, has a stronger influence on other network members. Multiple centrality measures have been defined in the literature, which are used for very different purposes (e.g., layout of local area networks, web search). Some of these measures exhibit high computational complexity, while others are simple to calculate. Recent literature on network analysis has shown that the centrality measures need to be matched to the network flow and application for which they are appropriate [16].

1.4. Focus of the Paper

Our main question is: "Given a customer network: How do we select the most important influencers and to what extent can messages be disseminated in the network via these highly influencing nodes." In other words, we analyze how well individuals distribute messages based on the topology of their network. This will also be referred to as the "performance" of a centrality measure.

In order to answer this question, we conduct large numbers of experiments based on field data of the customer network from a telephone operator. We measure how the selection of an initial set of customers influences the reach-out to other members of the network, measured by the number of customers reached after different models of message diffusion. We found that if the set of initial customers (represented as nodes in the network) is chosen according to SenderRank or out-degree measures, the number of reached customers is substantially higher than selecting the initial set of customers according to other centrality measures across different treatments. Thus, it is these measures that the designer of a marketing campaign should use in order to achieve a wide dissemination of marketing messages. Note that centrality is orthogonal to discrete choice models of customer preferences as they are typically used in marketing. In our analysis, we assume no significant differences in the customer preferences for particular messages. Typically, both measures would be used in combination in order to select appropriate target customers for a marketing campaign. For example, centrality can be used as a metric to rank-order customers that have a high probability of being interested in a message.

Therefore, the article is structured as follows. In section 2, we introduce related theory and empirical results from the field of network theory. This includes structural measures of influence

and diffusion theory in a network. In section 3 we describe the computational experiments to benchmark different centrality measures, and in section 4, we draw conclusions and provide an outlook on future research.

2. Network Theory

Traditionally, the study of complex networks has been the field of graph theory. While graph theory initially focused on regular graphs, since the 1950's large scale networks with no apparent design principles were described as *random graphs*, proposed as the simplest and most straightforward realization of a complex network [3]. Erdös and Rényi define a random graph as n labelled nodes $i \in N$ with $i = 1, .., n$ connected by a set of E edges which are chosen randomly from the $\frac{n(n-1)}{2}$ possible edges [25]. In these networks, the majority of nodes have a degree that is close to the average degree of the overall network.

Empirical analysis of large networks such as the Internet, the Web, or large phone networks found different characteristics [4]. In particular, the degree distribution mostly followed a power-law distribution. Networks that follow a power law distribution are called *scale-free networks*. In this section we will discuss properties of scale-free networks as well as structural measures of influence of nodes in a network, and models of message diffusion to provide the underlying rationale for our computational experiments.

2.1. Scale-Free Networks

Scale-free networks tend to contain centrally located and extensively high degree "hubs". They attach new members over time and the attachment prefers existing members that are already well connected. This principle of "preferential attachment" leads to interesting properties that have to be taken into consideration. Their behavior in terms of diffusion and communication processes is fundamentally different from that of random networks.

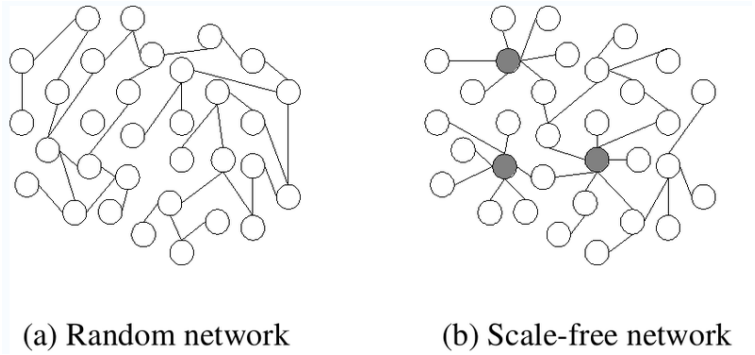

Figure 1 Random Networks versus Scale Free Networks [21]

In the past few years, there have been substantial advances in this area. Three concepts occupy a prominent place in contemporary thinking about complex networks. We will briefly discuss them in order to characterize our empirical data based on these characteristics:

1. **Degree Distribution**

The spread in the number of edges of a node, or node degree, is characterized by a distribution function $P(k)$, which describes the probability that a randomly selected node i has exactly k_i edges. Empirical results show that for most large networks, including the World Wide Web [4], Internet [26], or metabolic networks [36], the degree distribution follows a power law $P(k) = k^{-\alpha}$. Aiello et al. [2] analyzed the degree distribution of a phone call network and found a power law distribution with an exponent of $\alpha = 2.1$. The higher a power-law coefficient is, the fewer edges has a network

(with the same number of nodes). This also implies that with a higher power-law coefficient less people can be reached with a campaign.

2. **Clustering Coefficient**

A common property of social networks that can often be described by a power law distribution is that cliques are formed, representing groups of friends or acquaintances in which every member knows every other member. This inherent tendency to clusters is quantified by the clustering coefficient [59]. The clustering coefficient of a node i for an undirected graph is twice the ratio between the number of edges $|e_{jl}|$, which connect the k_i neighbors, divided by the total number of possible edges $k_i(k_i-1)$, where N_i is the set of neighbors of node i, and E the set of edges:

$$CC_i = \frac{2|e_{jl}|}{k_i(k_i-1)} : j,l \in N_i, e_{jl} \in E \tag{1}$$

The clustering coefficient of a whole network is the average clustering coefficient of all nodes $CC = \frac{1}{n}\sum_{i \in N} C_i$.

3. **Average Path Length**

The path length between two nodes of a network is defined as the number of edges between them. The minimal path length is the shortest path between two nodes (also called geodesic distance). The average path length is the average of all the minimum path lengths between all pairs of nodes in a network. The *small world concept* in simple terms describes the fact that despite their mostly large size, in most real networks there is a relatively short path between any two nodes. These small world graphs typically exhibit a number of characteristics, such as a low diameter and a low clustering coefficient.

Numerous studies have been initiated by the desire to understand various real systems ranging from communication networks to ecological webs. A comprehensive survey of different real networks and their characteristics is provided in Albert and Barabasi [3].

2.2. Structural Measures of Influence in a Network

In the social network community, a variety of measures were designed for the measurement of importance or prominence of nodes in a network [29, 14]. In the following, we will briefly summarize the most well-known centrality measures, as well as a number of link topological ranking measures, which describe possible candidate indicators for the power of influentials in message diffusion.

2.2.1. Centrality Measures: A centrality measure C is a function $C : N \to \Re$ that associates to each vertex $i \in N$ a non negative real number $C(i)$.

1. **Degree Centrality**

Degree centrality is the simplest centrality measure. The degree of a node i denoted by k_i, is the number of edges that are incident with it, or the number of nodes adjacent to it. For networks where the edges between nodes are directional, we have to distinguish between in-degree and out-degree. The out-degree centrality is defined as

$$C_{D_O}(i) = \sum_{j=1}^{n} a_{ij} \tag{2}$$

where a_{ij} is 1 in the binary adjacency matrix A if an edge from node i to j exists, otherwise it is 0. Similarly, the in-degree centrality is defined as

$$C_{D_I}(i) = \sum_{j=1}^{n} a_{ji} \tag{3}$$

where i describes the node i and a_{ji} is 1 if an edge from node j to i exists, otherwise it is 0.

2. **Closeness Centrality**

The closeness centrality measures how close a node is to all the other nodes in the set of vertices and is often used in social network analysis. As noted by Beauchamp [11], members occupying central locations with respect to closeness can be very productive in communicating information to the other members. Hakimi [34] and Sabidussi [53] developed a measure that central members are close, by stating that central nodes in a network have shortest paths to all other nodes. The closeness centrality index for directional relations is

$$C_C(i) = \frac{1}{\sum_{j=1}^{n} d(i,j)} \tag{4}$$

where $d(i,j)$ denotes the distance between node i and j, which is the minimum length of any path connecting i and j. Wassermann and Faust [58] proposed the following standardization of this measure to account for the size of the network:

$$C_C(i) = \frac{(n-1)}{\sum_{j=1}^{n} d(i,j)} \tag{5}$$

The problem with this definition of closeness centrality is that closeness is not defined unless the digraph is strongly connected, in such a way as each node has a direct path from i to j, otherwise, some of the $d(i,j)$ will be ∞ and the equation (5) would be undefined. Therefore, Lin [47] defined J_i as the number of nodes which are reachable from node i and propose to consider only the distances of these reachable nodes:

$$C_C(i) = \frac{J_i/(n-1)}{(\sum_{j=1}^{n} d(i,j))/J_i} \tag{6}$$

3. **Betweenness Centrality**

Interactions between two non-adjacent nodes might depend on the other nodes in the set of nodes, especially those nodes who lie on the path between the two. The node between the other two nodes can therefore control the interaction between the two non-adjacent nodes. The idea is that a node is central if it lies between other nodes on their geodesics, implying that, in order to have a large betweenness centrality, the node must be between many of the nodes via their geodesics. The betweenness centrality index is defined by Freemann [28] as

$$C_B(i) = \frac{\sum_{i \neq j \neq l} g_{jl}(i)}{g_{jl}} \tag{7}$$

where $g_{jl}(i)$ is the number of shortest paths linking the two nodes j and l containing node i. The computation of betweenness centrality is computationally expensive. Brandes [18] proposed an algorithm for betweenness that exploits the sparseness of typical networks to reduce the time complexity of this computation from $O(n^3)$ to $O(n^2+nk)$ and space complexity from $O(n^2)$ to $O(n+k)$. Moreover, other shortest-path based indices, like closeness, can be computed simultaneously within the same bounds.

4. **Eigenvector Centrality**

The eigenvector centrality is another measure of the importance of a node in a network. Here, the centrality of a node i is a function of the centrality of the nodes connected to i. Being nominated as powerful by someone seen by others as powerful should contribute more to one's perceived power. Let A again be the binary adjacency matrix of the network and $\vec{x}$ be the principal eigenvector corresponding to the maximum eigenvalue θ. The eigenvector centrality for a node i can be defined as a single element of the eigenvector calculated as:

$$C_E(i) = x_i = \frac{1}{\theta} \sum_{j=1}^{n} a_{ji} x_j \tag{8}$$

Here, each individual's status is merely proportional (not necessarily equal) to the weighted sum of the individuals to whom he is connected. Eigenvalues of large matrices are typically computed numerically [33]. For example, inverse iteration is an iterative eigenvalue algorithm to calculate the eigenvalues and eigenvectors of a matrix. The computational complexity of inverse iteration can be reduced to $O(n^2)$ if one first reduces the adjacency matrix A to a Hessenberg form.

5. **Edge-weighted Degree Centrality**

In the case of phone call networks, we have not only binary information about the communication of two members, but also a weighted digraph describing how often a member called another one, or how many short messages he sent. In this case, the entries of an adjacency matrix a_{ij} describe the numeric weights of a connection from node i to j. Each weighted graph can easily be transformed into a multigraph, where the same pair of vertices can be connected by multiple edges [50]. In this paper, we define edge-weighted degree centrality as:

$$C_{ED}(i) = \sum_{j=1}^{n} (a_{ij} + a_{ji}) \tag{9}$$

2.2.2. Link Topological Ranking Measures:

Most of the previously described centrality measures (except eigenvector centrality) disregard the type of node. There are very influential vertices to which a connection is more valuable than to others. With regard to social networks, a connection to a node with high centrality might be more valuable than to a node with only one neighbor. Web search engines leverage this information with HITS and PageRank probably being the most popular Web search algorithms.

Kleinberg [42] proposed a Web search algorithm called *HITS* (Hyperlink-Induced Topic Search) which identifies authoritative pages and a set of hub pages. Authoritative pages are pages which have many incoming links and hubs are pages that link to many related authorities. An iterative algorithm is used to find the equilibrium values for the authority and hub weights of a web page or node in a network respectively. This is reached if the difference of the weights between two iterations is less than a threshold value. For each page i a nonnegative authority weight $C_A(i)$ and a nonnegative hub weight $C_H(i)$ is associated. The weights of each type are normalized so their squares sum to 1 and are defined as:

$$C_A(i) = \sum_{j=1}^{n} a_{ji} C_H(j) \tag{10}$$

$$C_H(i) = \sum_{j=1}^{n} a_{ij} C_A(j) \tag{11}$$

where a_{ji} is 1 if an edge from node j to i exists otherwise 0. An iterative algorithm has been defined to find the equilibrium values for the weights. When the equilibrium is reached the most central nodes are those with the highest authority weight.

The *PageRank* algorithm, which was originally developed by Brin and Page [19], the founders of the Google search engine, maintains only a single metric for each web page. The so called PageRank is transmitted from the source page to the link target, and the value depends on the PageRank of the source page. So a link from a page that has large PageRank, such as the Yahoo home page, contributes more than a link from a page with low PageRank. The PageRank of page or node i is the sum of contributions from its incoming links or edges. A constant damping factor f is the

probability at each page that the "random surfer" will get bored and requests another random page. Additionally $(1-f)$ is added to each node. This is done because if a node has an out-degree of zero then his PageRank would be zero. This zero-value would be passed down to the original node. To avoid this, a constant value is added to the PageRank. The PageRank can be defined as:

$$C_{PR}(i) = (1-f) + f \sum_{j \in M_i} \frac{C_{PR}(j)}{C_{D_O}(j)} \tag{12}$$

where M_i is the set of source pages that link to i and $C_{D_O}(j)$ is the out-degree of page j as described in the previous subsection. The damping factor f is often set to value of 0.85 [19].

The PageRank is a variant of the eigenvector centrality with the difference that instead of the adjacency matrix, the Markov matrix is used. This is a reason, why we will only consider the PageRank in our experiments. A Markov matrix is the transition matrix for a finite Markov chain. Elements of the matrix must be real numbers in the closed interval [0, 1], where each element represents the transition probability from one page to the other page. Hence, if a connection exists between page i and j, then the element of the Markov matrix in row i and column j is $1/C_{D_O}(j)$, where $C_{D_O}(j)$ is the out-degree or out-degree centrality of the page j.

Web search algorithms such as PageRank focus on incoming links, since they count only the weighted number of incoming links and ignore the outgoing links. For information diffusion models, the out-degree is the more important measure. For this reason, we introduce a measure called *SenderRank*, which is based on the PageRank calculation apart from the direction of influence.

$$C_{SR}(i) = (1-f) + f \sum_{j \in L_i} C_{SR}(j) \tag{13}$$

where L_i is the set of pages the page i links to.

2.2.3. Comparisons of Centrality Measures: Several authors compared the performance of the existing centrality measures, either on empirical or on simulated data. Wassermann and Faust [58] provide a review of the early comparative studies. The first study of centrality measures was conducted by Freemann [29]. He analyzes the consistency of centrality measures with intuitions and their interpretability (e.g., control of communication, or communication activity). Freemann et al. [30] evaluated three centrality measures on four different graphs, all with $n = 5$ and found that betweenness best measured which member in the set of members was viewed most frequently as a leader. Another observation was that degree and betweenness centrality are important indicators for group performance (with respect to efficiency of problem solving) while closeness centrality was not even vaguely related to their experimental results on communication activity.

Bolland [13] studied four centrality measures. He examined a network data set giving influence relationships among forty people involved in educational policy-making. In addition, he conducted a Monte Carlo analysis by adding random and systematic variation to the network to obtain a number of noisy networks. Bolland's findings supported the earlier work of Freemann [29]. Specifically, degree-based measures of centrality were sensitive to small changes in network structure. Betweenness-based measures of centrality were considered useful and capable of capturing small changes in the network. Closeness centrality was found to be very sensitive to network change.

Costenbader and Valente [23] evaluated the stability of centrality measures when networks are sampled in the face of inaccurate or incomplete network data. It turned out that the most robust centrality measure is eigenvector centrality as a simple raw score followed by indegree centrality. They are less affected by sampling than outdegree and betweenness centrality.

Recently, Koschtuezki and Schreiber [43] calculated five different centrality measures for all the vertices of two networks and ordered the vertices according to their centrality. Using these rankings, they calculated the correlation of centrality measures and found that these correlations differ

between the two networks. The correlation between eigenvector centrality and degree centrality was however high in both networks.

Only recently, researchers have started to analyze and compare centrality measures based on the underlying network flow or diffusion model. Borgatti [16] showed how centrality measures can be matched to different kinds of network flow for which they are appropriate. A process can flow in different ways through a network. He classified network flow along different dimensions. For example, he makes a distinction, whether diffusion occurs via replication or transfer, if it is deterministic as in a computer network or undirected, describing a blind flow of traffic, as is the case with marketing messages. Based on the resulting classification, Borgatti [16] assigned respective centrality measures. The authors considered only flows with a pre-defined source and a target. In this paper, we will focus on centrality measures for the diffusion of marketing messages, where no predefined target exists, which is the most wide-spread application in viral marketing.

2.3. Diffusion in Networks

Closely linked to network theory are theories about the diffusion of messages or epidemics within networks. No matter if it is a virus disease, a computer virus, or a marketing message, they all have in common that they disperse through networks by spreading from one individual to their direct neighbors.

In the marketing literature, Bass [10] published an influential model on "A new product growth model for consumer durables" that motivated a stream of research on product innovation and diffusion. The Diffusion of Innovations Theory explains the dissemination of ideas; however, the actual spread of messages in a network requires different micro-level models closer in spirit to the models that have been developed to describe the diffusion of viruses.

2.3.1. Models for the Diffusion of Viruses: A number of models have been developed describing the spread of viruses of epidemics on a micro-level. Kermack and McKendrick [40] developed a mathematical framework called SIR-model for the spread of epidemics. SIR stands for susceptible-infected-removed. The probability of an individual to change from status susceptible to infected is given by δ and the probability that the status of a person changes from infected to removed is given by ν. How an infection spreads depends on $\lambda = \delta/\nu$ and the structure of the network [35].

Pastor-Satorras and Vespignani [51] analyzed the spreading properties of viruses by using data reported by the Virus Bulletin. They analyzed in particular the survival probability of homogeneous groups of viruses. For larger time frames, they found an exponential decay of the virus diffusion. Chrisley et al. [22] simulated the transmission of infections to identify high-risk individuals. The authors analyzed centrality measures for their ability to identify high-risk individuals and found that degree centrality appeared to perform at least as well as alternative measures for this application.

2.3.2. A Model for Message Spreading in Social Networks: In the following, we will introduce a general model describing the spread of marketing messages in a customer network. We define the probability $P_j(i)$ of a node with k_i neighbors passing on a message to $j \leq k_i$ of these neighbors based on a binomial distribution and a communication probability κ. The j nodes are then selected randomly among the neighbors.

$$P_j(i) = \binom{k_i}{j} \kappa^j (1-\kappa)^{k-j} \tag{14}$$

The difference between spreading of viruses and the spreading of messages is that virus infection can be cured, while once a message is received, a node stays "infected". Instead we do have a decay of the signal because a message gets old with time and might not be forwarded with increasing age.

To account for that, a decay parameter τ is defined. Initially every new message has a value of 1 that will be multiplied with a decay parameter τ for every hop in the network. This decay describes the level of trust a person puts into a message. Two different diffusion models were used, a model with exponential decay, and one with power law decay. In the *model with exponential decay*, the signal strength β_i of a message at node i which is q hops away from the original creator of the message is defined as

$$\beta_i = \tau^q \tag{15}$$

with $0 < \tau < 1$. In contrast, the *model with power law decay*, $\tau > 0$, uses

$$\beta_i = (q+1)^{-\tau} \tag{16}$$

For example, a power law decay with a power law coefficient of 1.75 has been found by Wu et al. [62], who provided a study in another domain of textual similarity of Stanford student homepages.

A node only forwards a message the first time he receives it. If he receives a message for the second time from another source, it is considered old and will not be propagated any more. However, if a node receives a message several times, the trust level of a message at node i adds up to σ_i. If a message was received s times at a node, then

$$\sigma_i = \sum_{r=1}^{s} \beta_r \tag{17}$$

In our computational experiments, we have set a threshold value ε, and if $\sigma_i > \varepsilon$, then a node has received the message. In addition, ω limits the number of times each message can be passed on to a neighbor, e.g., only 10 transmissions are allowed.

3. Computational Experiments

In a first step, we would like to find out, how well individuals distribute messages based on the topology of their network. For this purpose, in our experiments we sent a message to a set of customers selected by a particular centrality measure and then analyzed, how many customers could be reached overall based on a particular diffusion model.

3.1. Experimental Setup

We have developed a software framework for the simulation of diffusion processes in social networks which consists of two main components. A *network model* generates different types of networks. Based on the degree distribution and the number of nodes, we can generate different instances of networks adhering to the characteristics of ER networks or scale-free networks. In addition, we had anonymized data about the customer network of a telecom company.

The second component is a set of *diffusion models* that models communication and message distribution between customers. We used the model with exponential decay and the one with power law decay, as described in section 2.3. Different treatments were evaluated in the experiments: different networks, different diffusion models with different parameter-settings and different centrality measures. In the following we will introduce these treatments in more detail and then present the results.

3.1.1. Network Models: We got two real world networks (call detail records) from a mobile phone provider. Their main characteristics are described in the first lines of Table 1. The samples NW1 and NW2 were drawn following a breadth-first search, i.e., by selecting a random node and including all its neighbors in the network and then all the neighbors of these neighbors until a certain number of nodes was reached. The two samples NW1 and NW2 have different initial nodes. The networks included a number of boundary nodes, mostly customers from other phone providers, whose communication behavior is not available. The network information consisted of the two anonymized phone-numbers of the calling customer and the callee. Furthermore, we had monthly aggregates about the communication behavior between the customers which included following attributes: the number of voice calls, the number of minutes of a voice call, the number of short messages (sms), the number of multimedia messages (mms), and the average communication usage (number of sms and voice calls) of the callee. We also gathered attributes about the calling customer, certainly only if the calling customer is a customer of the telecommunication provider. The attributes included the gender, the age, and the zip-code of the customer. The first network aggregated one month of call detail records for the edge weights (which was calculated as number of sms+number of voice calls+number of mms) of the network and the second network two months.

We have also drawn two smaller samples, NW3 and NW4 from NW1 using breadth-first search. In NW4 the boundary nodes from external networks were excluded, while in NW1 - NW3 they were included. We only used anonymized data in this study. Table 1 also provides the clustering coefficients and the average path lengths for different networks.

In addition to the real world networks, we generated scale-free networks NW5 - NW7 with different power-law coefficients resembling real-world networks that have been analyzed in the literature (2.1, 2.5, and 2.8) and two ER-networks (NW8, NW9). This allowed us to evaluate centrality measures based on networks with other power-law coefficients. NW5 to NW7 did also not have the large number of boundary nodes, which are a consequence of the sampling in NW1 to NW4 and simulate a complete scale-free network.

In addition, the table provides characteristics of networks from the literature about a network of actors in the same movies, co-authorship in Mathematics, and the WWW to be able to compare them against the characteristics of NW1-9.

We analyzed the degree distribution of customers from our telecom provider. Figures 2 and 3 illustrate the degree distributions of NW1. Based on the logarithm of the observed values, we could identify a power law coefficient $\alpha = 2.8$ for the out-degree distribution, and 2.4 for the in-degree distribution. These distributional assumptions could be confirmed using a Kolmogorov-Smirnov test and a significance level of 0.05. In comparison, Aiello et al. [2] analyzed a phone network of 53,000,000 nodes and observed a power law coefficient of 2.1 for both in- and out-degree distributions. As the power law coefficient decreases a larger number of customers has a very high number of degrees.

The power-law coefficients of the WWW, which were analyzed by different authors [4, 45, 20] were between 2.4 and 2.7 for the out-degree distribution and 2.1 for the in-degree distribution. The network of movie actors follows a power law distribution with $\alpha = 2.3$ for in- and out-degree [8].

3.1.2. Diffusion Models: In our computational experiments, we have combined different levels of the decay constant (τ), the threshold for believing a message (ε), the probability for transmission (κ), and the lifetime of a message (ω) as treatments (see Table 2) to explore the parameter space. We evaluated both kinds of diffusion models (exponential and power law) with different model parameters. For the power law diffusion we used decay factor τ of 1.75 as in Adamic and Adar [1]. In the exponential diffusion model we used a decay constant of 1 which indicates no diffusion, a decay of 0.5, 0.65, and 0.7. For the threshold value ε we used low values of 0.3 and 0.2. For the probability of transmission κ we evaluated values ranging from 23 to 100 percent. The lifetime ω of the message was set at 10 or 20. In each line of Table 2, we changed only one or two parameters, in order to analyze the impact of these changes.

Table 1 Characteristics of network topology.

Network (NW)	Number of nodes	Number of edges	Type of network	External provider (boundary nodes)	Clustering coefficient	Avg. path length
NW 1	29,000	55,013	Real-world	Included (23,905)	0.1849	7.782
NW 2	54,839	118,475	Real-world	Included (42,227)	0.193	10.789
NW 3 (part of NW 1)	6,721	11,309	Real-world	Included (5,449)	0.182	10.44
NW 4 (part of NW 1)	6,492	16,240	Real-world	Excluded (1,115)	0.3265	10.794
NW 5 (power law: 2.1)	5,000	14,192	Simulated	-	0.0018	8.06
NW 6 (power law: 2.5)	5,000	12,987	Simulated	-	0.0010	8.6
NW 7 (power law: 2.8)	5,000	12,375	Simulated	-	0.0008	8.9
NW 8 (ER, max. out dgr. 5)	5,000	14,861	Simulated	-	0.0014	7.78
NW 9 (ER, max. out dgr. 10)	5,000	27,814	Simulated	-	0.0025	5.32
Movie actors*	225,226	-	-	-	0.79	3.65
WWW**	153,127	-	-	-	0.108	3.1
Math co-authorship***	70,975	-	-	-	0.59	9.5

* [59], ** [1], *** [9]

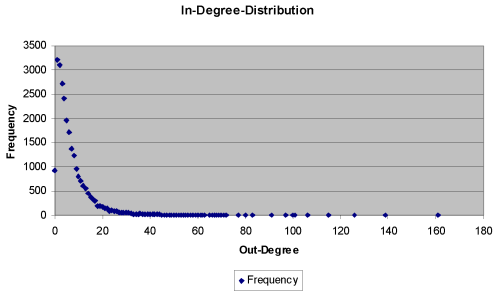

Figure 2 In-degree distribution of our data sample

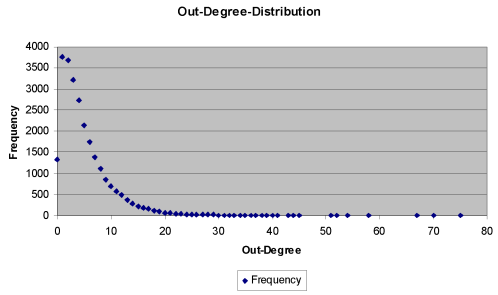

Figure 3 Out-degree distribution of our data sample

Table 2 Treatment Variables

Model	type	τ	ε	κ	ω
1	exponential	0.5	0.3	60	10
2	exponential	1	0.3	60	10
3	exponential	1	0.3	30	10
4	exponential	0.5	0.3	100	10
5	exponential	0.5	0.2	100	10
6	exponential	0.7	0.3	40	10
7	exponential	0.65	0.3	23	10
8	power law	1.75	0.3	40	10
9	power law	1.75	0.3	60	10
10	power law	1.75	0.3	100	10
11	power law	1.75	0.3	40	20

3.1.3. Centrality Measures: In addition to the centrality measures introduced in section 2.2.1, we implemented two more measures: the weighted PageRank and the weighted SenderRank. These measures take the weights of the edges into account. The formulas equal 12 and 13, with the difference that the PageRank and SenderRank are multiplied by the communication intensity, which is the sum of multimedia messages, short messages and voice calls received or sent respectively. The weighted PageRank $C_{WPR}(i)$ is defined as:

$$C_{WPR}(i) = (1-d) + d \sum_{j \in M_i} \frac{C_{PR}(j)(a_{ij} + a_{ji})}{\sum_{i=1}^{n} a_{ji}} \tag{18}$$

The weighted SenderRank $C_{WSR}(i)$ is defined as:

$$C_{WSR}(i) = (1-d) + d \sum_{j \in L_i} C_{SR}(j)(a_{ij} + a_{ji}) \tag{19}$$

3.2. Results

We have combined all treatments resulting in 1,089 (9 networks * 11 models * 11 centrality measures) different results. Gain curves, aka lift charts [61], are regularly used to evaluate direct marketing campaigns and describe the percentage of target customers reached based on the percentage of the customers that were addressed in a campaign. In our gain charts we will use absolute numbers and plot the number of customers reached on average after 10 iterations against the number of customers that were initially addressed, i.e. the ones selected based on centrality.

In order to avoid bias in the evaluation of the random selection, we applied the message-spreading 10 times on each network and calculated averages. We selected the most important results and will present them in two steps. First we will evaluate different networks, while keeping the other treatments constant. Then the effect of different diffusion models will be analyzed, keeping the network constant.

3.2.1. Evaluation of Different Networks

Real Phone Networks:

Within real networks the performance of the centrality measures was similarly independent of the network. As an example, Figure 4 illustrates the evaluation of NW2 with the following parameters: $\tau = 0.5$, $\varepsilon = 0.3$, $\kappa = 60\%$, $\omega = 10$. The histogram visualizes the results of a random selection and selections based on the 11 different centrality measures when for example 30 and 90 customers are initially addressed. This type of histogram allows us to display the performance of all metrics in one chart. In addition, Figure 5 presents the gain curves of a number of selected centrality measures. Each point on a gain curve illustrates the average of 10 simulation rounds. The probability of transmission κ leads to some randomness and is responsible for the fact that the number of customers reached can also stay constant or even decrease in some cases although more customers have been selected initially.

The Figures 4 and 5 illustrate that centrality based selection leads to a significant gain in reaching people compared to a random selection of customers, no matter which centrality measure is considered. The best centrality measures were out-degree centrality and SenderRank, which performed equally well across all quantiles. The weighted SenderRank and betweeness centrality were second best followed by the edge-weighted centrality and closeness centrality. The worst centrality measure for information diffusion of this sort were in-degree, authorities, and PageRank, which all stress the in-degree. Very similar results could be established for NW1, NW3 and NW4.

Simulated Scale-Free Networks

The power-law simulated scale-free networks (NW 5-7) provided results similar to the ones of the real phone networks NW 1-4 (see Figures 6 and 7). The best measures were SenderRank, out-degree and also weighted SenderRank. Edge-weighted degree centrality was followed by closeness.

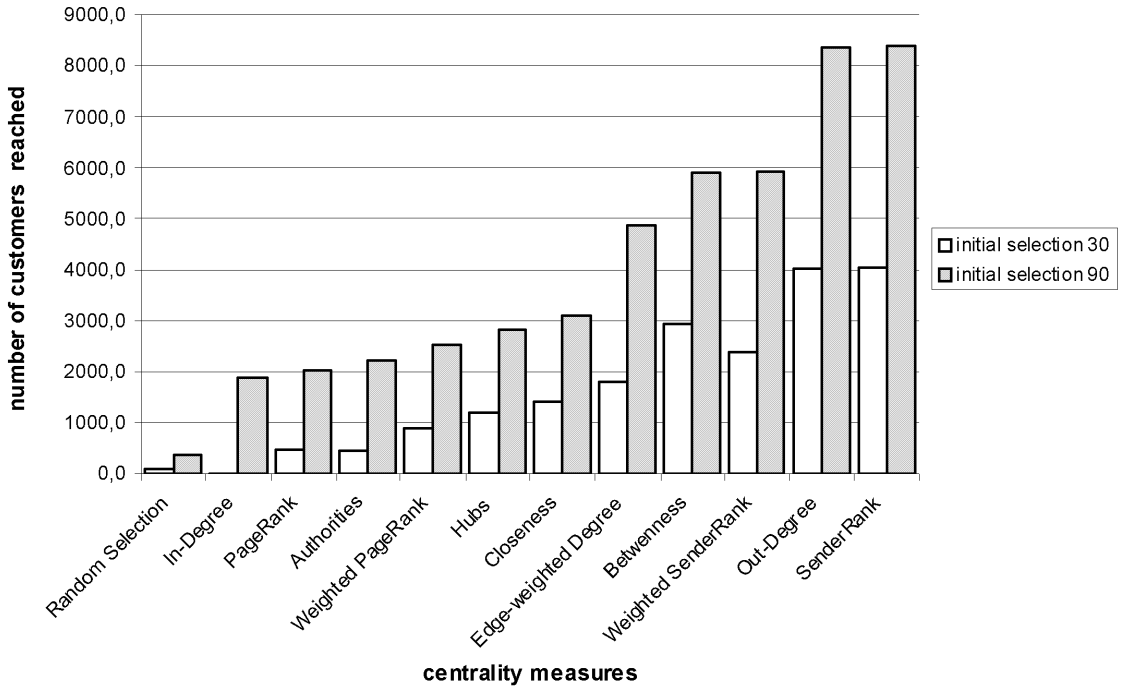

Figure 4 **Evaluation of centrality measures of real world network NW2, histogram**

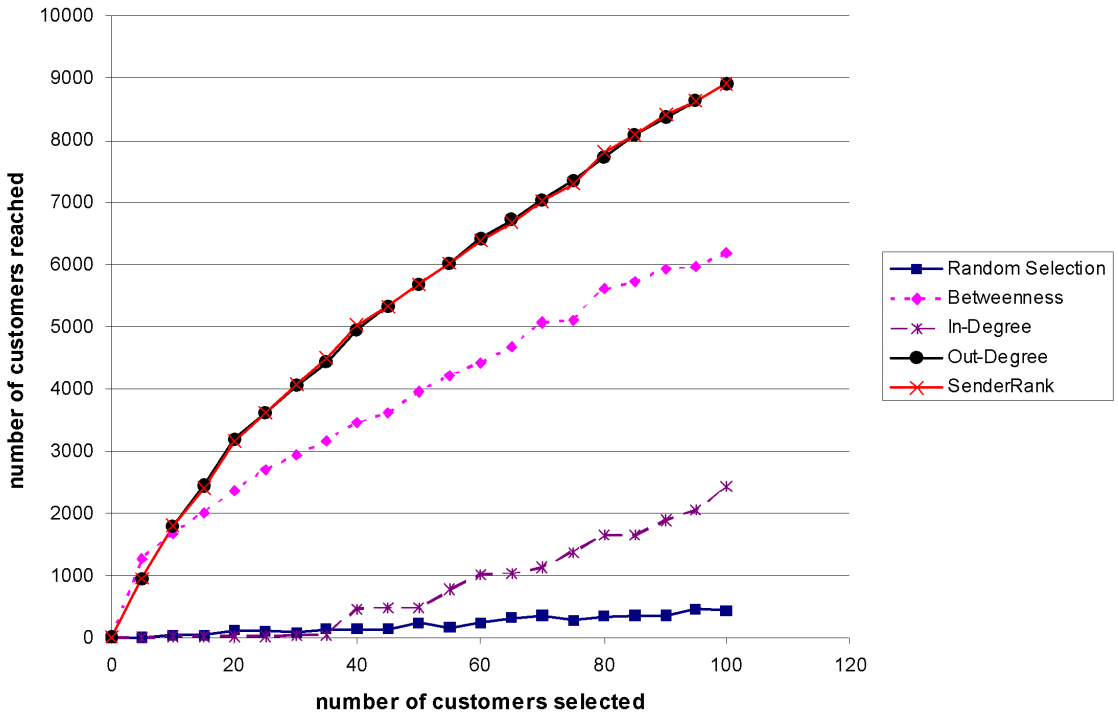

Figure 5 **Evaluation of centrality measures of real world network NW2, gain-chart**

A difference to real world networks in these networks was that closeness centrality performed better than betweenness centrality. This might be due to the fact that real networks contain more clusters

than simulated ones. Betweenness centrality better identifies *bridge*-nodes connecting different clusters of a network.

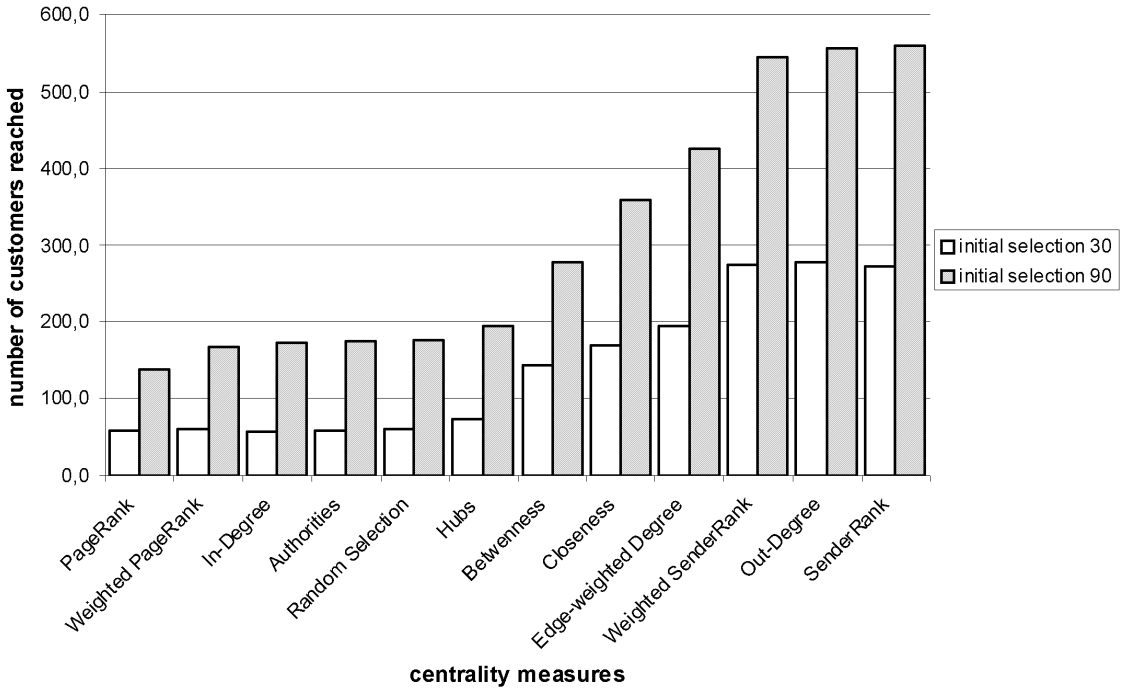

Figure 6 Scale-free network NW6, histogram

Simulated ER-Networks

ER-networks were mainly used for comparison, since most real-world networks are scale-free. In ER-networks without clusters (NW 7 and 8) the best measure was again the SenderRank followed by closeness and out-degree centrality (see Figures 8 and 9). However, in case of ER-networks there was less difference between the different centrality measures.

We have also looked at the correlation of the different centrality scores. While the correlation was typcially low between 0 and 0.5 for the different networks. The notable exception was the correlation between SenderRank and out-degree centrality, which was sometimes even close to 1. Also in-degree, authorities and PageRank were highly correlated (0.8-0.9). In other words, out-degree centrality turns out to implement the SenderRank in many scale-free networks with considerably lower computational complexity.

3.2.2. Evaluation of Diffusion Models: In the previous section we could already see that SenderRank and out-degree centrality performed very well. Much of this might be due to the assumptions of the diffusion model (signal strength τ, probability of message transmission κ, threshold ε, power law and exponential diffusion). Our diffusion model described in section 2.3 is very generic, and allows a variety of parameter settings modeling very different real-world diffusion processes. We analyzed different diffusion models based on network NW 2, in order to provide a sensitivity analysis of this finding with respect to different assumptions about the message diffusion. We have analyzed the parameter settings also on the other networks but since the results were similar we present these results as an example. We have evaluated all the parameter settings in Table 2. Due to space restrictions we will illustrate addtional treatments in the Appendix. Interestingly, out-degree and SenderRank were the top-ranked centrality measures in almost all treatments.

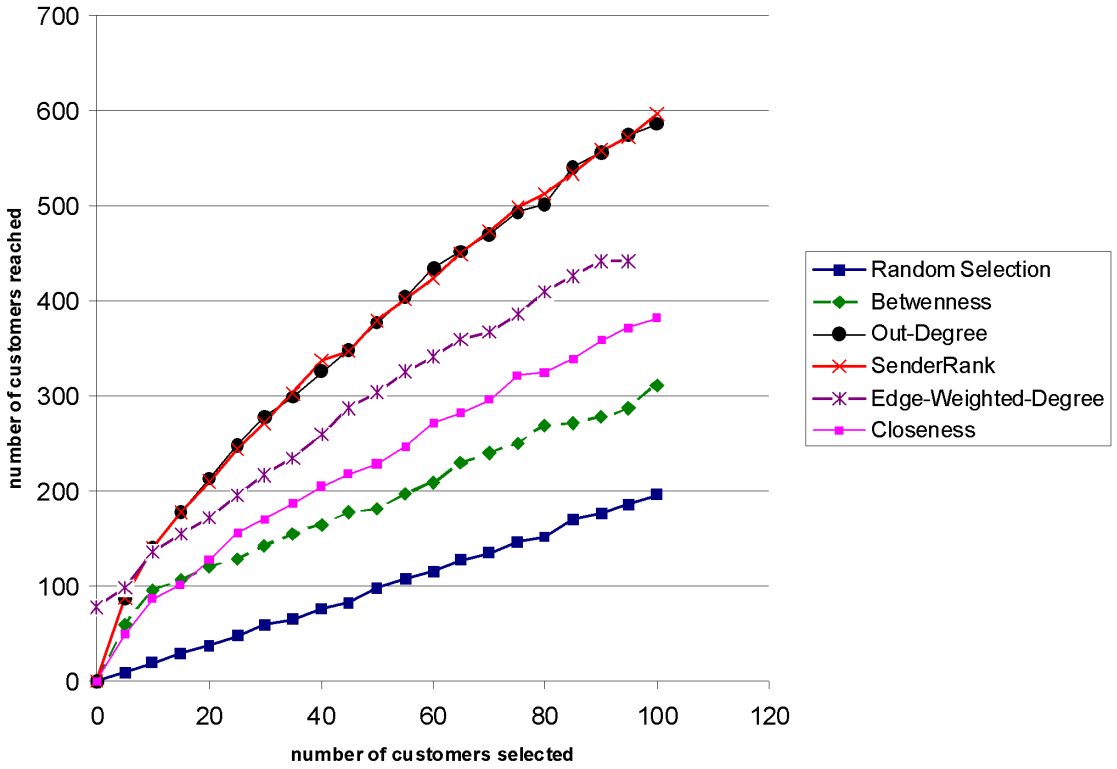

Figure 7 **Scale-free network NW6, gain chart**

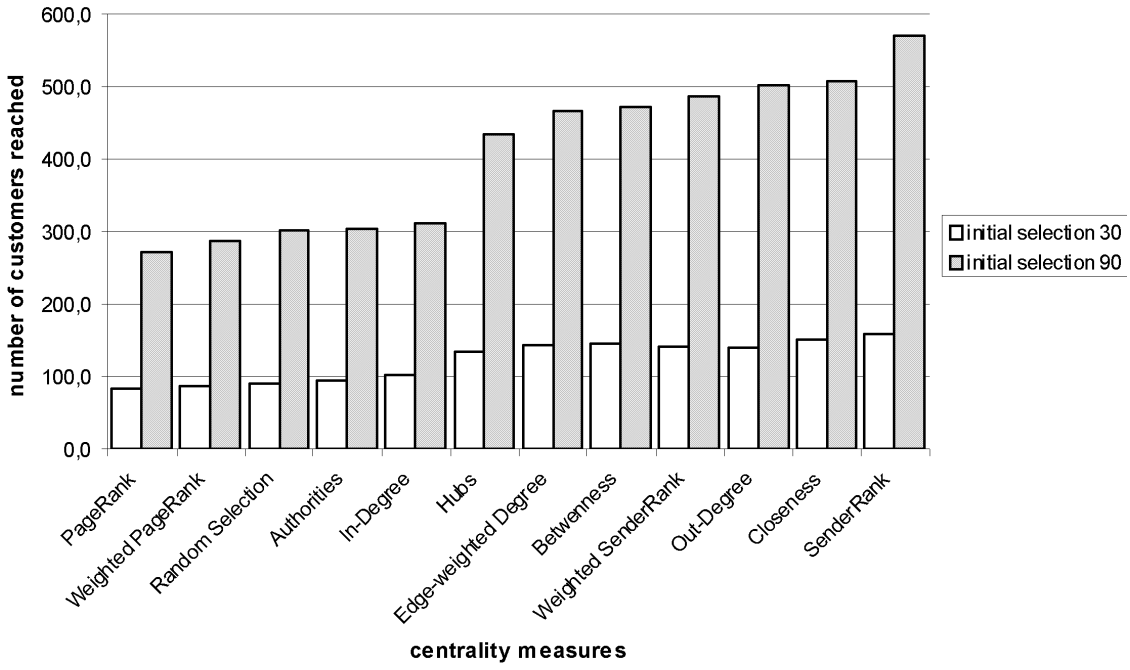

Figure 8 **ER network NW8, histogram**

3.2.3. Evaluation of Cluster Centralities: We have already seen, that based on the above network topology and diffusion models, the local vicinity of a node does play a role. So instead

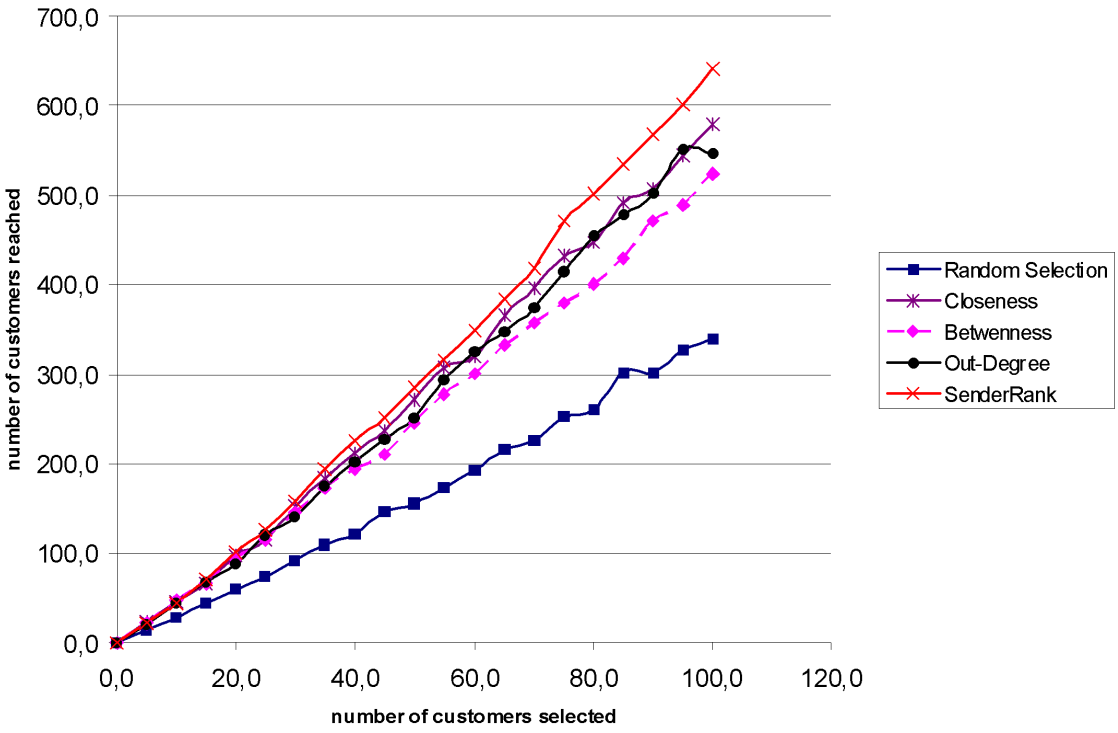

Figure 9 **ER network NW8, gain chart**

of calculating the global centrality for the entire network, an alternative might be to first derive clusters of highly connected nodes (e.g., groups of friends) and then find the most central customer among them.

The following analysis is based on NW 1 with diffusion model parameters $\tau = 0.5$, $\varepsilon = 0.3$, $\kappa = 60\%$, $\omega = 10$. Different cluster centrality measures were evaluated. The best performing measures of the previous evaluations were chosen for cluster centralities, namely out-degree, degree (which is the sum of in-degree and out-degree), SenderRank, betweenness and closeness centrality. The Minimum Spanning Tree clustering algorithm (see [57, 52, 44]) was used, because it has low computational complexity and is therefore applicable to large networks. We got 752 clusters, most of them having less than 30 nodes, but also a few big clusters of several hundred nodes. Only clusters with at least 5 nodes were taken into account, which has shown to improve the performance of the approach.

There are several ways, how one can rank-order these central cluster nodes. One way is to sort the clusters based on their size. During the selection, we iterated through all clusters, which exhibit the minimum size, and selected the most central node in the largest cluster followed by the most central node in the second largest cluster, until the most central nodes of all clusters were selected. In the next iteration we selected the second node ranked by centrality, etc. When one third of the nodes in a cluster was selected the cluster was not considered any more.

The results of the cluster centrality based selection compared to a network centrality based selection of NW1 are illustrated in Figure 10. Cluster centralities are marked with the prefix "Clustered". The cluster centrality measures which performed best are also out-degree, degree and SenderRank. However, clustering centralities performed worse than their equivalents calculated on the entire network. One reason might be the network topology. Sending a message to a node which has the highest out-degree of a cluster might ignore the fact that nodes with much higher out-degree centrality exist in another cluster, but were ignored in the iterative selection process described above. Overall, the nearest vicinity of a node turns out to be most important for various diffusion models. So, even modified strategies for the selection of initial nodes have little impact

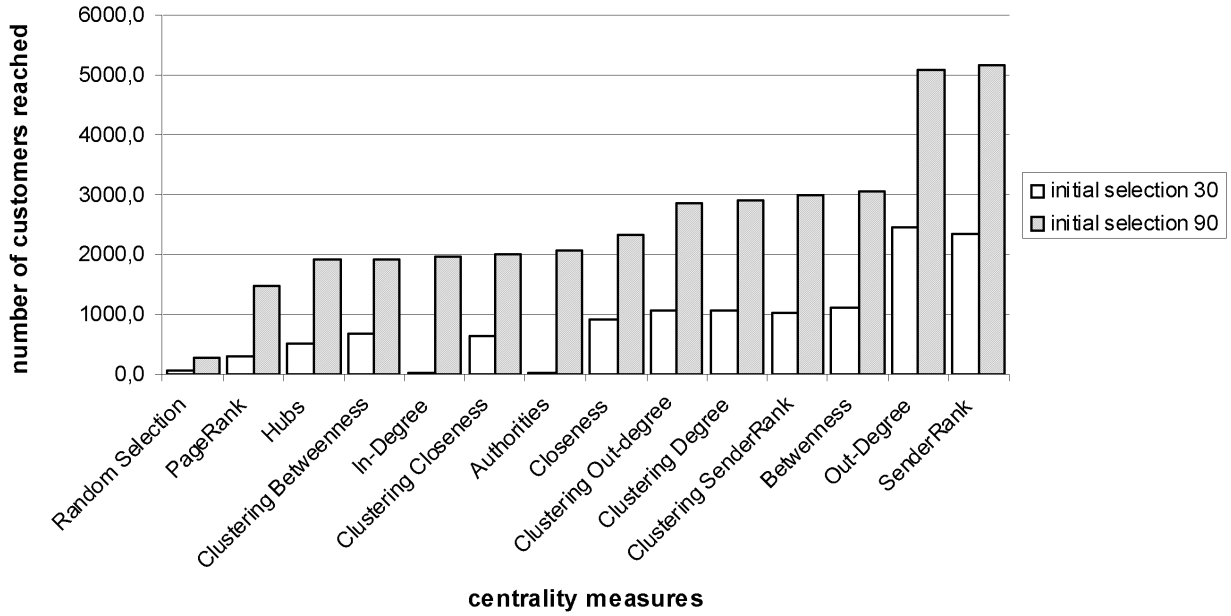

Figure 10 Evaluation of cluster centralities in NW1

on the result. Additional analyses of different networks with different parameter settings confirmed these findings.

3.2.4. Diffusion Processes over Time: In particular, for campaigns that need to reach many people in very short time, it can be of interest to analyze how many customers can be reached after only a few transmissions. If we assume each transmission is one step in time, we can analyze, which is the fastest way to reach customers. In Figure 11 we assumed a transmission probability $\kappa = 100\%$ and no decay (τ). For the illustration of these effects, we generated a small power law distributed network (with a power law coefficient $\alpha = 2.1$) with 300 nodes and 458 edges only, where the performance differences on the first few transmissions are easier to identify.

Figure 11 shows that in steps 2-5, the complex SenderRank performed best, and was slightly better than the simple out-degree measure. In step 4 betweenness and closeness outperform out-degree and in step 6 they also outperform SenderRank. This might be due to the fact that these centralities take into account all the nodes in a network and not only the close vicinity. In step 7, out-degree outperforms SenderRank and closeness centrality. After 15 steps the centrality measures converged since all nodes were reached. Since betweenness and closeness centrality had exactly the same values we described only one of them in the chart.

These experiments provide also information on how well the different centrality measures are doing, if you expect them to spread very far, i.e., the message is passed on many times from neighbor to neighbor, as might be the case for important information (such as hurricane warnings), or if you are talking about short-lived marketing messages about a new product, that might only be passed on a few times to friends that are interested in a particular product.

3.2.5. Summary: Overall, centrality measures achieved a very high lift of up to 20 compared to a random selection in our message spreading experiments. The lift can be calculated as the number of customers reached divided by the number of customers selected. In most cases SenderRank performed best, but is closely followed by or even equal to the simple out-degree centrality, which has a constant computational complexity. In scale-free networks the scores of both measures were highly correlated. Along the same lines, Fortunato et al. [27] showed recently for an analysis of Web search algorithms that the approximation of PageRank via the in-degree can be highly accurate.

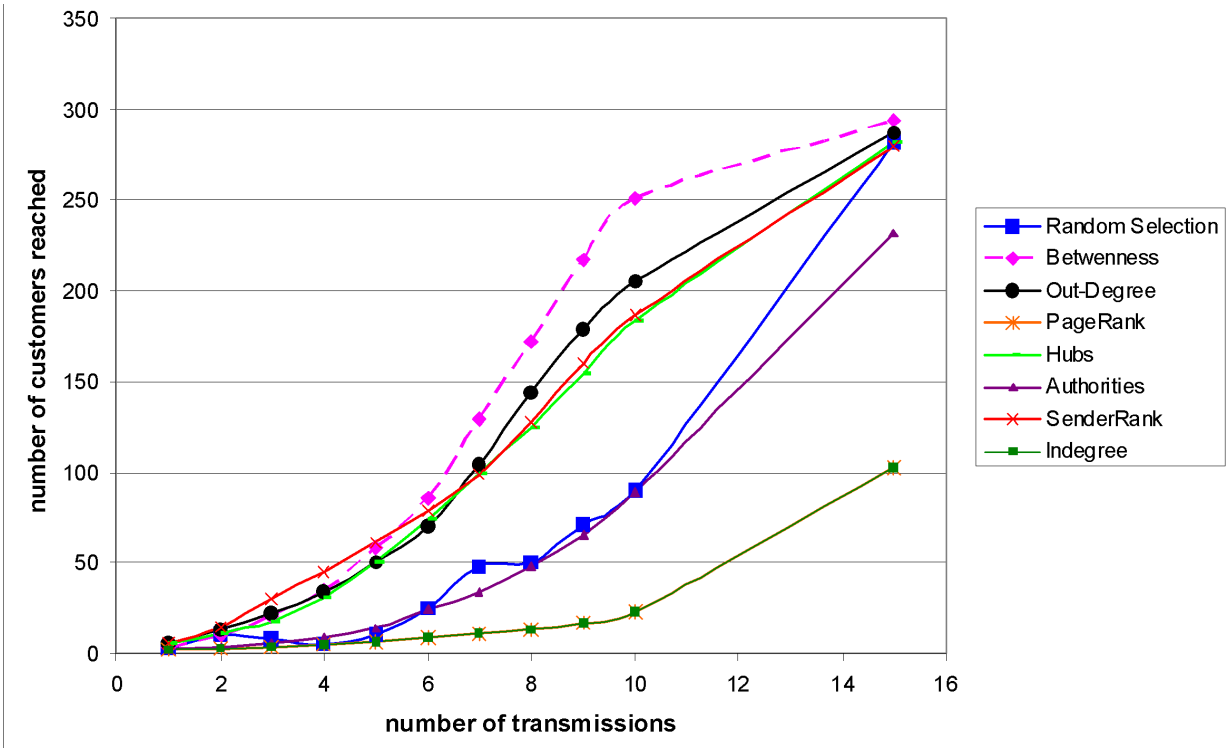

Figure 11 **Time-dependent analysis of centrality measures**

This result is robust against different types of network topologies and different assumptions about message distribution. Only for treatments with a very high signal strength (see the sensitivity analysis for τ in the Appendix) those metrics which emphasize not only the nearest vicinity perform much better. We did not find a different ranking of the centrality measures for the different samples of our real network, and also found a roughly similar ranking for simulated scale-free networks. Within a certain type of network (real, power-law simulated and ER simulated) the rankings of the metrics are highly correlated. Between different types of networks the correlation was lower, but still SenderRank and out-degree came out best most of the time. Also, the size of the network n did not lead to differences in the rankings.

The results provide marketing decision makers with a clear recommendation of what measure best describes the structural capability of a customer to spread a message. It is worthwile to emphasize, however, that this advice is suitable for viral marketing campaigns increasing the awareness of a product, as outlined in Section 1. In cases where we can assume that all consumers are aware of the product's existence, and seeding merely affects their expected value from adopting it, Sundararajan [55] has recently shown that under certain assumptions it might be a better strategy to seed the fringes rather than the hubs of a network.

4. Conclusions

Data mining, in particular classification, has become an integral part of decision support in CRM in areas such as campaign management or churn prediction. Respective CRM applications typically ignore the position of a person in the customer network. Many companies do not only have data about individual customer behavior, but also about the social network of customers. Although information about the content of a communication is typically not available, the frequency of interactions and the resulting topological information about the customer network can be leveraged

in viral marketing campaigns, for purposes of usage stimulation, or churn management (see Section 1.2).

Recent research has found that centrality measures need to be matched to the network flow for which they are appropriate. This topic has found little attention in the viral marketing literature, which exhibits diffusion patterns that are different to the ones in disease spreading or other domains. While much of the communication among customers cannot be observed, call detail records provide a sample of communication relationships that can be used to derive an estimator for the influence of a customer. Centrality can be used as such an estimator.

In this paper, we have compared different centrality measures with respect to their impact on message diffusion in social networks. We have evaluated existing measures and also introduced the SenderRank as a new one focused on message distribution in social networks. Based on a number of computational experiments on artificial and on real networks we observed a significant lift when using central customers in message diffusion, but also observed differences in the various centrality measures depending on the underlying network topology and diffusion process. We also found that the simple out-degree centrality achieves very good results compared to computationally more complex centrality measures. Only the SenderRank achieved a comparable performance.

There are a number of caveats, one might want to keep in mind. As indicated, call detail records can be used to derive an estimator for the social interaction pattern of a customer. However, this data reflects only parts of the social interaction of a customer and might be biased. Also, in this study, we ignore customer preferences for specific messages. The centrality of a customer describes only his basic capability to distribute a message in a network. It does say nothing about the tastes or preferences of this customer. It is, therefore, important to complement these metrics with the results of other predictive models, such as logit models or decision trees as they are regularly used nowadays in campaign management to predict the affinity of a customer for a certain product or brand. The centrality of a customer in his network is orthogonal information. It might be used to select target customers in combination with the likelihood of a customer of responding to a message, or one might also include only those customers in the network analysis that have a certain likelihood of being interested in a message.

While knowledge about the merits of different centrality measures can be helpful, when important messages should be distributed in a network, data privacy is a crucial aspect in all campaign management applications. The concerns that people have over the collection of personal data naturally extends to any analytic capabilities applied to the data. Respective applications are typically regulated by privacy laws and company guidelines. Misuse of campaigns can always have many adverse effects such as customer churn that marketers need to consider. On the other hand, intelligent use of analytical techniques can help companies with large numbers of customers to find and address the right customers with information that is of interest to them.

References

[1] L. A. Adamic, E. Adar, Friends and neighbors on the web, Social Networks 25(3) (2003) 211–230.

[2] W. Aiello, F. Chung, L. Lu, A random graph model for massive graphs, in: Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, 2000.

[3] R. Albert, A. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74.

[4] R. Albert, H. Jeong, A. Barabási, Diameter of the world-wide web, Nature 401 (1999) 130–131.

[5] E. W. Anderson, Customer satisfaction and word of mouth, Journal of Service Research 1 (1) (1998) 5–17.

[6] C. Apte, B. Liu, E. Pednault, P. Smyth, Business applications of data mining, CACM 45.

[7] H. S. Bansal, P. A. Voyer, Word of mouth processes within a services purchase decision context, Journal of Service Research 3 (2) (2000) 166–177.

[8] A. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 no. 5439 (1999) 509 – 512.

[9] A. Barabási, H. Jeong, E. Ravasz, Z. Nda, A. Schubert, T. Vicsek, Deterministic scale-free networks, Physica A 311 (3-4) (2002) 590–614.

[10] F. M. Bass, A new product growth model for consumer durables., Management Science 18 (1969) 215–227.

[11] M. Beauchamp, An improved index of centrality, Behavioral Science 10 (1965) 161–163.

[12] M. Bichler, C. Kiss, A comparison of logistic regression, k-nearest neighbor, and decision tree induction for campaign management, in: Tenth Americas Conference on Information Systems (AMCIS), New York, 2004.

[13] J. M. Bolland, Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks, Social Networks 10 (1988) 233–253.

[14] P. Bonacich, Power and centrality: a family of measures, American Journal of Sociaology 92 (1987) 1170–1182.

[15] P. F. Bone, Word of mouth effects on short-term and long-term product judgements, Journal of Business Research 32 (3) (1995) 213–223.

[16] S. P. Borgatti, Centrality and network flow, Social Networks 27 (2005) 55–71.

[17] D. Bowman, D. Narayandas, Managing customer-initiated contacts with manufacturers: The impact on share of category requirements and word-of-mouth behavior, Journal of Marketing Research 38 (2001) 291–297.

[18] U. Brandes, A faster algorithm for betweenness centrality, Journal of Mathematical Sociology 25 (2) (2001) 163–177.

[19] S. Brin, L. Page, The anatomy of a largescale hypertextual web search engine, in: 7th International Word Wide Web Conference, Brisbane, Australia, 1998.

[20] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajalopagan, R. Stata, A. Tomkins, J. Wiener, Graph structures in the web, WWW9/Computer Networks 33(16) (2000) 309–320.

[21] M. Castells, Informationalism, Networks, and the Network Society: A Theoretical Blueprint, The network society: a cross-cultural perspective, Northampton, 2004.

[22] R. M. Chrisley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett, J. Turner, Infection in social networks: Using network analysis to identify high-risk individuals, American journal of epidemiology 162 (10) (2005) 1024–1031.

[23] E. Costenbader, T. W. Valente, The stability of centrality measures when networks are sampled, Social Networks 25 (2003) 283–307.

[24] P. J. Danaher, R. T. Rust, Indirect financial benefits from service quality, Quality Management Journal 3 (2) (1996) 63–75.

[25] P. Erdoes, A. Renyi, On random graphs, Publicationes Mathematicae 6 (1959) 290–297.

[26] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, ACM SIGCOMM, Comput. Commun. Rev. 29 (251).

[27] S. Fortunato, M. Buguna, A. Fammini, F. Menczer, How to make the top ten: Approximating pagerank from in-degree, in: WWW 2006, Edinburgh, UK, 2006.

[28] L. C. Freemann, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.

[29] L. C. Freemann, Centrality in social networks: I. conceptual clarification, Social Networks 1 (215-239).

[30] L. C. Freemann, D. Roeder, R. R. Mulholland, Centrality in social networks: Ii. experimental results, Social Networks 2 (1980) 119–141.

[31] D. Godes, D. Mayzlin, Using online conversation to study word-of-mouth communication, Marketing Science 23 (4) (2004) 545–560.

[32] D. Godes, D. Mayzlin, Using online conversations to study word-of-mouth communication, Marketing Science 23 (4) (2004) 545–560.

[33] G. H. Golub, C. F. Van Loan, Matrix Computations, 3rd ed., Johns Hopkins University Press, 1996.

[34] S. Hakimi, Optimum locations of switching centers and the absolute centers and medians of a graph, Operations Reserach 12 (1965) 450–459.

[35] O. Hein, M. Schwind, W. Koenig, The impact of fat tailed degree distribution on diffusion and communication processes, Wirtschaftsinformatik 48 (4) (2006) 267–275.

[36] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. Barabsi, The large-scale organization of metabolic networks, Nature 407 (2000) 651.

[37] J.-J. Jonker, N. Piersma, R. Potharst, A decision support system for direct mailing decisions, Decision Support Systems 42 (2006) 915–925.

[38] P. Kannan, R. Rao, Introduction to the special issue: decision support issues in customer relationship management and interactive marketing for e-commerce, Decision Support Systems 32 (2001) 83–84.

[39] E. Keller, J. Berry, The Influentials, Free Press, 2003.

[40] W. Kermack, A. McKendrick, A contribution to the mathematical theory of epidemics, Proceedings of the Royal Society of London Series A 115 (1927) 700–721.

[41] J. Kirby, Connected Marketing, Butterworth-Heineman, an imprint of Elsevier, 2005.

[42] J. Kleinberg, Auth. sources in hyperlinked environment, in: CM-SIAM Symposium on Discrete Algorithms, 1998.

[43] D. Koschuetzki, F. Schreiber, Comparison of centralities for biological networks, in: German Conference Bioinformatics (GCB'04), vol. P-53, 2004.

[44] J. B. Kruskal, On the shortest spanning subtree and the traveling salesman problem, vol. 7, 1956.

[45] R. Kumar, P. Raghavan, S. Rajalopagan, A. Tomkins, Trawling the web for emerging cyber-communities, in: 8th International World Wide Web Conference, 1999.

[46] J. Leskovec, L. Adamic, B. Huberman, The dynamics of viral marketing, ACM Transactions on the Web 1.

[47] N. Lin, Foundations of Social Research, New York, 1976.

[48] W. Mangold, F. Miller, G. Brockway, Word-of-mouth communication in service marketplace, Journal of Service Marketing 13 (1) (1999) 73–88.

[49] K. B. Murray, A test of services marketing theory: Consumer information acquisition activites, Journal of Marketing 55 (1991) 10–25.

[50] M. E. J. Newman, Analysis of weighted networks, Physical Review E 70.

[51] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical Review Letters 86 (14) (2001) 3200–3203.

[52] R. C. Prim, Shortest connection networks and some generalisations, Bell System Technical Journal 36 (1957) 13891401.

[53] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (1966) 581–603.

[54] M. J. Shaw, C. Subrmaniam, G. W. Tan, M. E. Welge, Knowledge management and data mining for marketing, Decision Support Systems 31 (2001) 127–137.

[55] A. Sundararajan, Network seeding, in: Workshop on Information Systems and Economics, Evanston, IL, USA, 2006.

[56] C. Van den Bulte, Y. V. Joshi, New product diffusion with influentials and imitators, Marketing Science 26 (2007) 400–421.

[57] S. van Dongen, Graph clustering by flow simulation, Ph.D. thesis, University of Utrecht (2000).

[58] S. Wassermann, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.

[59] D. J. Watts, S. H. Strogatz, Collective dynamics of 'smallworld ' networks, Nature 393 (1998) 440–442.

[60] R. A. Westbrook, Product/consumption-based affective responses and post purchase processes, Journal of Marketing Research 24 (3) (1987) 258–270.

[61] I. H. Witten, E. Frank, Data Mining, Carl Hanser, Mnchen, Wien, 2001.

[62] F. Wu, B. A. Huberman, L. A. Adamic, J. Tyler, Information flow in social groups, Tech. rep., Physics Department, Stanford University HP Laboratories (2003).

Appendix. Evaluation of Centrality Measures for Different Diffusion Models

In the following, we provide more detailed results with respect to the different treatment variables in our diffusion models in Section 3.2.2.

Decay of signal strength τ

First we show in Figure 12 - 15 how changes in the decay factor τ impact the outcome by describing Model 1 and Model 2 of Table 2, i.e., if the decay parameter $\tau = 0.5$ changes to $\tau = 1$.

The difference between the two decay factors is that a high number of persons are reached very fast when the decay factor $\tau = 1$, i.e., the strength of the signal or message does not get weaker. In this case, there is little difference between the top-ranked centrality measures. If we assume a message to get weaker, however, the close vicinity of a node is of higher importance and SenderRank and out degree centrality outperform all other measures. The randomness in κ and the structure of the network lead to non-monotonous gain curves in Figure 15. We have observed the same pattern for the other networks.

Probability of message transmission κ

Figure 16 - 19 presents the differences of Model 9 and Model 10 of Table 2, with the transmission probability changing from $\kappa = 60\%$ to $\kappa = 100\%$.

Except for closeness centrality and the hubs score, we get the same ranking, although, as expected, with a higher κ also many more customers could be reached. Again, SenderRank and out degree were best if a certain minimum number of initial customers was addressed.

Threshold ε

The threshold ε describes the signal strength at which we consider a message received. In the previous experiments it was set to 0.3. Figure 20 - 23 show how the ranking of centrality measures when ε changed from 0.3 to 0.2, as in Model 4 and 5.

A noticeable difference is that the betweenness centrality did best when the threshold was only 0.2. We could, however, not observe this in the simulated networks or with other treatments, were always SenderRank

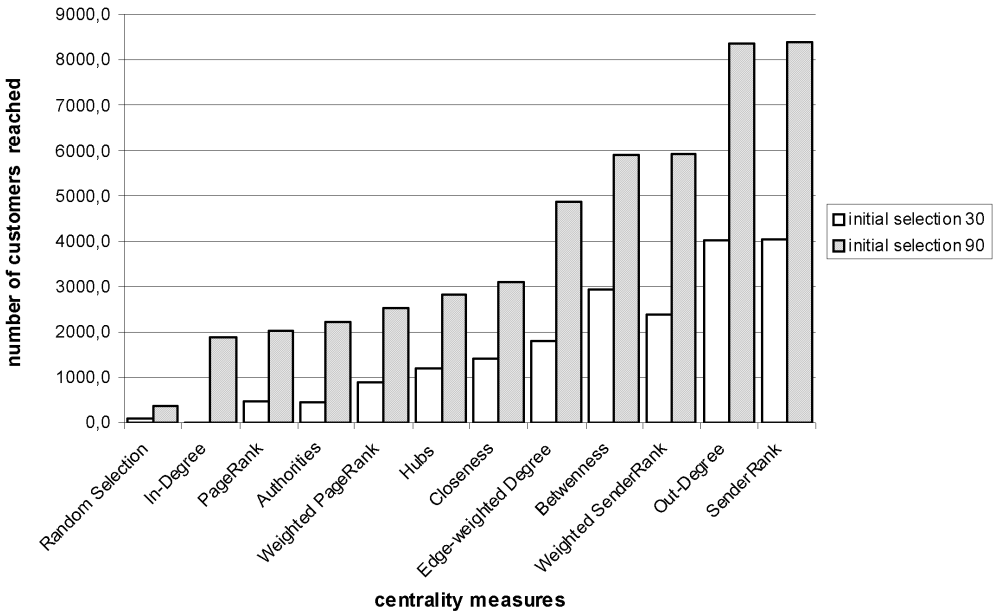

Figure 12 **Model 1, $\tau = 0.5$, histogram**

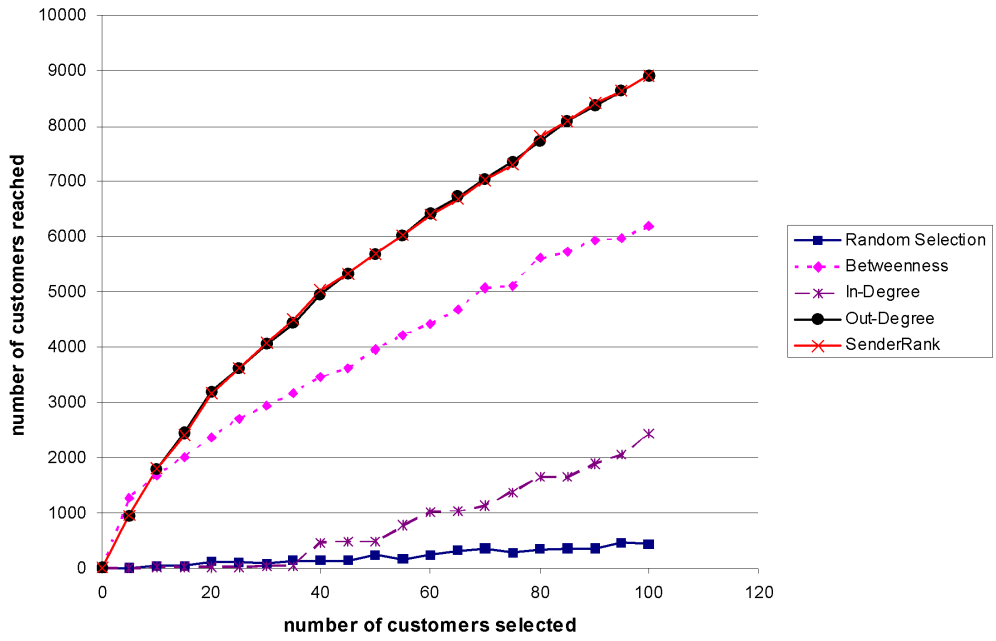

Figure 13 **Model 1, $\tau = 0.5$, gain-chart**

and out-degree centrality performed best. Overall, betweenness centrality did well when the threshold ϵ and the probability for transmission κ were set such that the message spreads widely. Different values of ω did not exhibit a significant impact on the ranking of the centralities.

Power Law Diffusion versus Exponential Diffusion

We could also find no significant differences in the ranking of centrality measures using different diffusion models. This can be seen, for example, in Figure 17 for a power law decay ($\tau = 1.75$, $\varepsilon = 0.3$, $\kappa = 60\%$, $\omega = 10$), and in Figure 13 with an exponential decay ($\tau = 0.5$, $\varepsilon = 0.3$, $\kappa = 60\%$, $\omega = 10$).

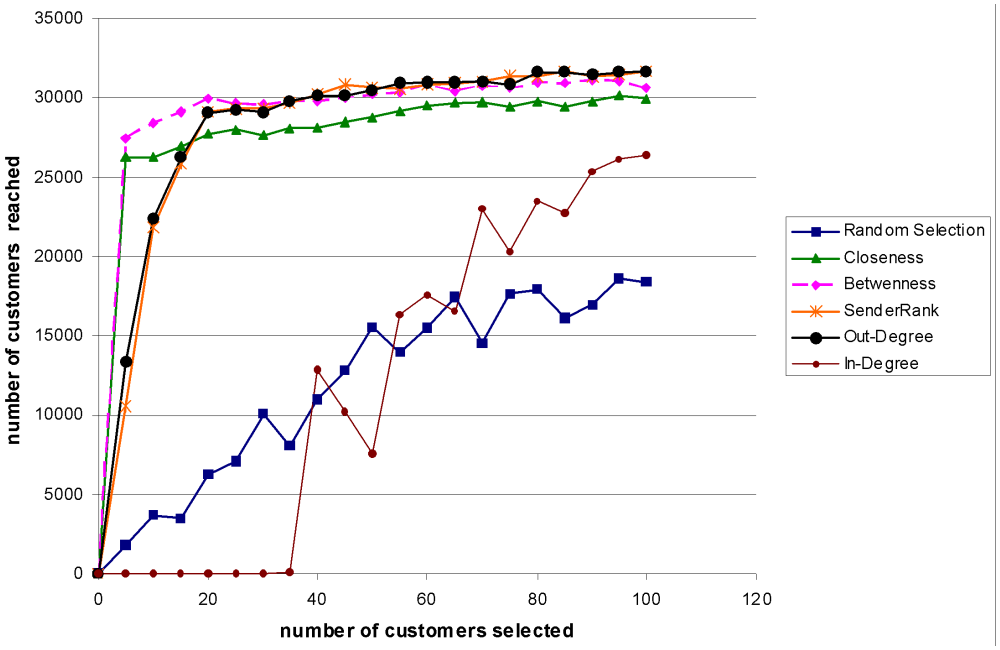

Figure 14 **Model 2, $\tau = 1$, histogram**

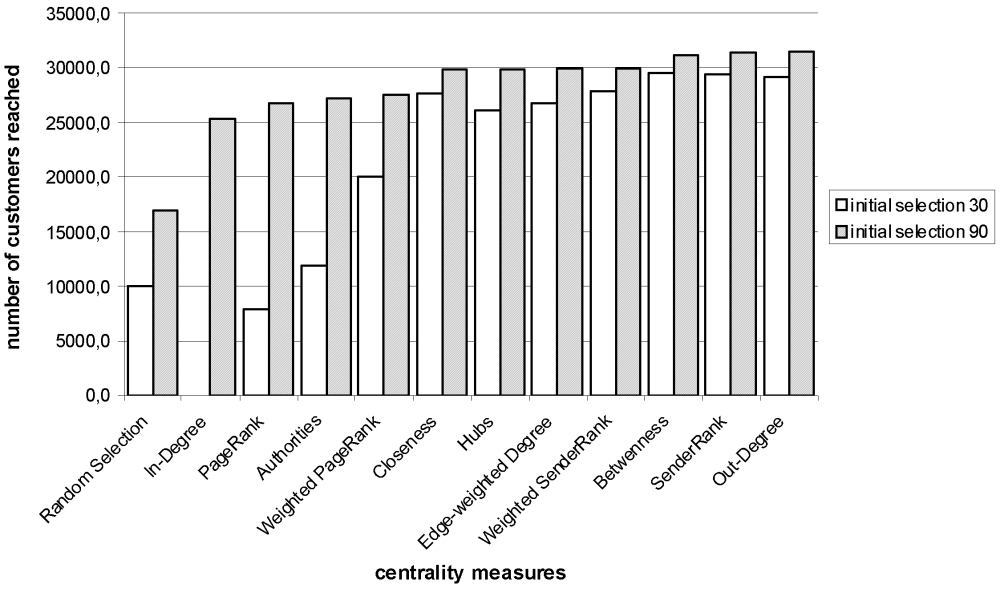

Figure 15 **Model 2, $\tau = 1$, gain-chart**

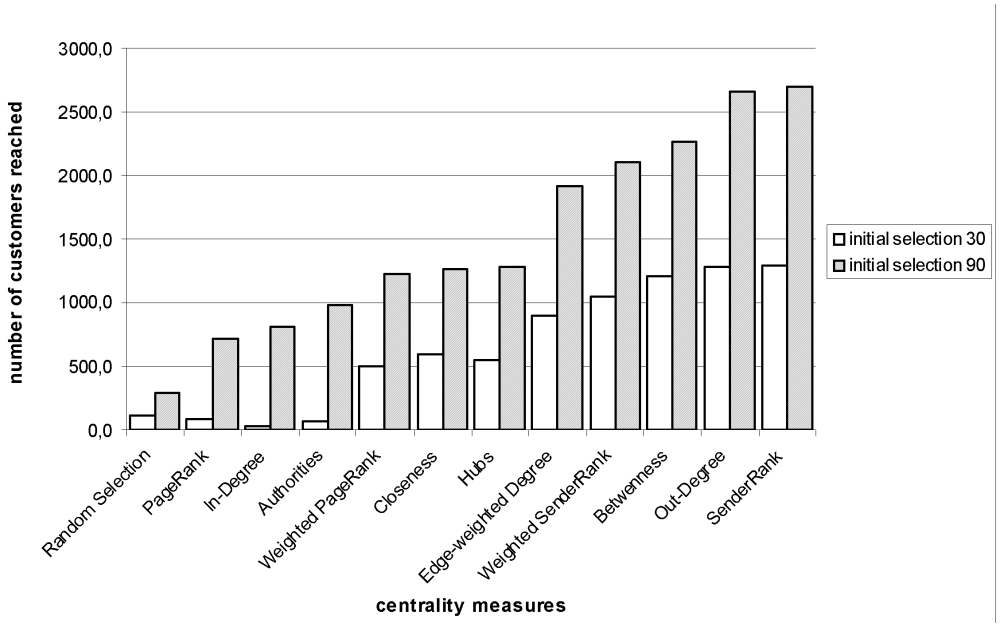

Figure 16 **Model 9, $\kappa = 60\%$, histogram**

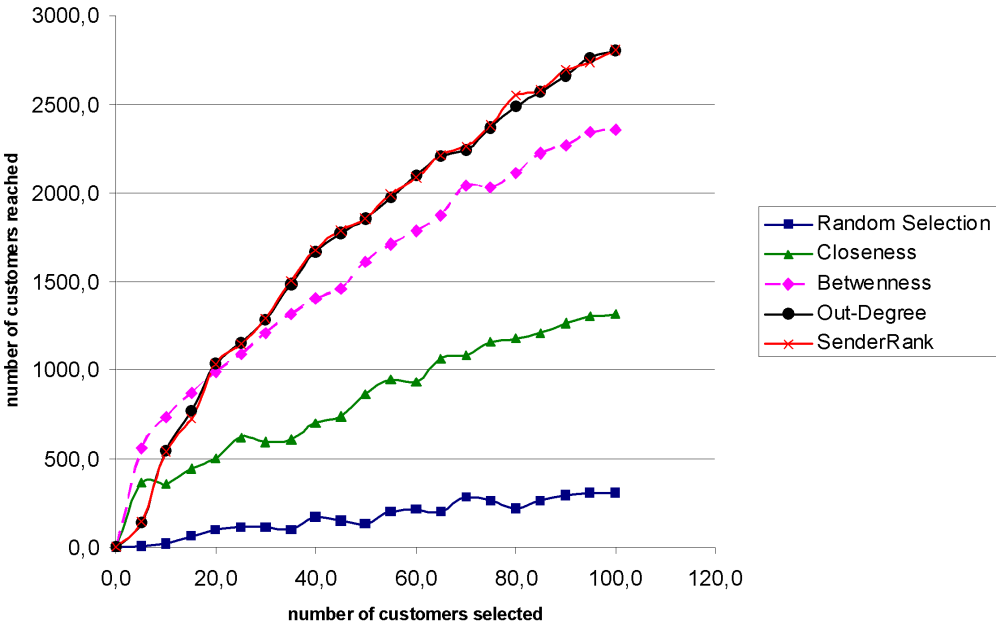

Figure 17 **Model 9, $\kappa = 60\%$, gain-chart**

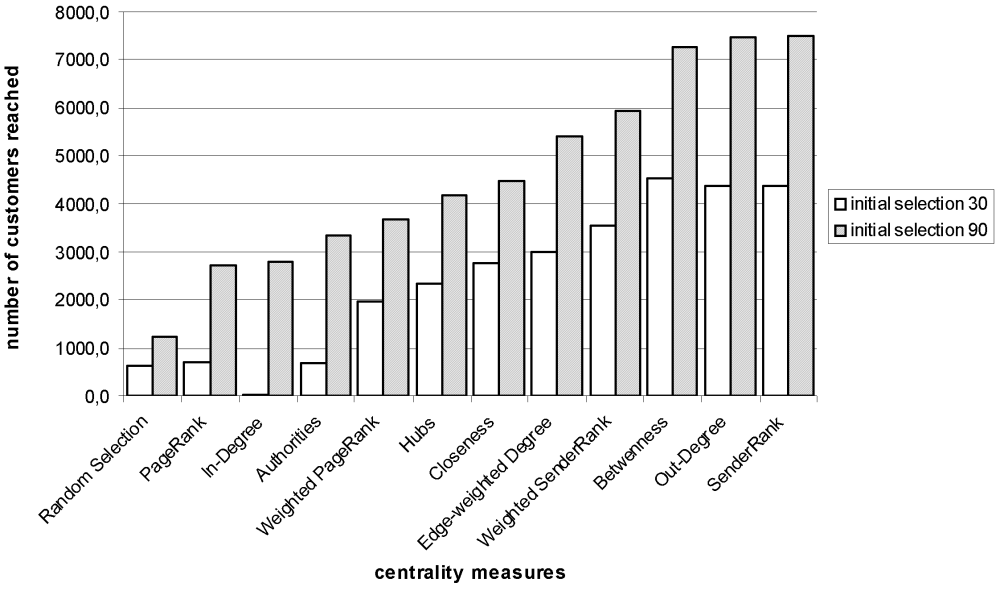

Figure 18 **Model 10, $\kappa = 100\%$, histogram**

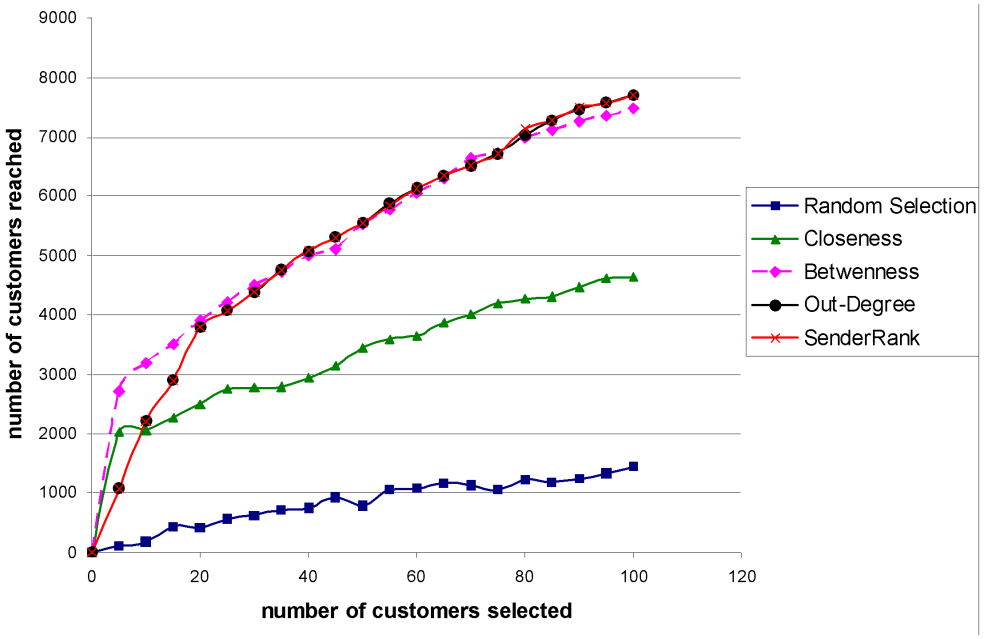

Figure 19 **Model 10, $\kappa = 100\%$, gain chart**

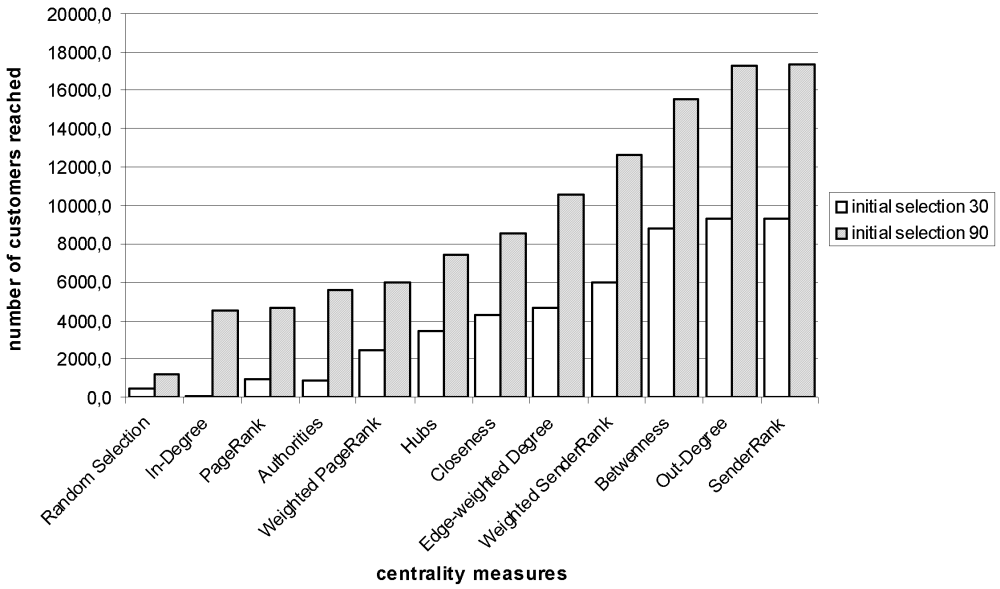

Figure 20 **Model 4, $\varepsilon = 0.3$, histogram**

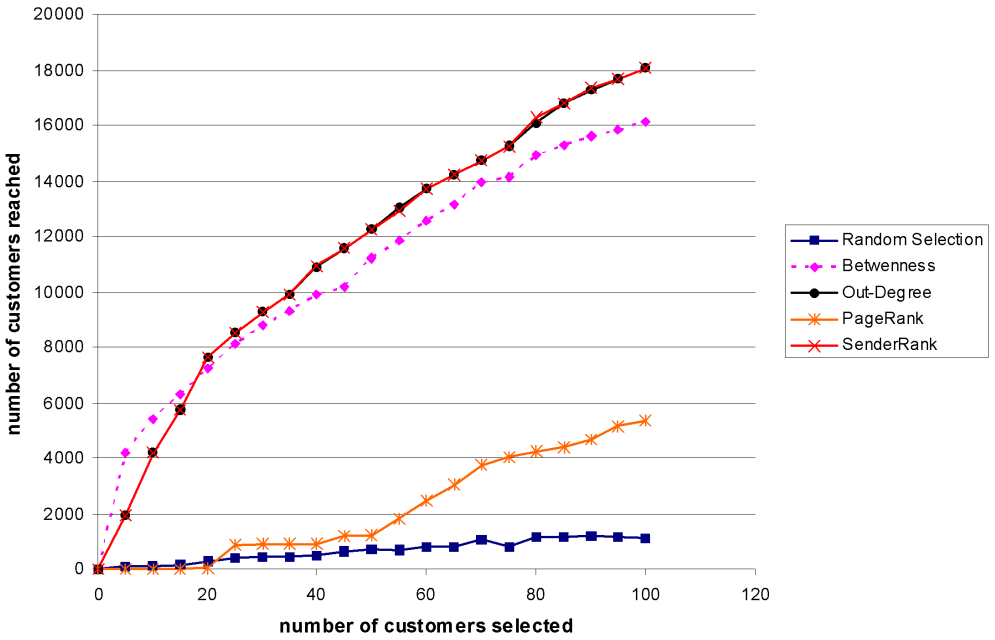

Figure 21 **Model 4, $\varepsilon = 0.3$, gain-chart**

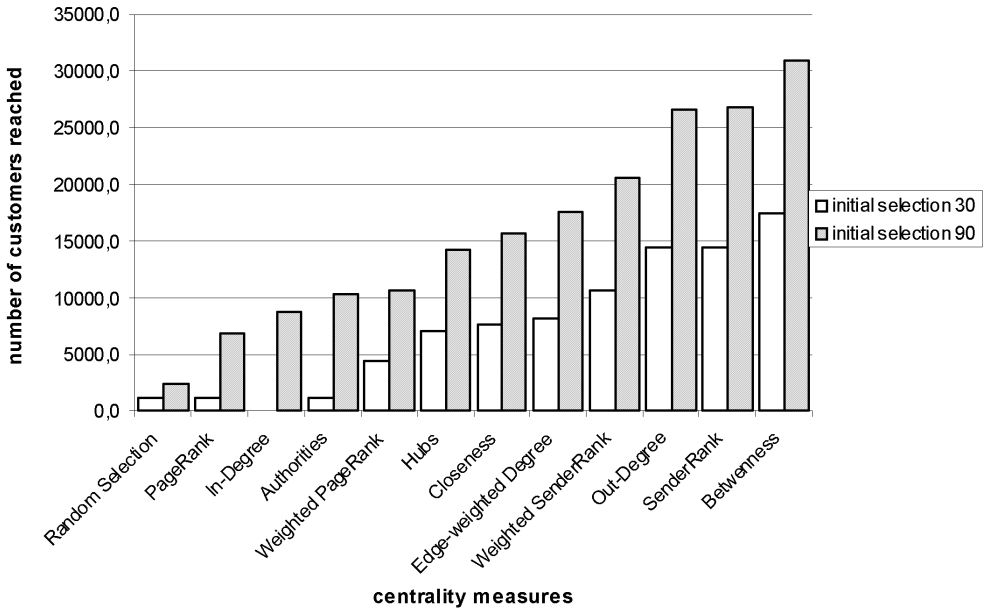

Figure 22 **Model 5, $\varepsilon = 0.2$, histogram**

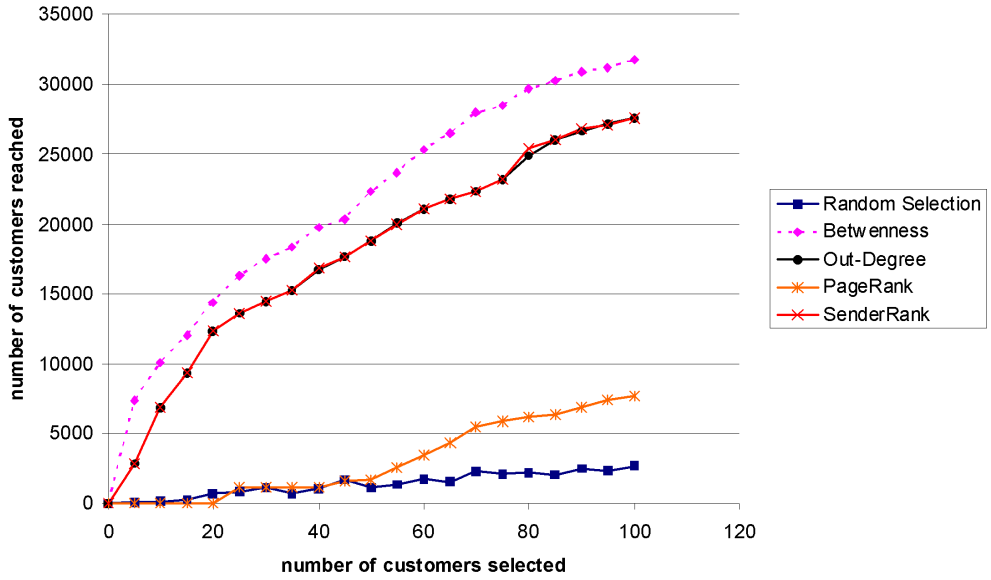

Figure 23 **Model 5, $\varepsilon = 0.2$, gain-chart**