THE EFFECT OF NETWORK CHARACTERISTICS ON COMPETITIVE PERFORMANCES OF CENTRALITY-BASED HEURISTICS IN INFLUENCE MAXIMIZATION

by

Cansu Özerim

B.S., Industrial Engineering, Istanbul Technical University, 2017

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfilment of
the requirements for the degree of
Master of Science

Graduate Program in Industrial Engineering

Boğaziçi University

2020

ACKNOWLEDGMENTS

Foremost, I would like to express my deep and endless gratitude to Assoc. Prof. Gönenç Yücel, my thesis supervisor, for his guidance, support and patience throughout this study. Besides his guidance, I am grateful to his encouragement that always motivated me to struggle against difficulties I have been faced during this study. I would also like to thank my committee members Prof. Kuban Altınel and Assist. Prof. Evren Güney for their precious time and valuable comments.

I wish to thank my dearest friend Elif Gürbüz for her unconditional support during my master study. Besides my master study period, it has always been good to know that since high school she has been there with her energy, enthusiasm and inspiration whenever I need. I am grateful for her existence in my life, and I know we will accomplish many more goals by supporting each other.

I wish to thank Baturalp Keskin specifically for his extraordinary patience, emotional support and encouragement during the last but never ending months of this study. His support has a crucial role on this accomplishment and in my life.

I would also like to thank Kaan Özdinçer and Hasan Eren Bekiloğlu for their friendship, support and helpful discussions on my study.

Finally, I would like to extend my deepest gratitude to my parents, Zühre and Mehmet Eray Özerim, and to my sisters Gökçen and Gülcan for their never ending support, understanding and affection.

ABSTRACT

THE EFFECT OF NETWORK CHARACTERISTICS ON COMPETITIVE PERFORMANCES OF CENTRALITY-BASED HEURISTICS IN INFLUENCE MAXIMIZATION

Online social networks enable people to disseminate a variety of information to masses easily, but under the budget and time constraints. To overcome these constraints, many seed selection approaches have been proposed in the literature to start dissemination from a small subset of people containing the most influential set of users in the network, named as Influence Maximization (IM) problem. The studies in IM have been conducted under the assumption that there is only one type of information spreading over a network. However, in the real world, there are more than one opposing information spreading over networks and performances of approaches are also affected by the competition between them, which is named as Competitive Influence Maximization (CIM) problem. More recently, it has been revealed that performances of approaches are also affected by the network topology. This thesis aims to investigate the direct effects of the network characteristics on performances of seed selection approaches in CIM. In this regard, a simulation-based study is conducted on 13 real-world network datasets by using an extension of Linear Threshold Model. The effects of the five network characteristics (average clustering coefficient, average degree, normalized average path length, normalized degree variance and density) are investigated. Furthermore, performances of the four most commonly used centrality-based heuristics (Betweenness Centrality, Degree Centrality, Closeness Centrality and Eigenvector Centrality) are compared in terms of their non-competitive and competitive performances. To interpret how the performances of the heuristics change in response to the network characteristics, regression tree method is used. According to our concrete findings, heuristics are sensitive to the existence of an opponent in the network and the type of the opponent as it is expected. Furthermore, the effects of the network characteristics differ from non-competitive to competitive environment. The findings emphasize the importance of an integrated perspective in the effects of network characteristics and Competitive Influence Maximization.

ÖZET

ETKİ ENBÜYÜKLEMESİNDE AĞ YAPISI ÖZELLİKLERİNİN MERKEZİYET TABANLI SEZGİSELLERİN REKABET PERFORMANSLARI ÜZERİNE ETKİSİ

Sosyal ağlar zaman ve bütçe kısıtları altında insanların çok çeşitli bilgileri kolaylıkla büyük kitlelere yaymalarını sağlar. Birçok araştırmacı bu kısıtlara karşılık çeşitli çekirdek küme seçim yaklaşımları önermişlerdir. Etki Enbüyüklemesi olarak adlandırılan bu problemde önerilen çekirdek küme seçim yaklaşımları yayılımı ağdaki en etkili kullanıcıları içeren küçük bir alt gruptan başlatmayı amaçlamaktadır. Ancak gerçek hayatta ağda yayılmakta olan ve birbiri ile rekabet eden birden fazla bilgi vardır ve çekirdek küme seçim yaklaşımları rakiplerin birbiri ile olan etkileşiminden de etkilenmektedir. Bu problem Rekabetçi Etki Enbüyüklemesi olarak adlandırılır. Ayrıca yakın zamanda önerilen yaklaşımların performanslarının ağın yapısal özelliklerinden de etkilendiği ortaya çıkarılmıştır. Bu bağlamda bu tezin amacı ağ yapısal özelliklerinin çekirdek küme seçim yaklaşımlarının Rekabetçi Etki Enbüyüklemesi' ndeki performansları üzerindeki doğrudan etkisini araştırmaktır. Bu amacı gerçekleştirmek için 13 gerçek ağ veriseti üzerinde Doğrusal Eşik Modeli' nin bir versiyonu kullanılarak benzetim tabanlı bir çalışma gerçekleştirilmiştir. Beş ana ağ yapısı özelliğinin (ortalama kümelenme katsayısı, ortalama derece, normalleştirilmiş ortalama yol uzunluğu, normalleştirilmiş derece varyansı, yoğunluk) etkisi araştırılmıştır. En yaygın şekilde kullanılan dört merkeziyet tabanlı sezgiselin (arasındalık merkezliği, derece merkezliği, yakınlık merkezliği, öz vektör merkezliği) rekabetçi olan ve olmayan Etki Enbüyüklemesi' ndeki performansları kıyaslanmıştır. Sezgisellerin performanslarının yapısal karakteristik değişimlerine karşı nasıl değiştiği regresyon ağacı metodu kullanılarak yorumlanmıştır. Somut bulgular beklendiği üzere sezgisellerin performanslarının ağdaki bir rakibin varlığına ve bu rakibin tipine karşı duyarlı olduğunu göstermiştir. Ayrıca ağ yapısal özelliklerinin sezgiseller üzerindeki etkilerinin rekabet olan ve olmayan etki enbüyüklemesi arasında farklılık gösterdiği gözlemlenmiştir. Bulgular ağ özelliklerinin etkilerinin ve Rekabetçi Etki Enbüyüklemesi' nin birlikte ele alınması gereken iki konu olduğunun önemini vurgulamaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

A	Adjacency matrix
a	Adjacency matrix component
d	Diameter of a network
E	Edge set
G	Number of shortest paths between two given nodes
I_A^t	A_active neighbors of a node step t
I_B^t	B_active neighbors of a node at step t
I^t	All neighbors of a node at step t
k	Average degree of a given node
n	Total number of nodes in a given network
s	Seed set size
t	Threshold of a given node
v	Node
w	Edge weight

LIST OF ABBREVIETIONS AND ACRONYMS

ACC	Average Clustering Coefficient
AD	Average Degree
BC	Betweenness Centrality
CART	Classification and Regression Trees
CC	Closeness Centrality
CELF	Cost-Effective Lazy Forward
CIM	Competitive Influence Maximization
D	Density
DC	Degree Centrality
EC	Eigenvector Centrality
IC	Independent Cascade
IM	Influence Maximization
LT	Linear Threshold
NDV	Normalized Degree Variance
PR	PageRank
R	Random

1. INTRODUCTION

Online social networks connect people across the world. According to Global Digital Overview Report (2020), more than 4.5 billion people around the world have been using the internet at the beginning of 2020. Portable devices such as mobile phones and tablets enable people to connect to social networks easily via the internet. Especially the group aged 16-64 among the 4.5 billion people spend 89% of the internet use time on various social networks such as Facebook and Twitter (Global Digital Overview Report, 2020). As a stunning example, there have been 152 million daily active users on Twitter and 1.66 billion daily active users on Facebook according to quarterly reports of the companies (Twitter Inc., 2020; Facebook Inc. 2020).

People use a wide variety of social networks such as online social networks, collaboration and email networks to disseminate innovations, products, fashions, rumors and advertisements; to share emails, blog links and status updates (Harrigan *et al.*, 2012). To be more precise, a variety of new information is introduced on a social network every day, e.g., when a new model of mobile phone is introduced to market, a party declares some political statements, or rumors about famous people are reported. Individuals or organizations aim to disseminate information to as many people as possible. However, some of information is adopted by the masses while some of information is adopted by only a certain group of people. Therefore, understanding the dynamics of information diffusion over social networks, which can be affected by various factors such as social interactions between people, draws great interest of researchers. On the one hand, exploring the mechanism behind information diffusion helps the companies or politicians to disseminate information originated by them efficiently and effectively. Besides, there are more than one competing information spreading over the networks, thus the studies of information diffusion help to prevent undesirable information dissemination such as rumors, unconvincing news and even the advertisement of the rival company (Jiang *et al.*, 2014).

Although social networks provide ease of information diffusion, there are budget and time constraints to propagate an information over a network efficiently. The most critical point for the information diffusion is to start propagation from a small subset of people

containing the most influential set of users in the network. For example, there is evidence showing that a group of people change from a commonly adopted brand to another brand because their friends in a social network have chosen it (Sumith *et al.,* 2018). In addition to this, there are examples where people are strictly loyal to some brands because of the same reason (Sumith *et al.*, 2018). Finding the most influential set of users on the network is named as Influence Maximization (IM) problem (Kempe *et al.,* 2003). However, there are more than one type of information spreading over social networks simultaneously in the real world. Thus, IM problem is extended to a competitive version named as Competitive Influence Maximization (CIM) (Bharathi *et al.,* 2007), and some researchers have taken into consideration this reality in the couple of decades (e.g. Carnes *et al.,* 2007; Budak *et al.*, 2011; Wu *et al.,* 2015; Liu *et al.,* 2016; Zhao *et al.,* 2017). Besides, more recently, some researchers have explored the effects of network topology on the information diffusion and the performances of seeding approaches. Seeding approaches show varying performances on different types of networks in terms of the number of adopters at the end of the diffusion process (Peres, 2014).

This thesis contributes to the literature on Competitive Influence Maximization by conducting a simulation-based study on the 13 real-world network datasets to investigate the competitive performances of centrality-based heuristics in influence maximization under different network topologies. In the literature, there are numerous studies conducted on the information diffusion over social networks. In the next section, where I review the literature, the most representative studies relevant to this thesis are summarized under three groups: (i) Influence Maximization, (ii) Competitive Influence Maximization, (iii) Network Topology Effect on Information Diffusion & the Performances of Seeding Approaches.

2. LITERATURE REVIEW

Influence Maximization problem is firstly defined by Kempe *et al.* (2003). Influence Maximization is the problem of finding a small subset of nodes in a network, so as to maximize the final spread of information, subject to the case that there is only one type of information spreading over the network. Later, many researchers have focused on finding better solutions for Influence Maximization under two categories: (i) algorithmic heuristics with approximation guarantees while increasing the computation time (ii) metric-based heuristics which do not guarantee or give weaker guarantee while reducing the computation time.

2.1. Influence Maximization

The first category deals with Influence Maximization Problem is approximation guaranteed heuristics such as *Greedy Algorithm, Cost-effective Lazy Forward Optimization (CELF)* and *CELF++*. Firstly, Kempe *et al.* (2003) showed the optimal seed selection is a NP-Hard problem. They developed a greedy algorithm achieving (1-1/e) approximation by focusing on two fundamental diffusion models named Linear Threshold Model and Independent Cascade Model (see Section 4.4 for detailed information about the diffusion models).

Since the *Greedy Algorithm* is an expensive algorithm in terms of time, Leskovec *et al.* (2007) focused on an efficient algorithm named *CELF,* which is 700 times faster than the simple greedy algorithm. The idea behind *CELF* is that the marginal gain provided by a node in the previous iteration cannot be worse than the marginal gain provided by the node in the current iteration.

Later, Goyal *et al.* (2011) have proposed *CELF++* with an improvement on *CELF*. The idea behind *CELF++* is to avoid repeated calculation of marginal gain for an already selected node. They proved that *CELF++* has provided an improvement of 35-55% over *CELF* in terms of computational time efficiency.

Another improvement on the *Greedy Algorithm* is made by Kimura *et al.* (2010). The authors improved the computation time by using bond percolations and graph theory approximations. They tested the algorithm on both Linear Threshold and Independent Cascade Model, and then it is reported that their algorithm is 4600 times faster than the simple greedy algorithm, but it has limitations due to the need of extensive user profile information and interaction data.

Later, Jiang *et al.* (2011) have come up with a totally new approach using *simulated annealing*. They tested the proposed algorithm on four real-world networks. The results indicate that the proposed algorithm outperforms the simple greedy algorithm up to 2-3 orders of magnitude in terms of computation time. Besides, it improves the accuracy of simple greedy algorithm by around an average of 5%.

Estevez *et al.* (2007) proposed *set covering greedy algorithm* aiming to ensure that neighbors of selected seeds do not overlap, which means that selected seeds can influence a higher number of nodes. The authors report that set covering algorithm requires less time, and it is also superior to the simple greedy algorithm in terms of information diffusion.

Cheng *et al.* (2013) proposed an algorithm, which is grounded in the simple greedy algorithm. They focused on both of the computation time problem of the simple greedy algorithm and accuracy problem of heuristics at the same time. Their results showed that the proposed algorithm requires considerably less time than the simple greedy algorithm without sacrificing accuracy.

To deal with computation time limitations of greedy algorithm, a *random reverse reachable set* algorithm is introduced to IM by Borg *et al*. (2014). A reverse reachable set for a random node *v* is generated by first sampling a network *G(V,E)* from the distribution generated by removing each edge according to its propagation probability, and then taking the set of nodes in *G(V,E)* that can reach *v*. A random reverse reachable set is simply a random reverse set for a node selected uniformly at random. Random reverse reachable algorithm achieves reducing computation time to almost linear with the size of the network under Independent Cascade Model. Later, Tang *et al.* (2014) proposed *TIM/TIM+ (Two-phase Influence Maximization),* and showed that random reverse reachable algorithm can

applied to Linear Threshold Model too. Later, Nguyen *et al.* (2016) presented *SSA (Stop-and-Stare Algorithm)* and *D-SSA (Dynamic Stop-and-Stare Algorithm),* and Tang *et al.* (2018) introduced *OPIM-C (online processing of influence maximization-conventional),* which are proposed to reduce the number of random reverse reachable sets to achieve a more effective algorithm in terms of computation time without a sacrificing approximation guarantee. However, all these improved algorithms based on reverse reachable set have scalability issues in large scale networks.

Kitsak *et al.* (2010) proposed the *k-shell decomposition analysis* building on the heuristic degree centrality. In particular, the underlying mechanism behind k-shell decomposition is identifying hierarchies in a network. The method begins with grouping the nodes have only one neighbor in a network under 1-shell. This procedure continues as assignment of the nodes with k neighbors to k-shell group until there is no node to be grouped. Their results showed that the most influential nodes in a network are under the shell groups with higher k values.

Since there are limitations of the algorithmic heuristics above in terms of computation time and accuracy, Zhang *et al.* (2013) proposed a new method named as *k-medoid.* This method uses a partitioning technique which clusters a set of n nodes into k clusters, with the number of k of clusters assumed known a priori and the dissemination probability between all pairs of nodes. This method is advantageous for the networks with a community structure since it selects nodes from different communities in a network.

The solution approaches so far have limitations in terms of computation time. Therefore, there is another direction in the literature, which improve the computation time while sacrificing the approximation guarantee.

This second category is heuristics using different metrics about nodes in a network. The most frequently used metrics are centrality-based metrics, which are *betweenness centrality, degree centrality, closeness centrality* and *eigenvector centrality.* These metrics are used to detect the influential nodes in terms of locating in a critical position in the network or connecting nodes from different clusters. Selecting seeds among the nodes with the highest values of these metrics are named as metric-based seed selection heuristics.

Besides, specifically, the heuristics using centrality metrics to select seeds are named as centrality-based heuristics.

Firstly, the *betweenness centrality* is simply the ratio of the shortest paths that pass through the node of interest among all the shortest paths in a network. The betweenness centrality of a node *v* is defined as

$$BC(v) = \sum_{v \neq t} \sum_{v \neq u} \sum_{t \neq u} \frac{g_{ut}^{v}}{g_{ut}} \tag{2.1}$$

where g^{v}_{ut} is the number of shortest paths among the nodes *u* and *t* and g^{v}_{ut} is the number of shortest paths which pass through node *v* (Ghalmena *et al.*, 2018).

The *degree centrality* is simply the metric of number of connections a node has. The degree centrality of a node *v* is defined as

$$DC(v) = \frac{k_v}{n-1} \tag{2.2}$$

where n is the number of nodes on the network and n-1 is the maximum possible degree (Lü *et al.*, 2016).

The *closeness centrality* represents how close a node is to all other nodes in a network. The closeness centrality of a node *v* is defined as

$$CC(v) = \sum_{v \neq u} \frac{1}{(n-1)d_{vu}} \tag{2.3}$$

where d_{vu} is the geodesic distance from node *v* node *u*, which means that the number of edges in the shortest path between node *u* and *v* (Cohen *et al.*, 2014).

The *eigenvector centrality* indicates how well a node is connected to a network by considering the connectedness of its neighbors. The eigenvector centrality of a node *v* is defined as

$$EC(v) = \frac{1}{\lambda}\sum_{u} a_{vu} EC(u) \tag{2.4}$$

where a_{vu} is the value from the adjacency matrix of the network and it is equal to 1 if *v* and *u* are connected and 0 if not. λ is a predefined constant (Bonacich, 1987).

As an extension of eigenvector centrality, *PageRank* metric is proposed to measure the time spent on a node in a given network. When at a node, there is an equal chance for all the neighbors to be visited at the next step. However, there is an additional 15% chance of a random step, going to a node not necessarily linked to the current one. The *PageRank* is defined as

$$PR(v) = \frac{1-a}{\mathbf{n}} + a \sum_{u \in N(v)} \frac{PR(u)}{k_u} \tag{2.5}$$

where n is the number of nodes in the network, k_u is the degree of node *u*, *N(v)* is the neighborhood of node *v*. The *PageRank* is proposed with the motivation of distinguishing the importance of different websites in Google search engine by considering the quantity and the quality of the pages linked to the website of interest (Lü *et al.*, 2016).

Apart from the heuristics that use centrality-based metrics mentioned above, there have also been other heuristics which use different metrics and extended from centrality-based heuristics. One of these heuristics is *degree discount heuristic* which is similar to the degree centrality heuristic. The difference between these two heuristics is calculating the degree of the node of interest by excluding its already activated neighbors (Chen *et al.*, 2009). Another extended version of degree centrality, named *twostep,* is proposed by Stonedahl *et al.* (2010). Unlike degree centrality, two-step considers also the degrees of neighbors of neighbors for the node of interest.

As it can be concluded from the references above, a variety of seed selection heuristics have been proposed for the problem of IM. All these studies discussed above have

been conducted under the assumption that there is only one type of information spreading over a network, i.e. there is no competition. In the next section, the literature of the comparison and extension of these heuristics in Competitive Influence Maximization problem is reviewed. The studies for Competitive Influence Maximization are conducted under the reality that there are more than one type of information spreading over social networks.

2.2. Competitive Influence Maximization

The Influence Maximization problem has been extended to a competitive situation where there are two or more competitors aiming to maximize its influence in a network (Zhuang *et al.*, 2017). Firstly, Bharathi *et al.* (2007) proposed CIM problem, which tries to find a seed set to maximize the influence spread of their products' information while the competitor carries out the same strategy.

Carnes *et al.* (2007) addressed the problem of competition between two companies aiming to propagate their new products via viral marketing while one of the products is already being introduced. The authors assumed that the nodes can adopt only one of the products due to the limited budget. They proposed two diffusion models inspired by the Independent Cascade Model. The first one named A Distance-Based Model has been created under the assumption that a node will be more likely to adopt the product of an early adopter if the locations of these nodes are close enough in the network. In other words, the probability that node v adopts a product is assigned depending on the sets of paths from all seed sets to node v. In the second model, the same probabilities have been given to the nodes to adopt one of the two competing products regardless of the distance between the nodes. They conducted simulation experiments on one real-world collaboration network named HepTh to assess the effectiveness of selecting seeds based on maximum degree and low average distance while the initial seed set of the second competitor is known. The results of the experiments indicated that selecting seeds based on the maximum degree metric outperforms low average distance metric with both of the diffusion models.

Budak *et al.* (2011) addressed the problem of finding of a small subset of nodes that need to be persuaded to adopt the intended product so as to minimize the number of nodes

that adopt the undesired product. The authors experimentally compared the performances of greedy algorithm to various heuristics under a competitive version of Independent Cascade Model. They showed that simple centrality-based heuristics such as *degree centrality* are comparable to the *greedy algorithm*.

Later, Wu *et al.* (2015) extended the Linear Threshold Model to a competitive situation. They addressed the objective function as selecting an optimal seed set under the assumption that the seed set of the competitor has been known. The authors extracted 1000 nodes from a real-world network dataset and generated one new synthetic network. Then, they tested the functionality of the *greedy algorithm* on these datasets. Their results showed that the greedy algorithm can approximate the optimal result with 1/1-e.

Liu *et al.* (2016) proposed an extension of Linear Threshold Model for competitive information diffusion and named the model as the Diffusion-Containment model. The idea behind the model is that there is an information disseminating in the network to prevent the spread of other type of information, which are named as C-influence and D-influence respectively. The authors created a payoff strategy for the diffusion: If node *u* and node *v* adopt the opposite behavior, both of them get a payoff zero. If node *u* and node *v* adopt the same behavior (D-influence or C-influence), both of them get a payoff higher than zero. Then, they tested the effectiveness of *greedy algorithm*, *maximum degree* (choosing the seeds from the nodes with highest degree) and *Random* on two synthetic scale-free networks and one real-world network named Wiki-Vote. Their results indicate that the *greedy algorithm* performs better than *maximum degree* and *Random* heuristics to prevent D-influence under Diffusion-Containment model, but it is not appropriate for larger scale networks.

Zhao *et al.* (2017) approached the competitive influence maximization problem differently compared to all the literature above. The other studies are conducted under the same assumption that the initial seed set of the second competitor is known. In the study of Zhao *et al.* (2017), different centrality-based seed selection heuristics are chosen by the competitors at the beginning of the simulation, and the competitors select the initial seeds at the same size depending on them. A competitive diffusion model inspired by game theory perspective is used to simulate diffusion process. In this model, an inactive node becomes

A-active or B-active at step t+1, if it has only one type active neighbors at step t, i.e. A or B. If an inactive node has no active neighbor, it remains inactive until the next time step, and if it has both type of A and B active neighbors, it becomes resistant to the diffusion process, which means it cannot become A or B active until the end of the simulation. The authors conducted simulation experiments on the six real-world network datasets ranging from 379 to 10680 nodes to compare the effectiveness of centrality-based heuristics, which are *betweenness, degree, closeness, eigenvector* and *k-shell*. The results of the experiments showed that there is a general ranking on all of the datasets such as *betweenness* and *degree centrality* outperform the *closeness centrality*, besides *betweenness centrality* is more appropriate strategy against *closeness centrality*. In addition to this, they stated that *eigenvector* and *k-shell* centrality heuristics show always poorer performance than the other three heuristics.

The major studies in the Influence Maximization and the Competitive Influence Maximization fields are summarized above. As it is stated before, finding a small subset of nodes consisting of the most influential set of nodes in the network is the key point to reach higher information disseminations ultimately. The heuristics proposed for this purpose are summarized in Section 2.1. Later, the studies considering there is a competitive environment during the information diffusion are summarized. In these studies, performances of the seed selection heuristics are compared. Apart from the competition, the recent studies have revealed that ultimate information dissemination is affected by the network types and characteristics (e.g. Barthélemy *et al.*, 2004; Hussain *et al.*, 2013; Peres, 2014; Liu and Hong, 2018). Hence, in the next section, the studies exploring the effects of network characteristics on information diffusion processes are summarized.

2.3. Effect of Network Topology on Information Diffusion and the Performance of Seeding Approaches

Classical works in information diffusion and influence maximization have been conducted under the same assumption that seed selection algorithms and heuristics perform similarly on all types of the networks and the behavior of the nodes are identical under different network topologies (Peres, 2014). More recently, some researchers have begun to

investigate the relation between diffusion processes and the types of the networks, and the effect of network topology on the performances of seeding strategies.

Barthelemy *et al.* (2004) investigated the impact of the heterogeneity level of connectivity in large scale networks on the dissemination of epidemics. The authors associated the heterogeneity level of connectivity with the degree distributions of the networks, i.e. the normalized degree variances of the networks. As the degree variance of the nodes in a network increases, this network can be considered as more heterogeneous. In their study, the hypothesis that an infection spreads to more people in scale-free networks is tested by using the Susceptible Infected diffusion model on synthetic networks. The results of the study indicated that if the nodes with highest degrees are infected at the beginning of the epidemic, it can transmit to much more nodes in the network, and the number of these infected nodes becomes higher in the networks with higher degree variance. The authors also stated that this result can be associated with other diffusion domains such as innovation and information diffusion.

Hussain *et al.* (2013) evaluated different behaviors of the four centrality-based heuristics on four different types of networks, which are small-world, scale-free, hybrid networks of small-world & scale-free and random networks. The *degree*, *betweenness*, *closeness* and e*igenvector centrality* have been taken into consideration by the authors since they have outperformed many approaches according to the literature. Both Independent Cascade Model and Linear Threshold Model are used in the study. They conducted simulation experiments on eight networks, four of which are synthetic and four of which are real-world datasets. The datasets are classified into one of the four groups according to their average clustering coefficient, degree distribution and average path lengths. The results of the study demonstrated that all of the heuristics have performed similarly on the small-world and random networks. On the other hand, they stated that all heuristics performed better on the scale-free and hybrid networks of small-world and scale-free, but with the superiority of betweenness centrality over other heuristics. The degree centrality followed the performance of betweenness centrality. The closeness and eigenvector centrality showed better performances on Twitter networks and scale-free networks respectively, but their performances were behind betweenness and degree centrality.

Peres (2014) studied how the network characteristics affect the diffusion of a new product in a network by focusing on the effects of average degree, relative degree of social hubs (the ratio between the overall average degree of the network and the average degree of the most connected nodes in the network) and clustering coefficient instead of the effect of different network types on innovation diffusion. The author evaluated two cases where there is no competition in the network and there are two competing companies in the network by using an agent-based model. The simulation experiments have been conducted on the synthetic and real-world networks, which differ in terms of network characteristics taken into consideration in the research. The results of the study concluded that the average degree and the relative degree of social hubs affect the information diffusion positively while clustering coefficient has a negative impact.

Liu and Hong (2018) examined the effect of network topological characteristics on the performance of some centrality-based heuristics. The authors used the degree and k-shell centrality measures as the sequential seeding strategies, which means that seeds are activated at the subsequent time steps in a diffusion process. They conducted simulation experiments on seven different real-world network datasets which differ in terms of their density, degree distribution and assortativity coefficients. They used the classical Independent Cascade Model as the diffusion model. Their results demonstrated that both of the centrality-based heuristics perform better on the heterogeneous networks in terms of the degrees of the nodes. In addition to this, they stated that the average degree and the density of the networks are correlated and the final information dissemination increases as the density increases for both of the centrality measures. Besides, they concluded that the coverage of the two centrality measures decrease as the assortativity coefficient of the network increases, which means that the information dissemination is negatively affected when the nodes with small degree (high degree) tend to connect to the nodes with small degree (high degree).

To sum up, existing literature on the effects of network topology involve a limited perspective. Barthelemy *et al.* (2004) considered only the effect of degree variance of the networks on the information diffusion. Hussain *et al.* (2013) investigated the effects of different network types on the standalone performances (performance in a non-competitive environment) of some centrality-based heuristics by ignoring the fact that there are more than one type of information spreading over the networks in the real world. Similarly, Liu

and Hong (2018) worked in a non-competitive environment, and they investigated the direct effect of a few network characteristics on the performance of some centrality-based heuristics. Only Peres (2014) investigated the direct impact of network characteristics on information diffusion in both competitive and non-competitive environment, but the author assumed that the seeds of the competitors are selected based on the same seed selection approach.

Consequently, there is a gap in the literature in terms of investigating the direct impact of the network characteristics on the performances of different seed selection heuristics in a competitive environment. In other words, to our knowledge, there has been no study focusing on whether the network characteristics have an effect on the Competitive Influence Maximization problem or not, and if they have what is the impact of them.

3. PROBLEM STATEMENT AND OBJECTIVES

Many seeding approaches have been proposed and compared in the literature in the context of both Influence Maximization and Competitive Influence Maximization problems. More recently, the literature has highlighted that performance of a seed selection approach (algorithmic and metric-based heuristics) is affected by the network characteristics (Peres, 2014), and it is also affected by the interaction with the seeding approach of the opponent in the network (Zhuang *et al.*, 2017). Thus, in order to provide efficient and wider information disseminations with the seeding approaches, appropriate seeding approach should be selected according to both the network characteristics and to the opponent's seeding approach.

Although the impact of the network topological characteristics on social influences processes have been studied in a few studies (e.g. Barthélemy *et al.*, 2004; Hussain *et al.*, 2013; Peres, 2014; Liu and Hong, 2018), the direct impact of the network topological characteristics on the competitive performances of seed selection approaches has not been studied thoroughly in the network analysis field yet. In this regard, the primary objective of this thesis is to contribute to this gap in the literature by investigating the direct effects of the network characteristics on competitive performances of the seed selection approaches in influence maximization. Besides, the following secondary objectives have been set in order to achieve the primary objective:

- Compare the standalone performances of the seed selection approaches on various real-world networks, which means that the performances of the approaches in a non-competitive environment.
- Evaluate the general ranking of the approaches in terms of their standalone performances
- Examine the effects of network topological characteristics on the standalone performances of the approaches.
- Compare the competitive performances of the approaches on various real-world networks, which means that the performances of the seed selection approaches in a competitive environment.

- Evaluate whether there is a difference between the standalone performance ranking and the competitive performance ranking or not.
- Examine the effects of network characteristics on the competitive performances of approaches.
- Evaluate the similarities and differences between the effects of network characteristics on the standalone performances and on the competitive performances of the seeding approaches.

In this regard, a simulation-based study is conducted. The study covers the most commonly used seed metrics for seed selection, which are *betweenness centrality, degree centrality*, *closeness* centrality and *eigenvector centrality.* Besides, the study includes *average clustering coefficient, average degree, normalized average path length, normalized degree variance* and *density* in its scope as the network characteristics. The information diffusion procedure is designed as an extension of Linear Threshold Model, and then the experiments are conducted on 13 real-world network datasets, which differ in terms of their characteristics.

The rest of the thesis is organized as follows: Section IV describes the methodology of the study with the related literature. In Section V, the experimentation and analysis procedure is presented. In Section VI, the results are summarized and interpreted. Section VII discusses the results by combining both standalone and competitive performance results of the heuristics. Finally, Section VIII contains a conclusion and suggestions for future directions.

4. METHODOLOGY

This section presents the design of the study with the related literature. Detailed information about considered network characteristics, proposed centrality-based seed selection heuristics, how the network datasets are selected, what are the features of these networks, and the properties of utilized information diffusion model is given in the corresponding sub-sections.

4.1. Network Topological Characteristics

The primary aim of this thesis is to study the direct impact of network characteristics on the competitive performances of centrality-based heuristics. In this regard, five topological characteristics have been taken into consideration: average clustering coefficient (ACC), average degree (AD), normalized average path length (NAPL), density (D) and normalized degree variance (NDV). There have been various network characteristics which are investigated in the social network analysis field. These five metrics are considered in this study since they are capable of measuring global features of networks and they are most commonly used ones in the literature (Peres, 2014; Boccaletti *et al.*, 2006). Detailed information and mathematical definitions of these five topological characteristics are given below:

4.1.1. Average Clustering Coefficient

Average clustering coefficient is the overall mean of the local clustering coefficient of all nodes in the network. Before the average clustering coefficient, local clustering coefficient is firstly proposed by Watts and Strogatz (1998) to evaluate whether a network has small-world property or not. The local clustering coefficient of a node *v* is defined as:

$$CC(v) = \frac{2L(v)}{k_v(k_v - 1)} \tag{4.1}$$

where *L(v)* is the number of links between the neighbors of node *v* and k_v is the degree of node *v*. Thus, average clustering coefficient of a network *N* is defined as:

$$ACC(N) = \frac{1}{n}\sum_{v=1}^{n} CC(v) \tag{4.2}$$

where n is the total number of nodes in the network (Kemper, 2009).

Average clustering coefficient indicates the measure of a graph's transitivity which means the possibility that if nodes *v* and *i* are linked to each other, and nodes *i* and *u* are linked to each other, then nodes *v* and *u* are also linked. It has been observed that most of the complex networks have a tendency to cluster, within this context their average clustering coefficient are much greater than zero, despite the fact that they are still significantly less than one (Wang and Chen, 2003).

4.1.2. Average Degree

One of the simplest and the most significant characteristics of a node is its degree. The degree k_v of a node *v* is defined as the total number of its connections. Therefore, the greater the degree, the "more significant" the node is in a network. The average degree of a network is the mean of the degrees of all the nodes in the network, i.e. the average number of edges per node in the network, and it is also an indicator of the network's level of connectivity. Mathematical calculation of the average degree of a network is relatively straightforward as follows:

$$AD(N) = \frac{\sum_{v=1}^{n} k_v}{n} \tag{4.3}$$

where *n* is the total number of nodes in the network (Wang and Chen, 2003; Peres, 2014).

4.1.3. Normalized Average Path Length

Average path length *L(N)* of the network is defined as the average shortest distance between all pair of its vertices. This network characteristic indicates how many steps are required on the average to reach a node from another node and it is one of the most important metrics to classify networks (Chen *et al.*, 2008). However, since average path length is a distance-based metric, it cannot be a basis for comparison between different networks with different number of edges and nodes. Therefore, average path length should be normalized. Normalized average path length is one of the basic network characteristics and its computation is crucial in the analysis of complex networks (Chechik *et al.*, 2014). In the literature, diameter, which is the largest distance among all distances between any pair of nodes in the network, is used to normalize the average path length. Thus, normalized average path length *NAPL(N)* of a network *N* is defined as:

$$NAPL(N) = \frac{l(N) - 1}{d(N) - 1} \tag{4.4}$$

where *l(N)* is the average path length and *d(N)* is the diameter of the network *N* (Cancho *et al.*, 2004).

4.1.4. Density

The density *D(N)* of a network is defined as the ratio of the number of edges to the number of possible edges in a network with *n* nodes. For an undirected graph *G= (V, E)* with vertex set of size *n* and edge set *E*, graph density is defined as:

$$D(N) = \frac{2|E|}{n(n-1)} \tag{4.5}$$

For a directed graph *G= (V, E)* with vertex set of size *n* and edge set *E*, graph density is defined as:

$$D(N) = \frac{|E|}{n(n-1)} \tag{4.6}$$

Therefore, the maximum density of a network can be one which is reachable only for complete graphs, while the minimum density is zero, which is attained when *E= Ø*. Networks are grouped as dense and sparse according to their density: As the density approaches one, the network is defined as a dense network, on the other hand, as the density approaches zero, the network is defined as a sparse network (Wasserman and Faust, 1994).

4.1.5. Normalized Degree Variance

Degree distribution is one of the most important characteristics to classify a network as a random, small-world or scale-free network (Wang and Chen, 2003) The recent studies (Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003) have showed that most of the real-world networks consisting of collaboration networks, World Wide Web, email networks and social networks are highly heterogeneous in terms of node degree, in other words, degree variances of these networks are significantly high, which has a significant impact on information diffusion.

In this study, degree variances and average degree are used to observe the effect of degree distribution on the competitive performances of centrality-based heuristics. Two fundamental features of any distribution are mean value and variance. Since average degree represents the mean value and degree variance represents the variance of the degree distribution of the networks, these two metrics are taken into consideration to reflect the effect of degree distribution. However, degree variance cannot be used to compare different networks due to different number of nodes and edges. In this study, degree variance values of the datasets are normalized based on the equation below:

$$NDV(N) = \sum_{v=1}^{n} \frac{(k_v^2) - (AD(N))^2}{AD(N)} \quad (4.7)$$

where k_v is the degree of a node v, n is the total number of nodes in N and $AD(N)$ is the average degree of a network N (Zimmerman *et al.*, 2004).

4.2. Network Datasets and Properties

In this study, the standalone and competitive performances of seed selection heuristics are tested on 13 real-world networks, in which the minimum size is 379 nodes and the maximum size is 23370 nodes: three from Facebook, two from email communication, two from Twitter, three from collaboration, one from citation, one from Wikipedia and one from online social friendship networks. The datasets are extracted from three reputable libraries which are SNAP of Stanford University[1], Network Data Repository[2], and KONECT[3]. In this regard, thirteen different networks in Table 4.1 are chosen by paying attention to the diversity of the datasets in terms of network characteristics considered in this study.

Table 4.1. Network Characteristics of the Datasets.

No	Network	V	E	AD(N)	ACC(N)	D(N)	NAPL(N)	NDV(N)
1	*Ego-Facebook*	4039	88234	**43.69**	0.605	0.011	0.38	62.9
2	*Hamsterster*	2426	16631	13.71	0.538	0.006	0.29	28.8
3	*Email-Eu-Core*	1005	16706	33.24	0.399	0.033	0.26	41.9
4	*Ca-GrQc*	5242	14496	5.53	0.529	0.001	0.08	11.3
5	*Wiki-Vote*	7115	**100762**	28.32	0.141	0.004	**0.06**	117.0
6	*Ca-Netsci*	**379**	**914**	4.82	**0.741**	0.013	**0.56**	**3.2**
7	*Ca-Hepth*	9877	25998	5.26	0.471	0.005	0.10	7.3
8	*Cit-DBLP*	12591	49635	7.88	0.119	0.006	0.38	35.8
9	*ia-fb messages*	1266	6451	10.19	0.068	0.008	0.29	17.2
10	*Ego-Twitter*	**23370**	33101	2.80	0.069	**0.0001**	0.37	35.7
11	*rt-voteonedirection*	2277	2460	**2.16**	**0.001**	0.001	0.25	**142.1**
12	*socfb-Hamilton*	2314	96394	83.31	0.298	**0.036**	0.27	47.3
13	*Email-Univ*	1133	5451	9.62	0.220	0.009	0.37	9.1

[1] http://snap.stanford.edu/
[2] http://konect.uni-koblenz.de/
[3] http://networkrepository.com/

The network characteristics of the thirteen real-world datasets are given by Table 4.1. As it is seen from the table, the average degrees of the networks are in the range of 2.16-83.31, the average clustering coefficients are in the range of 0.001-0.741, and the densities are in the range of 0.0001-0.036. The maximum and minimum values of normalized average path length of the networks are 0.06 and 0.56 respectively. Lastly, the normalized degree of variances ranges between 3.2 and 142.1.

Network definitions and main characteristics are mentioned below and some of the networks are drawn visually to show the difference between the datasets in terms of clustering level. Other networks are not shown visually because they do not give insights into clustering levels of the networks. However, the figures below indicate that some of the datasets are highly clustered while others can be classified as non-clustered.

Ego-Facebook: is an undirected and unweighted network based on the friend lists from Facebook app. The network has 4039 nodes representing the users and 88234 edges representing the friendship relation among the users of the Facebook app. The average degree of the nodes is 43.69 (Leskovec and Mcauley, 2012).

Figure 4.1. Ego-Facebook Network with 4039 Nodes and 88234 Edges.

Hamsterster: is an undirected and unweighted network based on friendship and family links between users of the hamstersters.com website. The network has 2426 nodes with 13.71 average degree and 16631 edges (Ryan and Nesreen, 2015).

Email-Eu-Core: is a directed and unweighted network consists of email data from a large European research institution. The incoming and outgoing emails among the members of the research institution have been used to extract the data. This dataset has 1005 nodes with 33.24 average degree and 16706 edges representing that there has been an email communication between the nodes (Leskovec *et al.*, 2007).

Ca-GrQc: is an undirected and unweighted network based on the scientific collaborations between authors through papers submitted to Arxiv GR-QC (General Relativity and Quantum Cosmology) category. This network has 5242 nodes with 5.53 average degree and 14496 edges. If an author *u* co-authored a paper with author *v*, the network includes an undirected link between *u* and *v*. If the paper is co-authored by n authors, a completely connected (sub)graph on n nodes is created. This network data covers the papers from January 1993 to April 2003 (Leskovec *et al.*, 2007).

Wiki-Vote: is a directed and unweighted network consisting of 7116 nodes with 28.32 average degree and 100762 edges. The network data has been obtained from all the Wikipedia votes between the inception date of Wikipedia and January 2008. The nodes in the network represents the contributors of Wikipedia and a directed link between *u* and *v* represents that *u* voted on contributor *v* (Leskovec *et al.*, 2010).

Ca-Netsci: is a co-authorship network including 379 nodes with 4.82 average degree and 914 edges. The nodes in the network represent authors and an undirected link between *u* and *v* represents a paper co-authored by *u* and *v* (Ryan and Nesreen, 2015).

Ca-HepTh: is a collaboration network consisting of 9877 nodes with 5.26 average degree and 25998 edges. The dataset has been obtained from the e-print arXiv and it includes scientific collaborations between authors papers submitted to High Energy Physics – Theory category. If an author *u* co-authored a paper with author *v*, there is an undirected

link between *u* and *v*. If the paper is co-authored by *n* authors, this results in a completely connected subgraph of *n* nodes (Leskovec *et al.*, 2007).

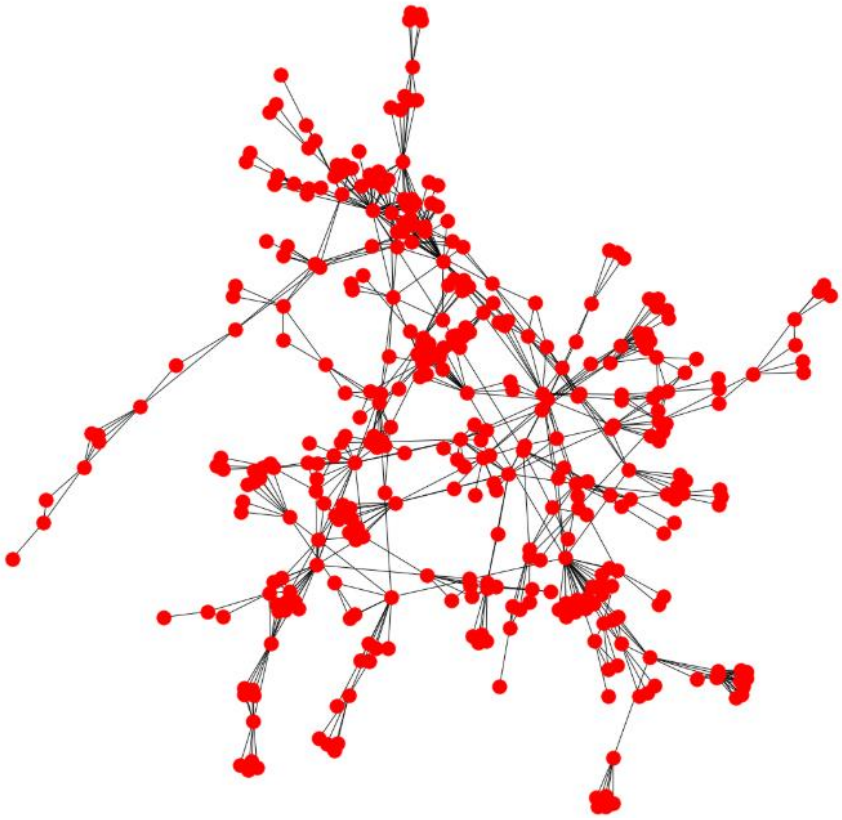

Figure 4.2. Ca-Netsci Network with 379 Nodes and 914 Edges.

Cit-DBLP: is a citation network of DBLP, a database of scientific publications, such as papers and books. The network dataset includes 12591 nodes with 7.88 average degree and 49635 edges. Each node in the network represents a publication, and each directed and unweighted edge is a citation of a publication by another. The graph contains loops since the publications are allowed to cite themselves (Ryan and Nesreen, 2015).

ia-fb-messages: is a Facebook-like social network including 1266 nodes with 10.19 average degree and 6451 edges. The dataset has been derived from an online community for students at a university in California, Irvine. Nodes denote the users of the application and an undirected edge between the user *u* and *v* represents that *u* or *v* sent at least one message to each other (Ryan and Nesreen, 2015).

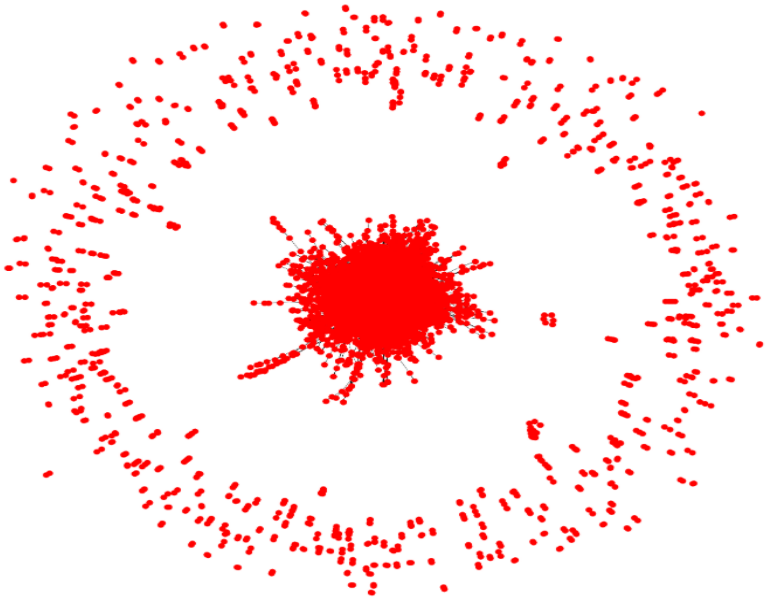

Figure 4.3. Ca-HepTh Network with 9877 Nodes and 25988 Edges.

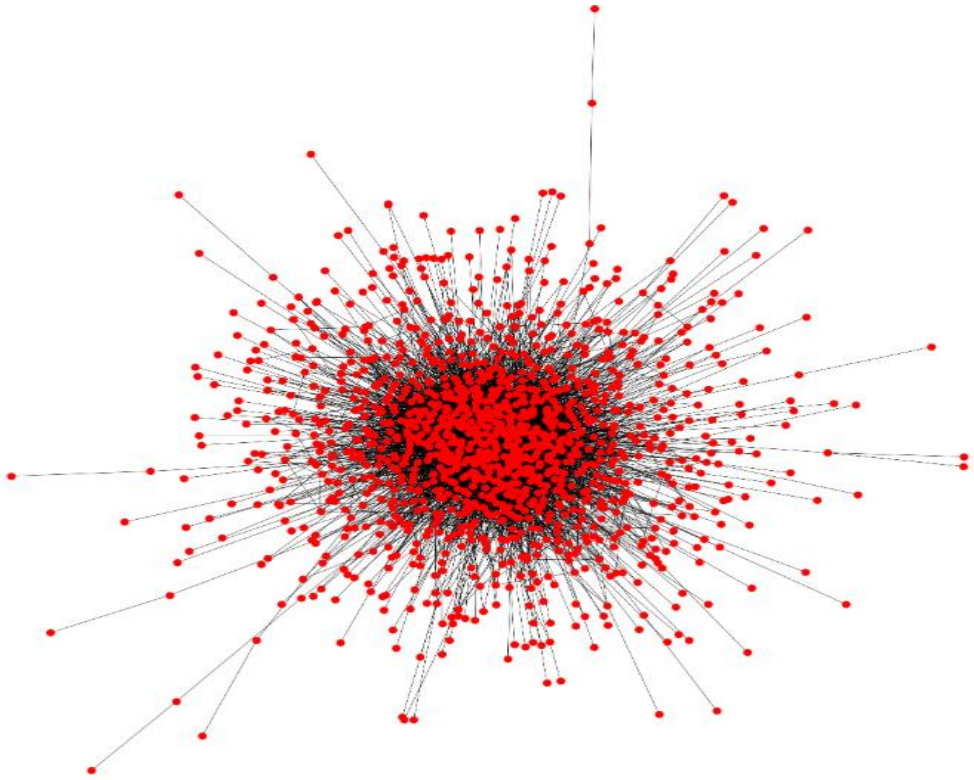

Figure 4.4. ia-fb-messages Network with 1266 Nodes and 6451 Edges.

Ego-Twitter: is a directed network consisting of 23370 nodes with 2.80 average degree and 33101 edges. It has been extracted from Twitter user-user following information. Nodes denote users of the Twitter application and edges indicate that the user *u* follows user *v* (Leskovec and Mcauley, 2012).

rt-voteonedirection: is an undirected and unweighted network under the temporal reachability networks category. The dataset has 2277 nodes with average degree 2.16 and 2460 edges. Nodes denote the users of Twitter application and directed edges represent that node *u* retweet a tweet about One Direction music group of user *v* (Ryan and Nesreen, 2015).

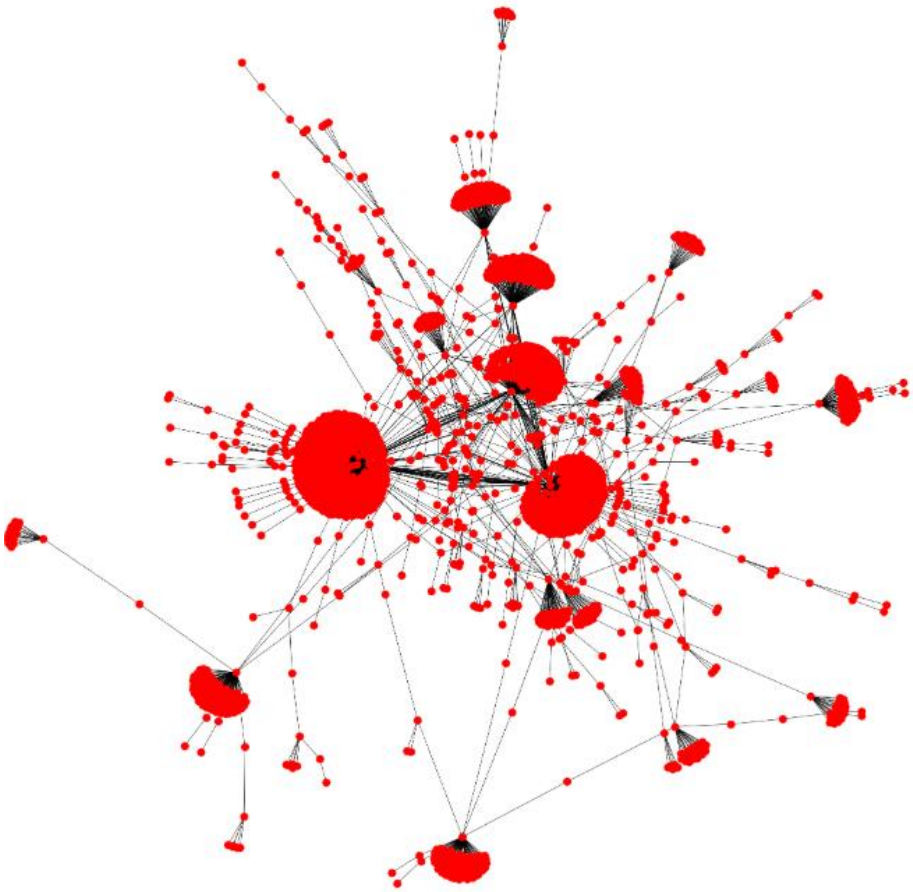

Figure 4.5. rt-voteonedirection Network with 2277 Nodes and 2460 Edges.

socfb-Hamilton: is a social friendship network extracted from Facebook. The dataset includes 2314 nodes with 83.31 average degree and 96394 edges. The nodes

represent the users of the Facebook app and undirected edges denote that user *u* and user *v* follow each other (Ryan and Nesreen, 2015).

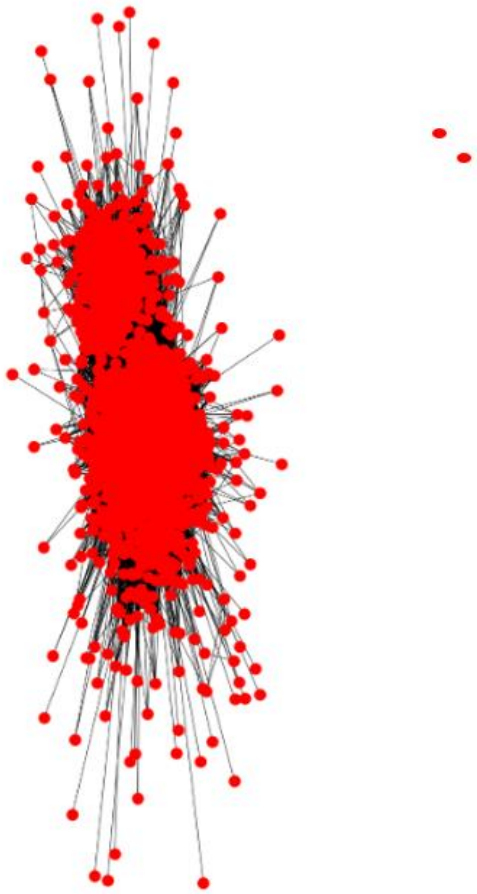

Figure 4.6. socfb-Hamilton Network with 2314 Nodes and 96394 Edges.

Email-Univ: is an email network including 1133 nodes with 9.62 average degree and 5451 edges. It has been extracted from email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. The nodes represent the users and a directed edge between user *u* and user *v* represent that user *u* sent at least an email to user *v* (Ryan and Nesreen, 2015).

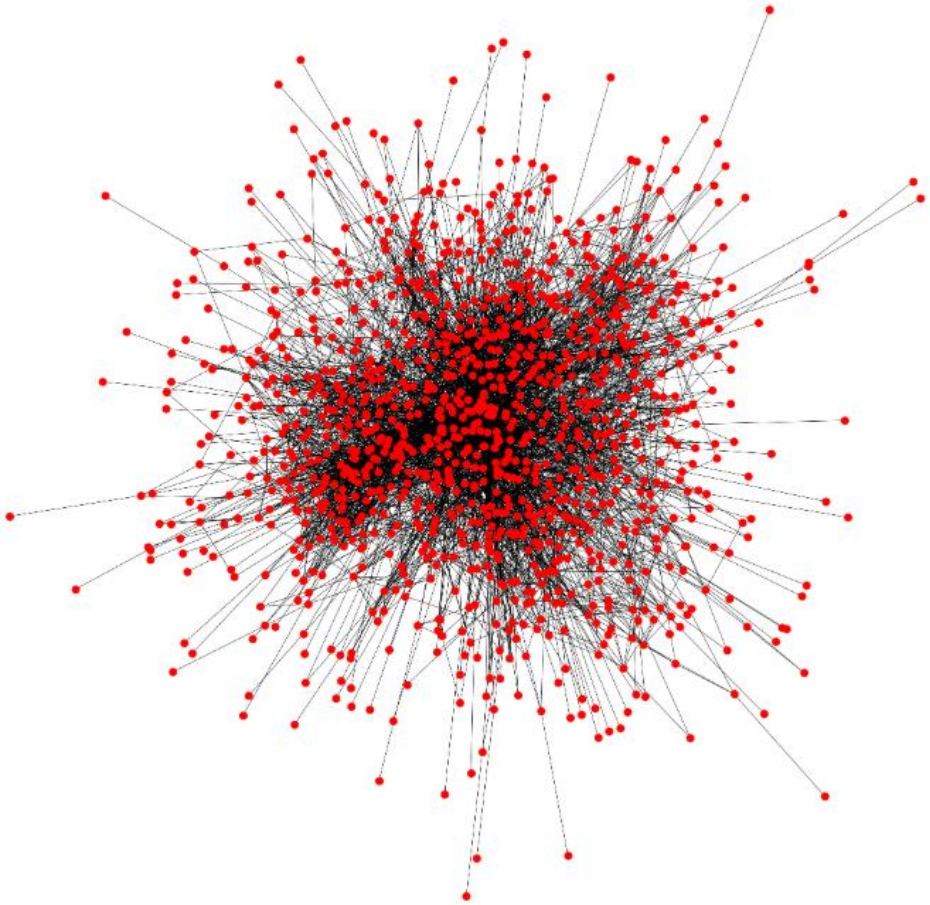

Figure 4.7. Email-Univ Network with 1133 Nodes and 5451 Edges.

4.3. Seed Selection Heuristics

In the literature, a variety of seed selection heuristics based on different metrics of nodes have been proposed. In this study, the heuristics based on *betweenness centrality*, *degree centrality*, *closeness centrality* and *eigenvector centrality* have been selected and *Random* seed selection has been used as the benchmark. In each centrality-based seed selection heuristic, the nodes with the highest metric of interest are chosen to be activated initially. These four centrality-based heuristics are the most commonly used heuristics in the literature (Borgatti and Everett, 2006), but the relation between their competitive information diffusion and the network characteristics has not been studied so far. Therefore, they are taken into consideration in this study in terms of their competitive performances and the relation between network topological characteristics and their performances. In the remainder of this section, all the heuristics will be explained in detail. The notation used in

the explanations for a simple network is *G (V, E)* with *V* being the set of nodes and *E* being the set of edges respectively. Besides, *v* represents the node which a metric is calculated for itself.

Random (R): Selecting seed set randomly.

Betweenness Centrality (BC): Selecting seeds among the nodes with the highest betweenness centrality in the network. Mathematically, betweenness centrality of a node *v* is indicated by *BC(v)* as

$$BC(v) = \sum_{v \neq t} \sum_{v \neq u} \sum_{t \neq u} \frac{g_{ut}^{v}}{g_{ut}} \tag{4.8}$$

where g_{ut} is the number of shortest paths among the nodes *u* and *t* and g^{v}_{ut} is the number of shortest paths which pass through node *v*.

Betweenness centrality is one of the well-known path-based centrality metrics which consider the ability of node *v* to control the information flow in a network (Ghalmane *et al.*, 2018). It is generally described as the ratio of the shortest paths that pass through the node of interest among whole the shortest paths in the network. The nodes with the largest betweenness centrality are crucial in the network since they act like a bridge between two communities of the network. If a node locates at the only way which other nodes have to go through for information diffusion, then that node should have a high betweenness centrality and should be significant for the ultimate information dissemination (Lü *et al.*, 2016).

Degree Centrality (DC): Selecting seeds among the nodes with the highest degree centrality in the network. Mathematically, $k_v = \sum a_{vu}$ which is the average degree of a node *v* and a_{vu} is the matrix component of adjacency matrix $A = \{a_{vu}\}$ of a network N. If *v* and *u* are connected, $a_{vu} = 1$ and 0 otherwise. Degree centrality of node *v* can be defined as

$$DC(v) = \frac{k_v}{n-1} \tag{4.9}$$

where *n* is the number of nodes on the network and *n-1* is the maximum possible degree. Degree centrality is one of the basic neighborhood-based centrality measures which depend on the significance of a node to its neighborhood topology. Generally speaking, when a node has more connections, it gets more influence on the network (Lü *et al.*, 2016).

Closeness Centrality (CC): Selecting seeds among the nodes with highest closeness centrality in the network. Closeness centrality of a node *v* can be defined as

$$CC(v) = \sum_{v \neq u} \frac{1}{(n-1)d_{vu}} \tag{4.10}$$

where d_{vu} is the geodesic distance from node *v* node *u*, which means that the number of edges in the shortest path between node *u* and *v*.

Closeness centrality is another measure to understand the importance of a node in terms of controlling information flow over the network. The closeness centrality of a node is the average distance to all other nodes in a connected network. If the graph is not connected, it is the distance of a node to all other nodes in the connected component. Closeness centrality removes the noise, which means that uncontrollable factors during information diffusion. To remove the noise, it summarizes the length of the shortest path from one node to another node in the network. In general, the node with larger closeness centrality is at a more central point of the network (Cohen *et al.*, 2014).

Both of the betweenness centrality and closeness centrality are path-based centrality metrics calculated based on the shortest paths between nodes. However, betweenness centrality is used to find the nodes undertaking bridging roles in the network, while the closeness centrality is used to find the closest nodes to other nodes in the network on average (Lü *et al.*, 2016).

Eigenvector Centrality (EC): Selecting seeds among the nodes with highest eigenvector centrality in the network. The eigenvector centrality for node *v* is defined by the equation

$$EC(v) = \frac{1}{\lambda}\sum_{u} a_{vu} EC(u) \tag{4.11}$$

where a_{vu} is the value from the adjacency matrix of the network and it is equal to 1 if *v* and *u* are connected and 0 if not. λ is a predefined constant. Eigenvector centrality is calculated by power iteration method. Each node is initialized as one in the first iteration. In the next iterations, every node shares its score with its neighbors and obtains new values. This process continues until the values of the nodes remain the same with the values of previous iteration.

Eigenvector centrality assumes that the influence of a node depends on the influence of each neighbor of the target node. In other words, the centrality of a node is correlated to total centrality of the nodes connected to the target node (Bonacich, 1987). The nodes with larger eigenvector centrality can influence many other nodes in the network through their connections since they have edges between well-connected actors of the network (Valente *et al.*, 2008).

4.4. Diffusion Model

In this thesis, to investigate the effects of network characteristics on the competitive performances of the heuristics, a simulation-based study is conducted. As required by the nature of the study, there are three steps to conduct the simulation experiments as described in Figure 4.8: (i) constructing the networks, (ii) selecting the seeds, and (iii) simulating the diffusion process. In this regard, firstly, the 13 real-world networks are constructed through NetworkX, which is a Python package to analyze social networks. This package provides opportunities to create, manipulate, and study the structural dynamics and functions of complex networks for users. In addition, the real-world network datasets in different formats such as txt, mtx and json can be easily implemented into NetworkX. Furthermore, a simulation model is constructed in Python for the procedures of seed selection and information diffusion.

All of the steps are executed in Python. The procedures of seed selection and information diffusion are explained in Section V with the pseudocodes. In this section, the

conceptual design of information diffusion model is described with the background information of Information Diffusion models.

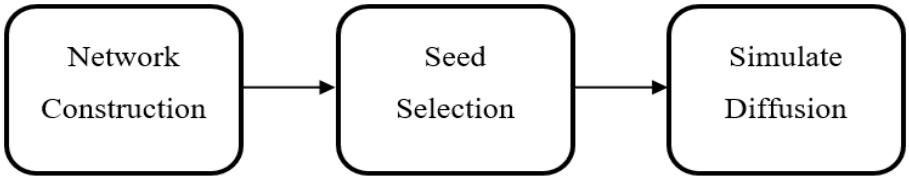

Figure 4.8. Procedure Flow Chart of the Simulation Model.

There are two diffusion models that are commonly used in standalone information diffusion experiments - the experiments which are simulated without a competitor- in the literature; namely Independent Cascade Model (IC) and Linear Threshold Model (LT). In the following two paragraphs, LT and IC models are explained. Then, the extension of LT model used in this study is described in detail.

The Independent Cascade Model is firstly proposed by Goldenberg *et al.* (2001). In IC model, a weight is assigned to each link, and the weights are classified as weak and strong in the model. There is an alternative usage of the weights as continuous numeric values in the range of [0,1]. These weights are considered as probabilities: When a node *v* is activated, it triggers an information dissemination to all its inactive neighbors. Then, propagation succeeds with w_{vu} which is defined as the probability that node *v* influence node *u.* If node *v* becomes active, it tries to activate all its inactive neighbors. The same procedure continues until the whole active nodes try propagating their inactive neighbors.

Linear Threshold Model is firstly proposed by Granovetter and Schelling (1978), and then it is generalized by Kempe *et al.* (2003). In the Generalized LT Model, each link has a weight w_{vu}, such that $\sum_u w_{vu} < 1$, which means the influence power of node *v* to activate node *u* , i.e. this weight can be considered as the impact of node *v* on node *u*. Unlike IC model, in the LT Model there is a threshold value $t_{v,}$ which is randomly selected and uniformly distributed in the interval [0,1]. Besides, there is another study in the literature where the threshold values are defined equal for all nodes to observe only the effect of link strengths on the information diffusion (Berger, 1999). The threshold value is assigned to each node as a characteristic of this model. The condition for an inactive node *v* to become

active is as follows: $\sum_u w_{vu} > t_v$, where node u is an active neighbor of node v. The propagation continues until the number of active nodes at t and t+1 is equal.

Later, Borodin *et al.* (2010) introduced two natural extensions of LT model for competitive information diffusion (propagation of more than one type of information in the network), which are *The Weight-Proportional Competitive Linear Threshold Model* and *The Separated-Threshold Model for Competing Technologies*. Firstly, the process in the Weight-Proportional Competitive LT model progresses as follows: The same uniformly distributed thresholds and weights in the interval [0,1] for the two competitors A and B are assigned to each node, i.e. they are assigned independently from the competitor type. Later, in every step t, each inactive node checks the total weight incoming from its active neighbors. Once the total weight coming from A_active and B_active nodes exceeds the threshold value of node *v*, it becomes A_active (B_active) with probability that the ratio of the total weight incoming from A_active (B_active) neighbors to the total weights incoming from all active neighbors. The same process continues until the number of active nodes at t and t+1 becomes equal.

Secondly, the process in the Separated-Threshold Model for Competing Technologies progresses as follows: Seeds at the same size are selected for competitors A and B. Then, different uniformly distributed threshold and weight values in the interval [0,1] are assigned to each node and each edge for A and B. For step t, an inactive node *v* becomes A_active (B_active) if the total weight coming from A_active (B_active) neighbors exceeds the threshold for A (B). If both threshold values are exceeded for inactive node *v*, it becomes A or B active based on the flip-coin procedure.

The LT model used in this study has more similarities between Weight-Proportional Competitive LT Model. The same weights (w_{vu}) and thresholds (t_v) for both competitors, which are uniformly distributed in the interval [0,1] are randomly assigned to the edges and the nodes respectively as in the Weight-Proportional Competitive LT model. Thus, the equal conditions for the two opponents are created to compare their competitive performances fairly. However, there are differences with the Weight-Proportional Competitive LT Model in terms of the information diffusion process. For every step t, each inactive node *v* checks the ratio between the total weight incoming from A_active and B_active neighbors and the total weight coming from all inactive and active neighbors, i.e. mathematically, checks the

ratio $\sum_{u \in I_A^t} w_{vu} \,/\, \sum_{u \in I^t} w_{vu}$ for the opponent A, and $\sum_{u \in I_B^t} w_{vu} \,/\, \sum_{u \in I^t} w_{vu}$ for the opponent B. I_A^t and I_B^t represent the A_active and B_active neighbors of node *v* at step t while I^t represent all neighbors of node *v*. If only the ratio for A or B exceeds t_v, it becomes A_active or B_active. However, if both of the ratios exceed the threshold of node *v*, it becomes A_active or B_active depending on the magnitude of the ratios. If $\sum_{u \in I_A^t} w_{vu} \,/\, \sum_{u \in I^t} w_{vu}$ is higher than $\sum_{u \in I_B^t} w_{vu} \,/\, \sum_{u \in I^t} w_{vu}$, node *v* becomes A_active. Otherwise, it becomes B_active. Since it is more realistic to adopt the influence of more influential competitor, in this study, selecting A or B randomly has not been found appropriate when both of the weight ratios exceed t_v. Later, the process terminates when the number of active nodes at t and t+1 becomes equal.

Referring one of the secondary objectives of the thesis which is comparing the effects of the network characteristics on the standalone and competitive performances of the heuristics, the simulation experiments for the standalone performances of the centrality-based heuristics are also conducted. In this regard, the same competitive information diffusion model is used. However, this time, the seeds are selected and activated only for one of the opponents, and 0 (inactiveness status) is assigned to whole nodes for the other opponent. Thus, the model behaves appropriately also for the information diffusion in a non-competitive environment.

5. EXPERIMENTATION AND ANALYSIS PROCEDURE

As it is stated in the previous section, there are three steps of this simulation-based study (see Figure 4.8). Before proceeding the experiments, all the three steps of the simulation model is tested in terms of the consistency between the conceptual design and the execution. After ensuring the accuracy of construction of networks, seed selection procedure and the information diffusion procedure, the model is simulated in Python for both standalone and competitive experiments. In this section, the criteria and the steps of the model testing are detailed. Later, the steps of the seed selection and diffusion procedures are described with the pseudocodes. Lastly, the procedure of analyzing the results of simulation experiments are explained.

5.1. Model Testing

To ensure the accuracy of network construction step, the network datasets topological characteristic values seen in network libraries are compared with the values seen on NetworkX (number of edges, number of nodes, average clustering coefficient, average degree, average path length and density). As a result, it was observed that NetworkX reflected the datasets accurately.

To test the second step, it is checked whether the number of selected seeds is the same with the initially defined seed set size (*s*) and whether the selected nodes are chosen according to the selected heuristic properly or not. In this regard, all of the heuristics are checked firstly for the standalone experiments, and then each heuristic combination is also checked for the competitive experiments. The smallest dataset *Ca-Netsci* is utilized for the test. In order to explain the steps of this test procedure, *BC vs. DC* heuristic combination is chosen as an example and the steps of this test are detailed through this example. The seed set size (*s*) is defined as 5% for all of the experiments depending on a value that works well enough for the study. If *s* is too high, then whole the heuristics are expected to reach a very high influence spread. On the contrary, if *s* is too low, all of the heuristics achieve a limited influence spread (Erkol and Yücel, 2017). Thus, the number of initially active nodes should be 19 in *Ca-Netsci* (0,05×379= 19) for both of the seed sets chosen based on *BC* and *DC*. A

node can have highest betweenness and degree centralities at the same time. Besides, although there is a low probability, seed selection priority might always belong to the same opponent during the seed selection since it depends on a flip-coin procedure. Thus, in the test conducted on *Ca-Netsci,* the same 19 nodes may have the highest betweenness and degree centralities at the same time. Meaning that, if the selection priority belongs to the same opponent during the seed selection, seeds can be chosen at most from the first 38 nodes with highest betweenness centralities based on *BC* and the first 38 nodes with the highest degree centralities based on *DC*. Besides that, a flip-coin procedure is defined for the seed selection process for the competitive experiments. The competitors select the seeds one by one, and the seed selection process is completed when both of the competitors reach predefined *s.*

For the test, the first 38 nodes with the highest betweenness and degree centralities of *Ca-Netsci* are listed. When the simulation model is run to activate nodes in *Ca-Netsci* based on *BC* and *DC*, it seen that both of the seed sets contain the nodes belong that list. Thus, the model selects seeds as it is designed. Apart from this observation, the selected nodes based on the heuristic *BC* and *DC* do not overlap as it is planned. The same procedure is repeated for each heuristic combination and also for each standalone case, and then it is observed that seed selection procedures are working properly.

Lastly, the same dataset is utilized to verify the competitive diffusion model used in this study since it has the smallest number of nodes and also it is a connected network, which means that it is possible to activate all inactive nodes in the network. Two simulation experiments are carried out by assigning zero and one combinations to the node thresholds and the edge weights. Firstly, competitive information diffusion between *BC* and *DC* is simulated when the node thresholds are zero and edge weights are one. At the end of the simulation, all nodes became A_active (the first competitor) or B_active (the second competitor) as it is expected. On the contrary, when the node thresholds are one and the weights of all edges are zero, the number of active nodes did not exceed the seed set size as it is planned. Moreover, a synthetic network is created on NetworkX, which is also connected and have only 15 nodes. The same verification procedure is repeated on this network with each standalone case and for each heuristic combination. It is observed that if all threshold values are assigned zero and all edge weights are assigned one, all of the nodes in the network

are activated at the end of the simulation. If the assignment is vice versa, only the nodes in the seeds remain active.

5.2. Experimentation Procedure

For this study, thirteen different real-world network datasets have been selected. Standalone experiments are conducted on each network for each heuristic (*R*, *BC*, *DC*, *CC* and *EC*) with 40 replications. Thus, there have been 13×40=520 runs for the standalone performance of each heuristic. Since all of the other heuristics are able to exceed the standalone performance of *R*, they are compared against each other (*BC vs DC*, *BC vs CC*, *BC vs EC*, *DC vs CC*, *DC vs EC*, and *CC vs EC*). Competitive experiments are also conducted on each network with 40 replications, thus there have also been 520 runs for each paired heuristic combination. Thus, there are 5720 simulation runs overall, 520×5=2600 for the standalone experiments and 520×6=3120 for the competitive experiments. All experimentation procedures are implemented in Python 3.7.0 and the following experiments are conducted on Windows 64-bit OS with 2.90 GHz Intel Core i7-7500U and 8 GB memory.

At the beginning of each experiment, uniformly distributed node thresholds and the edge weights are assigned randomly in a range between one and zero. The same random seeds which are used to assign thresholds and weights for each of 40 replications are kept on NetworkX for starting to each simulation experiment under the identical conditions (a total of 80 random seeds are kept for 40 replications; 1 for thresholds and 1 for weights). After the assignment of thresholds and weights, *s* (seed set size) is defined as 5% of total numbers of nodes in the subjected dataset. Another step is to select the nodes to be activated initially based on a heuristic for the standalone experiments or based on the two heuristics for the competitive experiments. In Figure 5.1, *R* seed selection procedure for standalone experiments is presented. In Figure 5.2, seed selection procedures of centrality-based heuristics for standalone experiments are grouped and described. In Figure 5.3, seed selection procedure for the competitive experiments of a centrality based heuristic vs. another centrality-based heuristic are presented. In that figure, *c>= 0.5* statement presents the flip-coin rule. Since a node can have the highest values of different metrics at the same time, flip-coin rule is applied to give competitors equal chance to activate the nodes to be selected as seeds.

After all initialization steps are completed, the simulation is run according to the diffusion model used in our study. In Figure 5.4 and Figure 5.5, diffusion processes for standalone and competitive influence maximization experiments are described respectively.

for randomly selected *s* nodes **do:**
  activate
end for

Figure 5.1. Pseudocode of *Random* Seed Selection Procedure in Standalone Experiments.

for each node *v* **do:**
  set v_x: calculate *BC / DC / CC / EC*
end for
for *s* nodes with maximum v_x
  activate
end for

Figure 5.2. Pseudocode of Centrality-Based Heuristics' Seed Selection Procedure in Standalone Experiments.

```
a = 0  // to iterate the total number of A_active nodes
b = 0 // to iterate the total number of B_active nodes
for each node v do:
   set v_x:  calculate BC / DC / CC/ EC
   set v_y: calculate BC / DC / CC/ EC
end for
while a+b < 2s:
  if a < s and c>=0.5: // c is a random number in [0,1]
      pick v with maximum v_x from the list of inactive nodes
      activate v
      a=a+1
    else:
       if b < s:
           pick v with maximum v_y from the list of inactive nodes
           activate v
           b=b+1
    end if
end while
```

Figure 5.3. Pseudocode of Centrality-Based Heuristic vs. Another Centrality-Based Heuristic Procedure in Competitive Experiments.

```
for each inactive node v do:
  set m: calculate the number of A_active neighbors
  set n: calculate total weight coming from A_active neighbors
  set p: calculate total weights coming from inactive neighbors
  set n / (n+p): average influence of competitor A
  set t: threshold
end for
for each inactive node v do:
   if m!= 0:
      if (n/(n+p)) > t
            activate node v
end for
```

Figure 5.4. Pseudocode of Information Diffusion Process in Standalone Experiments.

```
for each inactive node v do:
    set m: calculate the number of A_active neighbors
    set n: calculate total weight coming from A_active neighbors
    set u: calculate the number of B_active neighbors
    set p: calculate total weights coming from B_Active nodes
    set r: calculate the number of inactive neighbors
    set y: calculate total weights coming from inactive neighbors
    set n / (n+p+y): average influence of competitor A
    set p / (n+p+y): average influence of competitor B
    set t: threshold
  end for
  for each inactive node v do:
    if (m+u)!= 0:
        if (n/(n+p+y)) > t and (p/(n+p+y)) > t:
            if  0.5 <= (n/(n+p)): // to activate v based on the opponent with higher weight
                activate node v as A
            else:
                activate node v as B
        if (n/(n+p+y)) > t > (p/(n+p+y)):
                activate node v as A
        if (p/(n+p+y)) > t > (n/(n+p+y)):
                activate node v as B
  end for
```

Figure 5.5. Pseudocode of Diffusion Process in Competitive Experiments.

5.3. Statistical Analysis Procedure

In this study, the results of the experiments contain one dependent variable and six independent variables. The dependent variable is the average influence spread of the heuristics while the independent variables are the five network characteristics, and additionally the opponent type for the competitive influence maximization experiments. To analyze how the average influence spreads of the heuristics change in response to the network characteristics, regression tree is used. Regression tree is one of the methods of decision tree methodology. In this section, background information for regression trees is given, and the related procedure followed in this study is explained. Besides, detailed information with an example which explains how to read regression trees can be found in Appendix A.

Decision tree is one of the methods of Machine Learning and covers both Classification and Regression trees (CART), which is commonly used as statistical tools for modeling and analyzing complex data. Classification and regression trees explain the variation of a single response (dependent) variable by one or more explanatory (independent) variables. Selection of classification or regression tree method depends on the structure of the response variable. Classification trees are utilized to explore categorical response variables, while regression trees are used for numeric response variables. Apart from the response variable, both methods can be used with categorical or numeric explanatory variables (Breiman *et. al.*, 1993). In this study, since the response variable representing the average influence spreads of the heuristics is numeric, regression tree is used.

Regression trees are formed by repeatedly splitting the data depending on a single explanatory variable. In each split, data is divided into two mutually exclusive groups which are always uniform in itself. Then, the splitting rule is applied repeatedly to each group and continues until all of the final groups named as terminal nodes are reasonably small and pure, which means that each terminal node consists of one sample case. The main aim is to partition the single response variable into uniform groups as possible, but also to obtain a tree at an acceptably small size. The size of a tree is the total number of its terminal nodes. In regression trees, terminal nodes are distinguished by the mean values of the response variable, size of the node (amount of the data included in the node) and a certain criteria of

explanatory variables. These certain criteria are defined by inequalities (for numerical response variables) or equalities (for categorical response variables) on the branch above the terminal nodes, and thus why some of the data is included in the terminal node of interest is explained according to these criteria (Breiman *et. al.*, 1993).

Regression trees are illustrated as a decision tree, in which there is a root (parent) node on the top which indicates the mean value of the non-partitioned data and there are terminal nodes at the bottom which show the mean values of the split response variable. Besides, the routes between the root node and each terminal node are named as branches (Breiman *et. al.*, 1993).

In this study, the results of the performances of heuristics are compared based on their averages over 40 replications on each network dataset. Each individual replication result is taken into consideration in the analysis as a percentage. The average influence spread of the heuristics based on the individual results is the dependent variable while average clustering coefficient, average degree, normalized average path length, normalized degree variance and density are the independent variables. Apart from the network characteristics, the type of the opponent spreading over the network is the additional independent variable in the analysis of competitive experiment results. On the parent node of the regression trees, the overall average influence spread of the heuristics are seen regardless of the effect of independent variables. In the following nodes, it is presented that a heuristic performs better or worse according to a decrease or increase in which network characteristic. Besides, how the performance of a heuristic change depending on the opponent type is presented in the regression trees for the competitive influence maximization experiment results.

To sum up, the standalone performances of all centrality-based heuristics exceed the performance of the heuristic *Random* on most of the datasets. Therefore, all the four centrality-based heuristics' both standalone and competitive performances are interpreted through regression trees created with RPART package of RStudio program. Overall, the four regression trees are created for the standalone and the four trees are created for competitive experiment results.

6. RESULTS

In this section, the results with the standalone and competitive performances of the heuristics are represented separately. The effects of network characteristics are interpreted on the standalone and competitive performances of each heuristic through regression trees. At the end of both standalone and competitive performance results' subsections, the findings are summarized in an overview section. All the findings of Section VI will be discussed further with respect to meaning, importance and relevance in Section VII for Discussion.

6.1. Standalone Performances of the Centrality-Based Heuristics

In this study, the seed selection heuristic *Random (R)* is used as the benchmark, as Zhao *et. al.* (2017) did. The authors studied competitive performances of several seed selection heuristics including *betweenness (BC), degree (DC), closeness (CC) and eigenvector centrality (EC).* They compared each heuristic against *R* on the two real-world network datasets which have 379 and 1133 nodes. Then, since *EC* performed worse than *R* on these two datasets, the authors eliminated *EC* from the comparative analysis of the standalone performances of centrality-based heuristics (Zhao *et al.*, 2017). In this study, simulated standalone performances are listed in Table 6.2 according the averages over the results of 40 replications, and then the heuristics are ranked in Table 6.1 based on these average influence spreads. In Table 6.1, "1" represents the heuristic with the best average standalone performance on the network of interest, while "4" represents the heuristic with the worst performance. According to Table 6.1, the heuristic *EC* performed worse than *R* on *Ca-Netsci* (N6), which is one of the two real-world datasets used in the study of Zhao *et al.* (2011). On the contrary, *EC* performed better than the heuristic *R* on the eight networks among all the 13 real-world datasets as shown in the Table 6.1. Besides, the heuristics *BC, DC* and *CC* also influenced more nodes than *R* on all of the datasets. Hence, all of the centrality-based heuristics are included in the analysis of standalone performances and competitive performances as well.

Table 6.1. Ranks of the Heuristics According to Standalone Performances.

	N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**
R	4	5	5	4	5	4	4	5	5	5	5	5	5
BC	1	1	1	1	1	2	1	1	1	2	2	1	1
DC	2	2	2	3	2	1	2	2	2	1	1	2	2
CC	3	3	3	2	3	3	3	3	4	3	3	4	3
EC	5	4	4	5	4	5	5	4	3	4	4	3	4

Table 6.2. Average Standalone Performances of the Heuristics on the Datasets (%).

No	**Network**	**mean(*R*)**	**mean(*BC*)**	**mean(*DC*)**	**mean(*CC*)**	**mean(*EC*)**
1	*Ego-Facebook*	20.2	39.7	32.5	28.7	**8.3**
2	*Hamsterster*	**33.8**	48.4	47.1	43.5	37.6
3	*Email-Eu-Core*	19.8	57.5	56.1	55.2	53.0
4	*Ca-GrQc*	15.6	**34.9**	**23.5**	**25.4**	12.6
5	*Wiki-Vote*	22.1	81.5	81.4	**77.0**	75.1
6	*Ca-Netsci*	17.5	37.1	41.7	28.8	12.8
7	*Ca-Hepth*	16.7	38.5	34.2	27.8	21.9
8	*Cit-DBLP*	19.2	68.3	63.0	48.8	44.8
9	*ia-fb messages*	19.1	63.0	63.0	59.9	**60.7**
10	*Ego-Twitter*	**15.9**	**93.6**	98.9	51.6	41.5
11	*rt-voteonedirection*	17.3	91.5	**95.5**	65.4	55.8
12	*socfb-Hamilton*	21.8	46.1	45.5	42.9	43.2
13	*Email-Univ*	19.1	48.1	46.9	41.7	35.4
Overall Mean		**19.9**	**57.6**	**56.1**	**45.9**	**38.7**

From Table 6.1, it is shown that there is a general ranking of the four heuristics as follows: *BC*>*DC*>*CC*>*EC*. However, when the results with *R*, *BC*, *DC*, *CC* and *EC* are interpreted according to Table 6.2, it can be clearly seen that the network type significantly affects the standalone performances of the heuristics within themselves. For example, all the four centrality-based heuristics show better standalone performances on *Wiki-Vote* dataset which has the lowest NAPL (normalized average path length) among the datasets. The heuristics *BC* and *DC* apparently perform better on *rt-voteonedirection* which has the lowest ACC (average clustering coefficient). On the other hand, *CC* and *EC* cannot increase their performances as much as *BC* and *DC* on *rt-voteonedirection*. Hence, the standalone performances of the heuristics are significantly affected by the network characteristics. The sub-sections below will be helpful in understanding which network characteristics are

influential on the performances of the heuristics. However, the reasons behind the effects of network characteristics will be detailed in Section VII for Discussion by combining standalone and competitive experiment results.

6.1.1. Betweenness Centrality

The results with the heuristic *BC* showed that average influence spread is 58% over all of the networks, which is the highest average performance among all of the heuristic performances. Considering the individual experiment results, the minimum performance is observed as 30% on *Ca-Netsci* while the maximum performance is seen on *Ego-Twitter* with 94% influence spread.

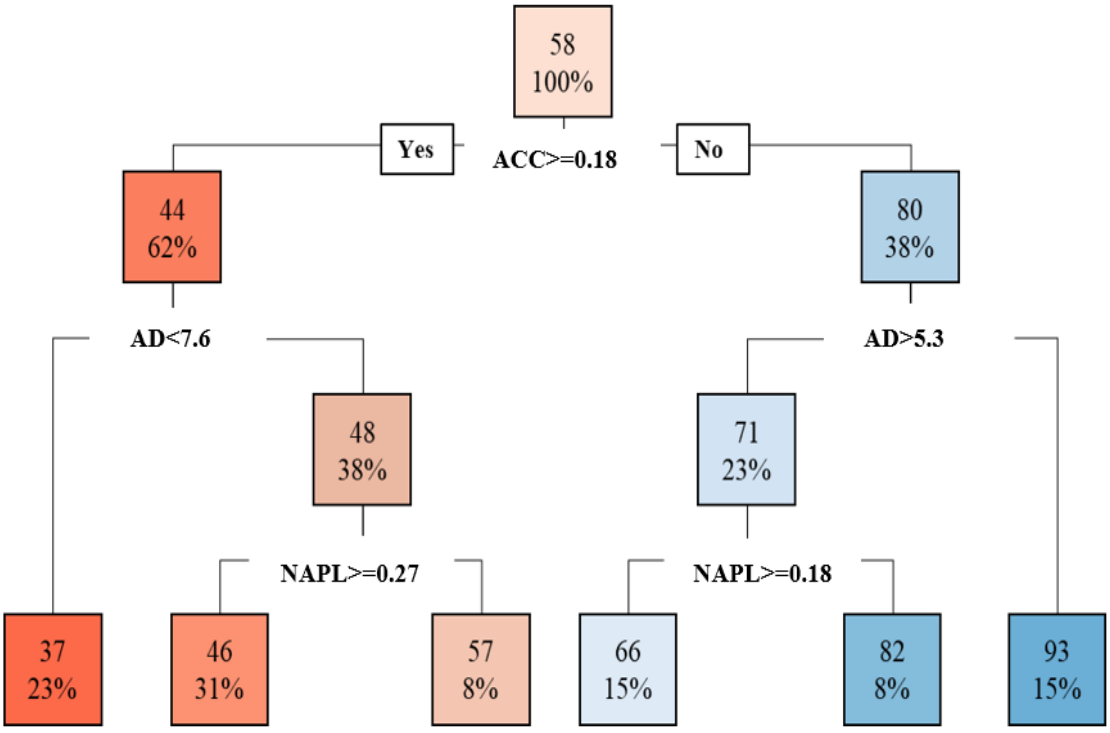

Figure 6.1. Regression Tree of *BC*'s Standalone Performance.

Figure 6.1 indicates that ACC (average clustering coefficient) is the most significant network characteristic on the standalone performance of *BC* because the topmost split point separates off 62% off the data with 44% average influence spread from 38% of the data with 80% average influence spread. Hence, it is found that the heuristic *BC* performance is better

on slightly clustered networks. Barthelemy (2004) stated that lower ACC maximizes betweenness centrality of some nodes since all shortest paths between different clusters are prone to go through the same nodes. However, when the ACC is high enough, shortest paths in a network go through much more nodes. Meaning that, the average betweenness centrality of the network is also high, and it is less possible for the nodes to make a difference in controlling communication on the network, i.e. being a node with higher *BC* is not a property that makes this node superior to other nodes for information diffusion (Barthelemy, 2004). Thus, selecting seeds based on the betweenness centrality of nodes is not sensible on highly clustered networks.

The AD (average degree) is found as the second most important network characteristic because the first split points of the left and the right branches in Figure 6.1 are on AD. The best performance of *BC* is seen on slightly clustered networks with lower AD. However, higher AD is more advantageous for *BC* on highly clustered networks (see the left branch in Figure 6.1). When the network is highly clustered, propagation becomes limited regardless of the seed selection heuristic. Thus, it can be concluded that the nodes with a large number of neighbors impacts the propagation.

The last split points of the *BC* performance are on the NAPL (normalized average path length). An increase in the NAPL negatively affects the performance of *BC* in standalone experiments. Since betweenness centrality is a path-based centrality metric Jahanpour and Chen (2013) expected to see that the nodes with higher betweenness centrality have more potential to control communication in the networks with lower NAPL. Their experimental results have supported their hypothesis (Jahanpour and Chen, 2013). The results of our study are also in line with the authors' claim. However, since the NAPL splits a little amount of the data compared to the ACC and the average degree in Figure 6.1, it can be concluded that it is not influential on the performance of *BC* as much as the ACC and the average degree.

There is no split point on the NDV (normalized degree variance) and the D (density) in the regression tree above. Meaning that, these two variables do not affect the performance of *BC* as much as the ACC, the average degree and the NAPL.

Overall speaking, the results indicate that standalone performance of *BC* is significantly affected by the ACC and the AD. To obtain the best propagations from the heuristic *BC*, it should be used in slightly clustered networks with low AD. A decrease in the NAPL also increases the performance of the heuristic BC, but with a less significance compared to the ACC and the AD.

6.1.2. Degree Centrality

The results with the heuristic *DC* showed that average influence spread is 56% over all of the networks, which is the second highest average influence spread result among all of the heuristic performances. Considering the individual experiment results, the minimum performance is observed as 23% on *Ca-GrQc* while the maximum performance is seen on *Ego-Twitter* with 99% influence spread.

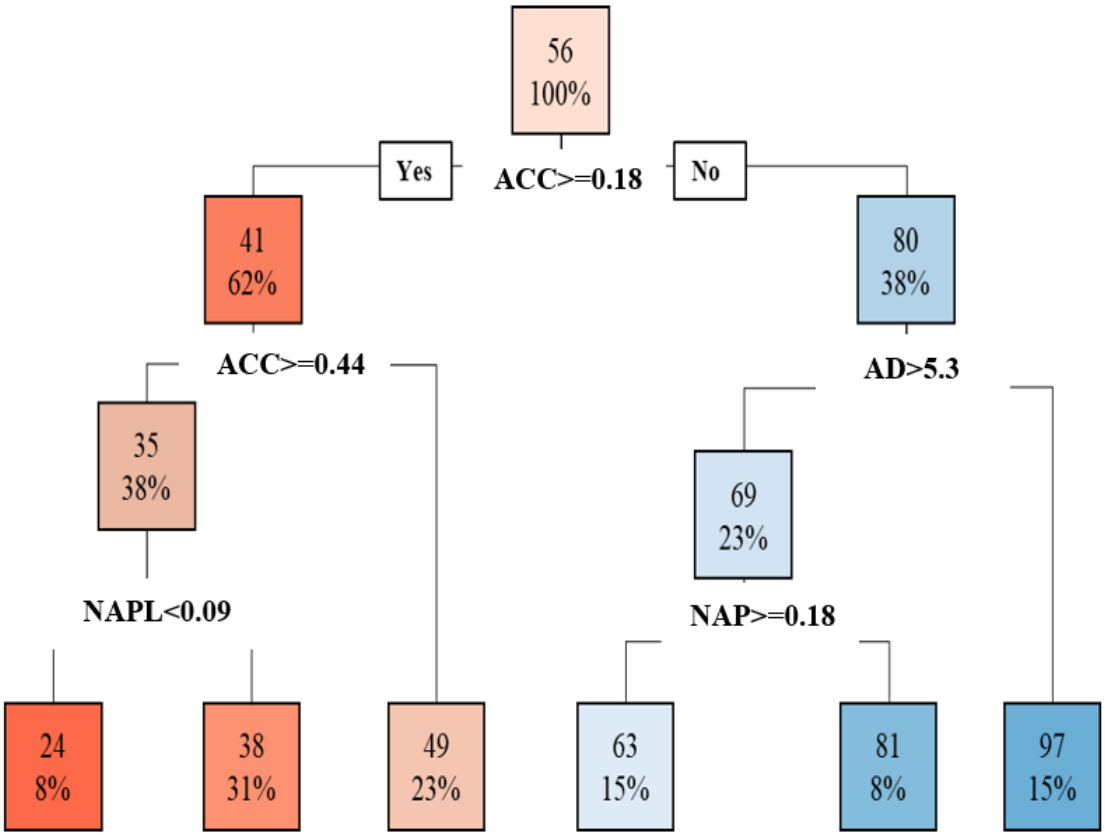

Figure 6.2. Regression Tree of *DC*'s Standalone Performance.

In Figure 6.2, the topmost split point, according to the ACC (average clustering coefficient), separates off 62% of the data with 41% average influence spread from 38% of the data with 80% average influence spread. As the ACC decreases, the standalone performance of *DC* increases too. Additionally, the first split point of the left branch is also on the ACC. The heuristic *DC* also shows its best performances on the slightly clustered networks.

The first split point of the right branch is on the AD (average degree). When the ACC of a network is lower, the AD becomes the most significant characteristic on the standalone performance of *DC*. The best performance of *DC* has been observed on slightly clustered networks with low AD, which are Twitter networks, namely *rt-voteonedirection* and *Ego-Twitter*. As it is presented in Subsection 6.1.1, *BC* performance is also affected negatively by an increase in the average degree of the slightly clustered networks. The related split points on the average degree of *BC* and *DC* have also the same thresholds AD>=5.3. Besides, it is observed that both heuristics showed approximate average performances (*BC:* 71%; *DC:* 69%) on the networks belong to these split points. Meaning that, the heuristics *BC* and *DC* are affected similarly by the ACC and the AD of the networks.

The third most influential characteristic is found as the NAPL (normalized average path length). It is interesting to see that an increase in the NAPL is advantageous for the performance of *DC* on the highly clustered networks while it is disadvantageous on the slightly clustered networks. However, the NAPL splits a little amount of the data compared to the ACC and the AD. Hence, the NAPL is not influential as much as the ACC and the AD for the performance of *DC*.

There is no split point on the NDV (normalized degree variance) and the D (density) in Figure 6.2. Meaning that, these two variables do not affect the performance of *DC* as much as ACC, AD and NAPL.

To sum up, the most significant characteristics affecting the standalone performance of *DC* are the ACC, the AD and the NAPL respectively. The best performance of *DC* is on the slightly clustered networks with lower AD. Moreover, when the AD of a slightly

clustered network increases, a decrease in the NAPL is a supporting factor for the heuristic *DC* to propagate better.

6.1.3. Closeness Centrality

The results with the heuristic *CC* showed that average influence spread is 46% over all of the networks, which is the third highest average influence spread among all of the heuristic performances. Considering the individual experiment results, the minimum performance is observed as 22% on *Ca-Netsci* while the maximum performance is seen on Wiki-Vote with 80% influence spread.

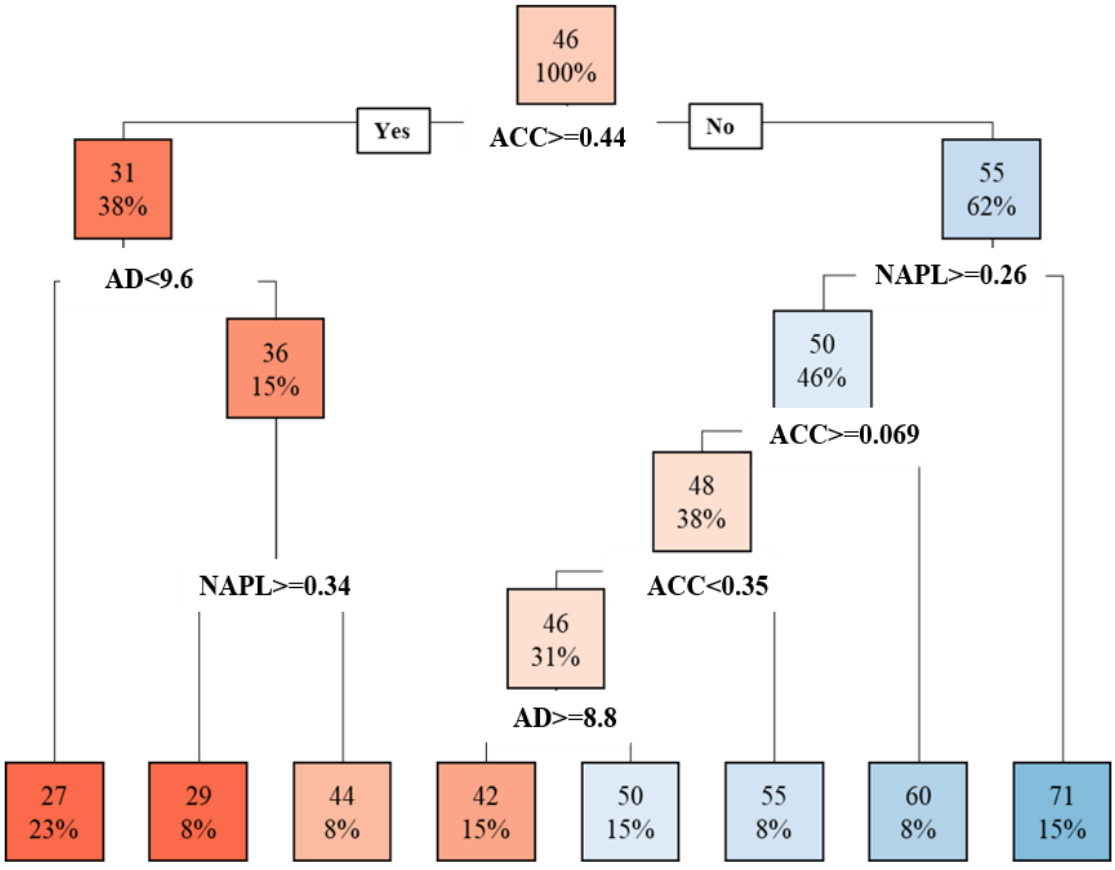

Figure 6.3. Regression Tree of *CC*'s Standalone Performance.

The topmost split point in Figure 6.3 is on the ACC (average clustering coefficient) which separates off 38% of the data with 31% average influence spread from 62% of the data with 55% average influence spread. As in the regression trees of *BC* and *DC* standalone

performance, *CC* also performs better on the networks with a lower ACC. However, the threshold value of the first split points is 0.44 for *CC* while it is 0.18 for *BC* and *DC*. The performances of *BC* and *DC* considerably decrease when the threshold of the networks exceed 0.18. However, the performance of *CC* does not significantly change until the threshold reaches 0.44. Thus, the results conclude that the ACC is the most significant network characteristic on the performance of the three heuristics, but it has a smoother effect on *CC* compared with the effects on *BC* and *DC*.

The first split point on the right branch is on the NAPL (normalized average path length), and it separates 46% of the data with 50% average influence spread from 15% of the data with 71% average spread. There is another split point on the NAPL on the left branch of the regression tree. Both of the split points indicates that the heuristic *CC* propagates better as the NAPL of the network decreases. Meaning that, the performance of *CC* is significantly and positively affected by a decrease in the NAPL regardless of the clustering levels of the networks. The closeness centrality of a node indicates the reciprocal sum of the shortest path distances from that node to all other nodes in a network. Therefore, if the NAPL of a network is low, the nodes with high *CC* value will have more capability to control the diffusion on the network (Yen *et al.*, 2013).

Another finding of Figure 6.3 is the fact that the AD (average degree) effect on the performance of *CC* arises when the ACC of the network increases. It is found that the worst *CC* performance occurs on the networks with highest clustering coefficient and lowest average degree.

As shown in Figure 6.3, there is no split point on the D (density) and NDV (normalized degree variance) in the regression tree. Thus, the results conclude that the performance of *CC* does not change significantly due to the effects of the NDV and the D.

Consequently, the most important network characteristics for the performance of *CC* are ranked as the ACC, the NAPL and the AD respectively. The effect of AD is important only on the highly clustered networks which the heuristic *CC* propagates limitedly. To obtain best influence spread performances, the heuristic *CC* should be used on the slightly clustered networks with lower NAPL.

6.1.4. Eigenvector Centrality

The results with the heuristic *EC* showed that average influence spread is 39% over all of the networks, which is the lowest average influence spread among all of the four heuristics' standalone performances. Considering the individual experiment results, the minimum performance is observed as 7% on *Ego-Facebook* while the maximum performance is seen on *Wiki-Vote* with 79% influence spread.

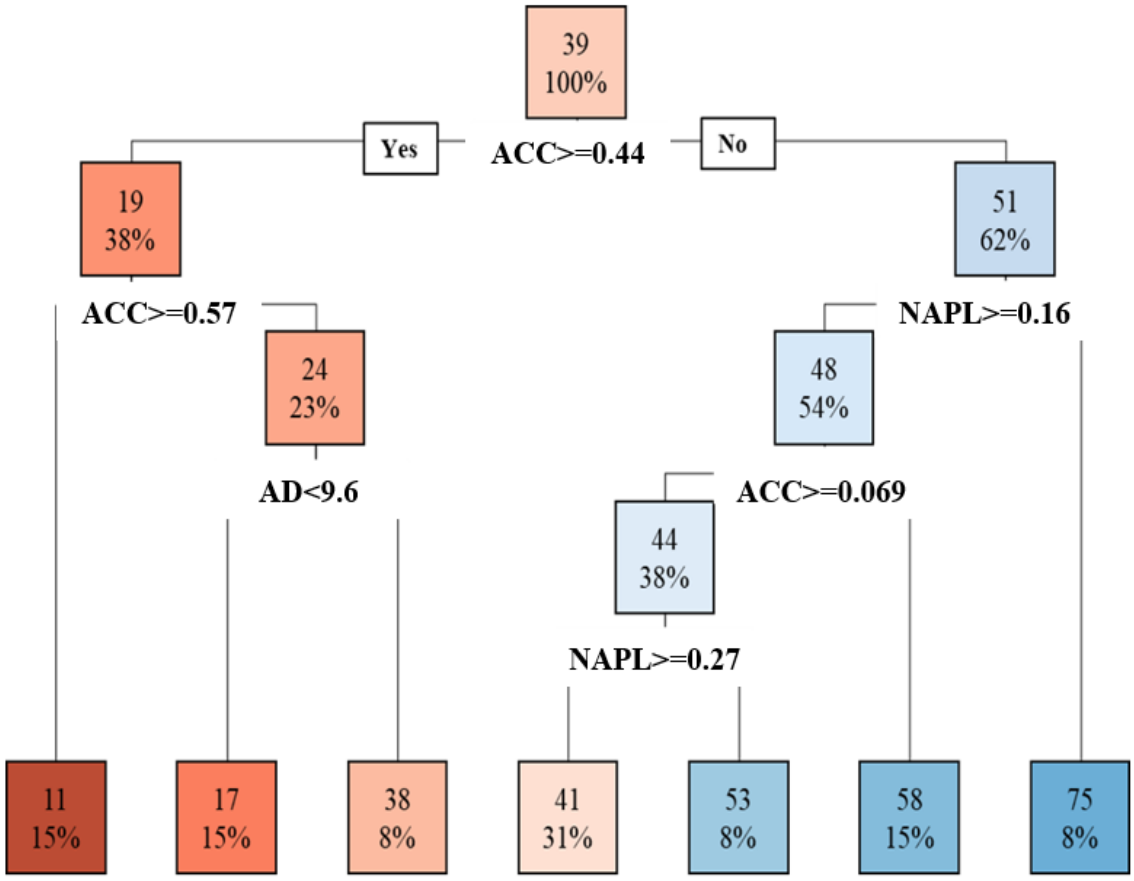

Figure 6.4. Regression Tree of *EC*'s Standalone Performance.

Figure 6.4 shows that the ACC (average clustering coefficient) has the highest effect on the performance of *EC* since the topmost and the first split point of the left branch are on the ACC. As the ACC of a network decreases, the heuristic *EC* reaches higher influence spreads. Similar to the heuristic *CC*, the threshold value for the topmost split point on the ACC is 0.44 which is higher than the value for *BC* and *DC* (0.18 for *BC* and *DC*; 0.44 for

DC and *EC).* Meaning that, the ACC has a smoother effect on also *EC* compared with the effects on *BC* and *DC.*

The NAPL (normalized average path length) of a network is observed as the second most significant characteristic on the performance of *EC*. The first split point on the right branch separates off the 54% of the data with 48% average spread from the 8% of the data with 75% average spread. Meaning that, the NAPL is the influential factor to obtain the best propagation from the heuristic *EC* on the slightly clustered networks.

The AD (average degree) splits a small percentage of the data on the left branch of the regression tree. The effect of the AD arises only on the highly clustered networks. An increase in the AD fosters the higher performance of the heuristic *EC*, but with a less impact. Moreover, there is no split point according to the NDV (normalized degree variance) and the D (density). Thus, it can be concluded that the NDV and D do not have a significant impact on the performance of the heuristic *EC.*

Overall, the ACC and the NAPL of the networks significantly affect the standalone performance of *EC*. To obtain the best performance from the heuristic *EC*, it should be used on the slightly clustered networks with lower NAPL.

6.1.5. Overview of the Standalone Performance Results

Before proceeding to the results of competitive performances of the heuristics, the findings of standalone experiments are summarized in Table 6.3. In the table, positive direction of correlation represents that as the value of network characteristic increases, the performance of the heuristic tends to also increase. If the direction is negative, the performance of the heuristic decreases with an increase in the network characteristic.

It is found that the average clustering coefficient (ACC) is the most influential characteristic on the performances of the heuristics as shown in Table 5.3. All the four centrality-based heuristics propagate better as the ACC of the network decreases. Besides, although the heuristics *BC* and *DC* performs better than the heuristics *CC* and *EC* on both

highly and slightly clustered networks, the results showed that the heuristics *BC* and *DC* are more sensitive to a change in the ACC compared to *CC* and *EC*.

Table 6.3. Summary Table of Network Characteristics' Effects on the Standalone Performances of the Heuristics.

	Significant Network Characteristics	**Direction of Correlation**
BC & DC	Average Clustering Coefficient (ACC)	Negative
	Average Degree (AD)	Negative
	Normalized Average Path Length (NAPL)	Negative
CC & EC	Average Clustering Coefficient	Negative
	Normalized Average Path Length	Negative
	Average Degree	Positive

The results indicate that the second most important characteristic for the standalone performance of *BC* and *DC* is the average degree (AD). A decrease in the AD results in higher influence spread for these heuristics on the slightly clustered networks. In other words, the best standalone performances of both heuristics are observed on the slightly clustered networks with low AD.

The second most influential network characteristic on the standalone performance of *CC* and *EC* is found as the normalized average path length (NAPL). These heuristics show better performances on the networks with lower NAPL. The best standalone performances of both heuristics are observed on the slightly clustered networks with lower NAPL.

The results show that the average path length (NAPL) is the third most significant characteristic for the performance of *BC* and *DC*. Compared with the ACC and the AD, it is a less significant characteristic for explaining the influence spread performance of these two heuristics. However, it is still important and a decrease in NAPL results in higher influence spread performances of *BC* and *DC* on the slightly clustered networks.

The third most influential characteristic on the performance of the heuristics *CC* and *EC* are observed as the AD. However, the effect of the AD arises only when the network has a low average ACC. An increase in the AD results in a better performance of the heuristics *CC* and *EC.*

The results highlight that the D (density) and the NDV (normalized degree variance) do not have a significant impact on the performances of all the four centrality-based heuristics. The data has never been split on these two characteristics in any regression tree.

Considering the significant characteristics on the standalone performances of the four heuristics, it can be concluded that all the four heuristics' standalone performances are affected by the same three network characteristics which are ACC, AD and NAPL.

6.2. Competitive Performances of the Centrality-Based Heuristics

The competitive performances of the heuristics *BC*, *DC*, *CC* and *EC* against each other are indicated in Table 6.4 and Table 6.5. In the tables, average influence spread percentages on the 13 real-world network datasets are listed, and overall average competitive performances are indicated at the end of each column.

According to tables, there are many interesting results compared to the results of standalone experiments. For example, it is expected to see that *CC* and *EC* would lose the most from their standalone performances against *BC* since the performance ranking in standalone performances is *BC*>*DC*>*CC*>*EC*. However, according to the results, the heuristics *CC* and *EC* lose the most influence spread against the heuristic *DC* instead of *BC*. The overall average competitive performance of *CC* against *BC* is 18.4 % while it is 17.5 % against *DC*. In addition, for *EC*, it is 16.2 % against *BC* and 14.7% against *DC*.

Interestingly, the most successful heuristic on a dataset in the competitive setup may differ from standalone experiments depending on the opponent type. Besides, the difference between heuristics' influence spread performances may become enlarger compared to the standalone results. For example, the heuristics *BC* and *DC* show their best standalone performances on the same network (*Ego-Twitter*) with 93.7% and 98.9% average influence

spread respectively. However, the heuristic *BC* shows 27.2% competitive performance against the heuristic *DC* on *Ego-Twitter* while the performance of *DC* is 71.2%. In the competitive experiments, influence spread difference between the heuristics *BC* and *DC* is quite large despite the fact that the standalone performance gap is much lower.

Moreover, the heuristic *BC* has showed roughly one point more influence spread than the heuristic *DC* (*BC:* 46.1% vs *DC:* 45.5%) on *socfb-Hamilton* in the standalone experiments, but *DC* has showed approximately six points more performance than *BC* in the competitive experiments (*BC:* 17.3% & *DC:* 26.2%). The same situation occurs also between the heuristics *CC* and *EC*. The standalone performances of these two heuristics are 65.4% and 55.8% respectively on the *rt-voteonedirection*. However, the heuristic *EC* has showed approximately five points more competitive influence against the heuristic *CC* on the same network (30.9% vs 36.0%).

Table 6.4. Average Competitive Performances of the Heuristics on the Datasets for the First Three Heuristic Combinations (%).

	BC vs. DC		***BC vs. CC***		***BC vs. EC***	
Network	**mean(*BC*)**	**mean(*DC*)**	**mean(*BC*)**	**mean(CC)**	**mean(*BC*)**	**mean(*EC*)**
Ego-Facebook	27.3	16.6	30.3	16.5	37.2	6.8
Hamsterster	25.8	22.6	27.4	18.6	30.2	16.9
Email-Eu-Core	25.0	23.9	25.6	22.2	27.1	21.2
Ca-GrQc	30.5	**9.6**	28.6	12.7	31.6	7.5
Wiki-Vote	38.9	33.6	39.2	**26.4**	46.0	**27.5**
Ca-Netsci	23.7	29.0	30.9	14.7	32.9	10.3
Ca-Hepth	25.7	19.8	30.6	14.4	32.2	10.1
Cit-DBLP	**41.8**	29.5	50.5	17.9	53.7	15.8
ia-fb-messages	28.8	27.6	33.8	21.5	33.0	22.2
Ego-Twitter	27.2	**71.2**	**80.1**	**14.3**	**87.6**	**6.6**
rt-voteonedirection	41.0	55.5	73.4	21.0	67.9	25.8
scofb-Hamilton	**17.3**	26.2	**20.7**	20.0	**19.2**	23.2
Email-Univ	27.8	25.1	30.5	18.6	35.2	16.7
Overall Mean	**29.3**	**30.0**	**38.6**	**18.4**	**41.1**	**16.2**

Table 6.5. Average Competitive Performances of the Heuristics on the Datasets for the Second Three Heuristic Combinations (%).

	DC vs. CC		***DC vs. EC***		***CC vs. EC***	
Network	**mean(*DC*)**	**mean(*CC*)**	**mean(*DC*)**	**mean(*EC*)**	**mean(*CC*)**	**mean(*EC*)**
Ego-Facebook	23.9	15.7	31.3	6.2	26.5	8.2
Hamsterster	28.1	15.0	33.0	12.7	27.6	15.7
Email-Eu-Core	27.3	18.3	32.6	15.0	30.7	14.4
Ca-GrQc	**13.5**	21.2	**20.8**	9.0	23.2	**8.0**
Wiki-Vote	48.4	**21.5**	49.4	25.6	36.2	28.7
Ca-Netsci	32.2	17.4	36.1	9.2	22.7	11.7
Ca-Hepth	26.8	14.6	30.7	9.9	23.2	10.8
Cit-DBLP	44.9	15.6	49.2	14.9	32.7	19.0
ia-fb-messages	35.1	20.6	34.9	20.9	25.2	28.4
Ego-Twitter	**84.5**	14.3	**94.0**	**5.3**	**46.7**	13.7
rt-voteonedirection	76.1	20.6	63.5	**33.2**	30.9	**36.0**
scofb-Hamilton	23.3	**14.1**	24.1	13.3	**16.1**	19.7
Email-Univ	29.2	18.3	35.5	15.5	29.9	17.5
Overall Mean	**37.9**	**17.5**	**41.2**	**14.7**	**28.6**	**17.8**

Another example is that *DC* is approximately four points more successful than *BC* on *rt-voteonedirection* (*BC:* 91.5% & *DC:* 95.4%) in the standalone experiment. In the competitive experiment, *DC* has showed approximately 15 points better performance against *BC* (*BC:* 41.0% & *DC:* 55.5%).

These interesting results of competitive experiments support the necessity for extending this study to the analysis of the relation between the competitive performances of the heuristics and the network characteristics. In the next subsections, this analysis is detailed based on each heuristic as in the standalone experiment results.

6.2.1. Betweenness Centrality

The results with the competitive performance of the heuristic *BC* showed that the overall average performance of *BC* is 36% in paired competitions with *DC*, *CC* and *EC* over all the datasets. The average standalone performance of *BC* has been observed as 58%. Meaning that, *BC* loses 37.9% on average from its performance due to the presence of an opponent in the network regardless of the opponent type. Considering the individual experiment results, the best competitive performance of *BC* is observed as 89% influence

spread against *CC* on *rt-voteonedirection*, and the worst performance is seen again on *rt-voteonedirection* with 5.7% influence spread against *DC*.

The AD (average degree) is found as the most significant characteristic for the competitive performance of *BC*. The first split point on AD separates of 85% of the data with 31% average influence spread from 15% of the data with 63% average influence spread. It concludes that *BC* competes better on the networks with lower AD regardless of the opponent type.

The second most influential characteristic is observed as the ACC (average clustering coefficient). It separates 23% of the data with 41% average influence spread from the 62% of the data with 28% average influence spread. The heuristic *BC* performs better against its three opponents on the slightly clustered networks.

According to the right branch of the Figure 6.5, the heuristic *DC* dramatically worsen the performance of *BC* on the networks with lower AD. Meaning that, the significance of the opponent type arises only when the AD of the network decreases.

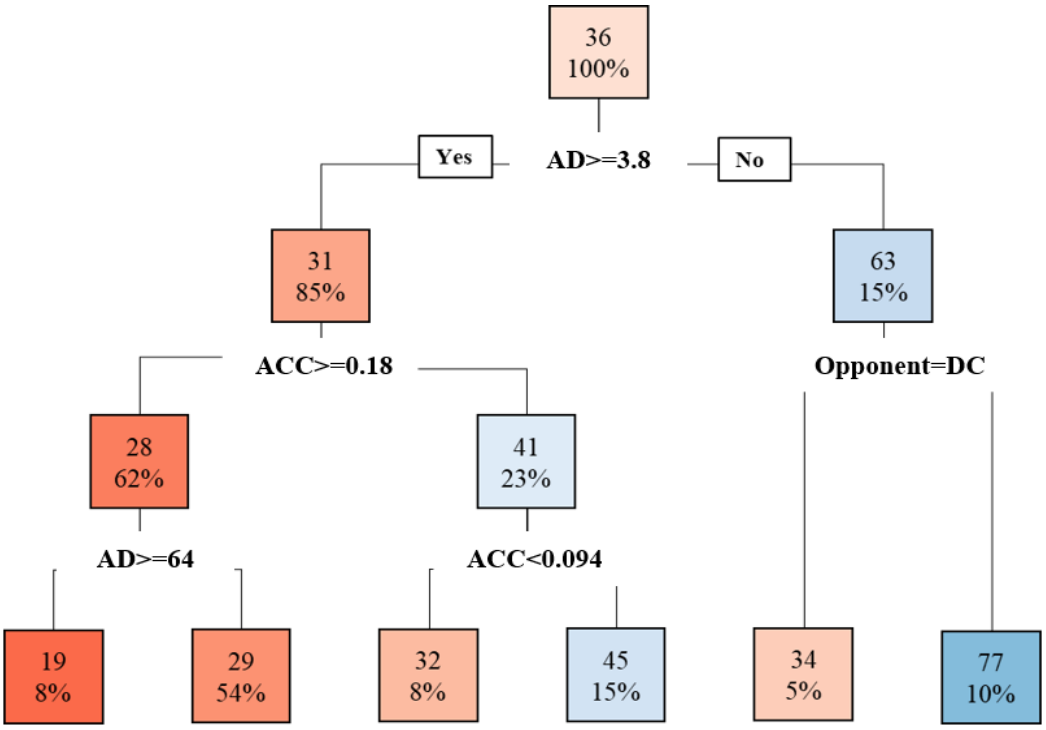

Figure 6.5. Regression Tree of *BC*'s Competitive Performance.

There is no split point on the NDV (normalized degree variance), the NAPL (normalized average path length) and the D (density). Hence, these three characteristics do not affect the performance of *BC* against its opponents as much as the AD and the ACC.

To sum up, the AD and the ACC are the most influential characteristics on the performance of *BC* regardless of the opponent type. The heuristic *BC* is robust to the opponent type since there is only one split point depending on the opponents in the regression three and the percentage of the data on that split point is only 15%. Considering the right branch in Figure 5.5, it can be concluded that the performance of *BC* could worsen significantly by only the heuristic *DC* on the networks with higher average degree.

6.2.2. Degree Centrality

The results with the competitive performance of the heuristic *DC* showed that the overall average performance of *DC* is 36% in paired competitions with *BC, CC* and *EC* over all the datasets. The average standalone performance of *DC* has been observed 56%. Meaning that, DC loses 35.7% on average from its performance due to the presence of an opponent in the network regardless of the opponent type. Considering the individual experiment results, the best performance of *DC* is observed as 94.1% influence spread against *EC* on *Ego-Twitter*, and the worst performance is seen on *Ca-GrQc* as 8.6% influence spread against *BC*.

The leftmost split point in Figure 6.6 shows that the NAPL (normalized average path length) splits 62% of the data, which means that the NAPL is one the most significant characteristics on the performance of the heuristic *DC* against its opponents. If the network is highly clustered and has a high AD, an increase in the NAPL becomes advantageous for *DC.*

The first split point on the right branch splits equally only 15% of the data into two groups, which means that the NDV (normalized degree variance) has not a significant effect on the performance of the heuristic *DC.* Besides, there is no split point depending on the D (density) in the regression tree. Meaning that, the performance of *DC* does not change significantly depending on a decrease or an increase in the density of a network.

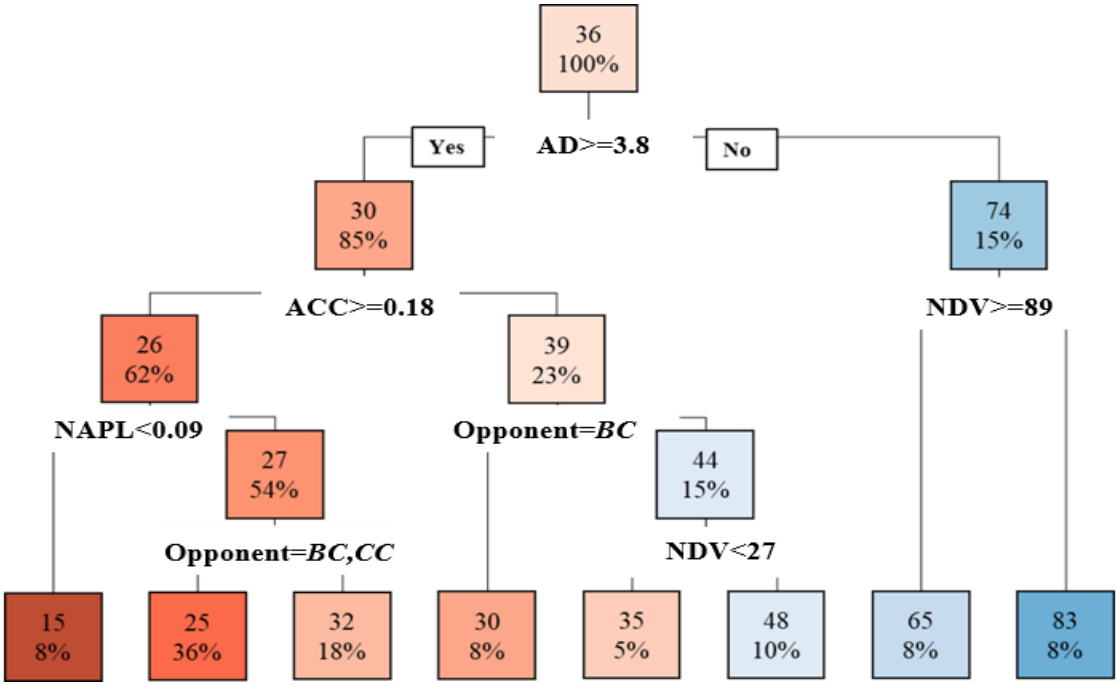

Figure 6.6. Regression Tree of *DC*'s Competitive Performance.

When all the split points in Figure 6.6 are evaluated, it is seen that the opponent type affects the performance of *DC* only on the highly clustered networks with high AD and NAPL (see the left branch of Figure 6.6). Considering such networks, it can be concluded that the heuristics *BC* and *CC* reduce considerably the influence spread of the heuristic *DC*.

There is no split point depending on the D (density) in the regression tree above. Meaning that, the performance of *DC* does not change according to a decrease or an increase in the density of a network.

Overall, the most significant characteristics on the competitive performance of *DC* are the AD and the ACC of the networks. An increase in the NAPL becomes influential when the network is highly clustered and has a high AD. The opponent type is observed as less significant on the performance of *DC* than the network characteristics. Lastly, the D and the NDV have not showed a significant effect as in the standalone performance results.

6.2.3. Closeness Centrality

The results on the competitive performance of the heuristic *CC* showed that the overall average performance of *CC* is 21% in paired competitions with *BC*, *DC* and *EC* over all the datasets. The average standalone performance of *CC* has been observed 46%. Meaning that, *CC* loses 54.3% on average from its performance due to the presence of an opponent in the network regardless of the opponent type. Considering the individual experiment results, the best performance of *CC* is observed as 56.1% influence spread against *EC* on rt-voteonedirection, and the worst performance is seen on again rt-voteonedirection with 5.9% influence spread against *DC*.

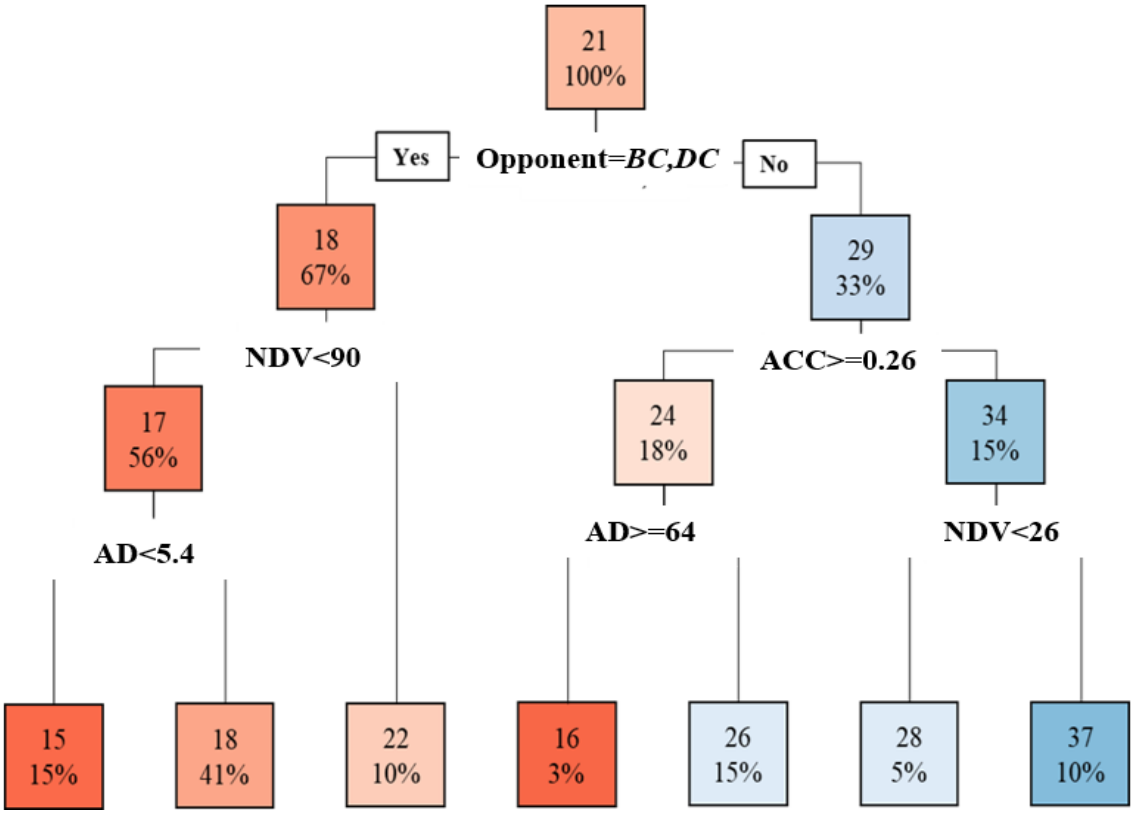

Figure 6.7. Regression Tree of *CC*'s Competitive Performance.

Referring Figure 6.7, the heuristic *CC* is the mostly affected heuristic by the opponent type since the topmost split point separates off 67% of the data with 18% average propagation from 33% of the data with 29% average propagation rate. According to this split

point, *CC* performs considerably better when the opponent is *EC* than the case that the opponent is *BC* or *DC*.

The first split point of the right branch showed that if the competition is with *EC*, the ACC (average clustering coefficient) of a network is the most influential characteristic for the competitive performance of *CC.* The heuristic *CC* propagates better against *EC* on the networks with low ACC. The other split points on the right branch are not taken into consideration because they separate a little amount of the data. Therefore, the most significant characteristic for the competitive performance of *CC* is the ACC when its competitor is the heuristic *EC.*

The first split point on the left branch separates 56% of the data with 17% average influence spread from the 10% of the data with 22% average influence spread based on the NDV (normalized degree variance) of the networks. As the NDV increases, the competitive performance of *CC* against the heuristics *BC* and *DC* also increases. Apart from the NDV, the regression tree points out that the AD (average degree) as the second most significant characteristic when the competitor of *CC* is *BC* or *DC*. As the AD of a network increases, the competitive performance of *CC* against the heuristics *BC* and *DC* increases also. It is mentioned in the previous sections that both of the heuristics *BC* and *DC* significantly perform better against their competitors on the networks with a lower AD. Hence, an increase in the AD becomes advantageous for the heuristic *CC* when the competitor of it is *BC* or *DC.*

According to Figure 6.7, the data has never been split on the NAPL (normalized average path length) and the D (density). Therefore, the results show that the performance of the heuristic CC does not change depending on the NAPL and the D of the networks.

To sum up, the type of the opponent is found as more influential on the performance of *CC* than the effect of the network characteristics. The heuristics *BC* and *DC* dramatically prevents *CC* from reaching higher influence spread in a network. Considering the network characteristics, a decrease in the ACC of a network is found as advantageous for the performance of *CC* against *EC*. Besides, the results show that *CC* competes better against

BC and *DC* on the networks with higher AD and higher NDV, but it still reaches less influence spread than *BC* and *DC* in paired competitions with them.

6.2.4. Eigenvector Centrality

The results with the competitive performance of the heuristic *EC* showed that the overall average performance of *EC* is 16% in paired competitions with *BC*, *DC* and *CC* over all the datasets. The average standalone performance of *EC* has been observed 39%. Meaning that, *EC* loses 58.9% on average from its performance due to the presence of an opponent in the network regardless of the opponent type. Considering the individual experiment results, the best performance of *EC* is observed as 53.18% influence spread against *BC* on *rt-voteonedirection* and the worst performance is seen on *Ego-Facebook* with 5.13% influence spread against *DC*.

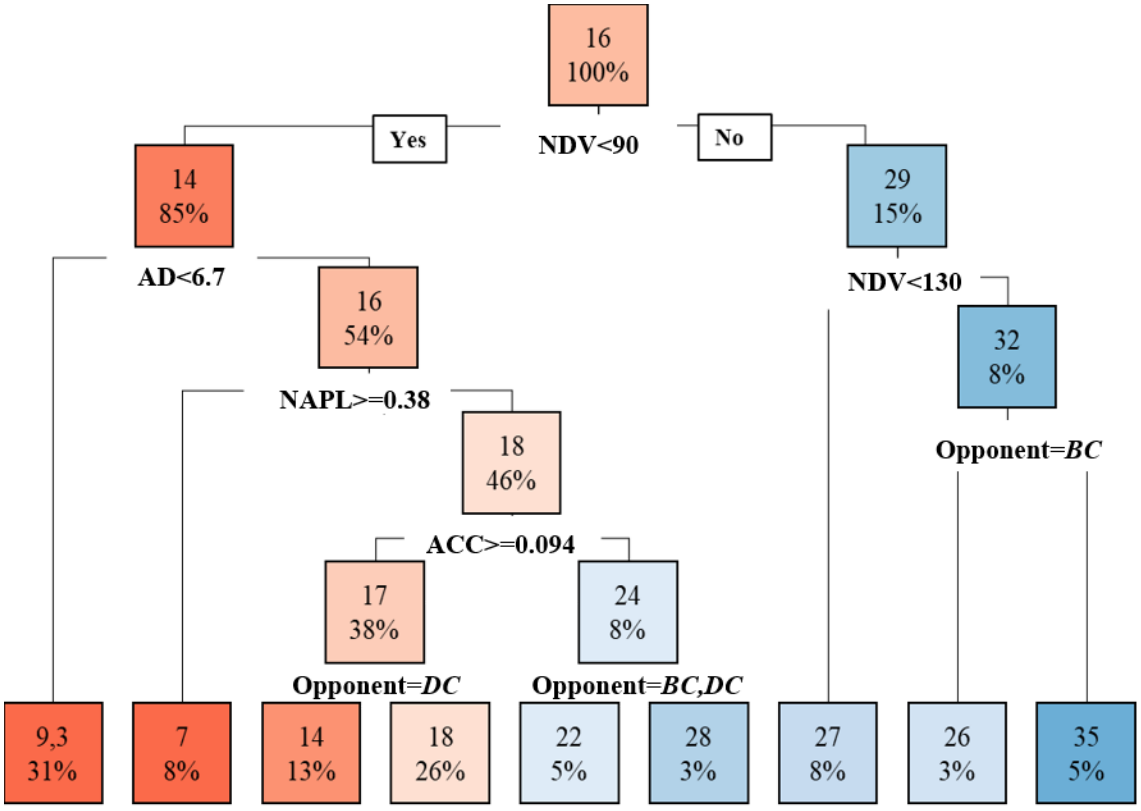

Figure 6.8. Regression Tree of *EC*'s Competitive Performance.

According to Figure 6.8, the topmost split point separates off 85% of the data with 14% average influence spread from 15% of the data with 29% average influence spread depending on the NDV (normalized degree variance) of the datasets. Meaning that, the competitive performance of *EC* increases regardless of the opponent type as the NDV of the networks increases.

From the first split point on the left branch, the AD (average degree) is the second most significant characteristic for the competitive performance of *EC*. The AD separates 31% of the data with 9.3% average influence spread from the 54% of the data with 16% average influence spread. As the AD of the networks increases, the competitive performance of *EC* increases too.

The other influential characteristics on the competitive performance of *EC* are the NAPL (normalized average path length) and the ACC (average clustering coefficient). The heuristic *EC* performs better on the networks with a lower NAPL and a lower ACC regardless of the competitor type.

The data is split depending on the all network characteristics except the D (density). As in the results of the heuristics *BC*, *DC* and *CC*, the results show that the D has also not a significant effect on the performance of *EC* against its competitors.

To sum up, the heuristic *EC* performs the worst competitive performance among all of the four heuristics. The NDV and the AD are the most influential characteristics on its competitive performance and an increase in their values increases the competitive performance of *EC* as well. While the ACC and the NAPL affect the competitive performance of the heuristic *EC*, D has not showed a significant effect on its performance. Besides, the opponent type is not significant for the competitive performance of *EC* since the heuristic EC usually lose much from its influence regardless of the opponent type.

6.2.5. Overview of the Competitive Performance Results

Before proceeding to Section VII for Discussion, the findings of competitive experiments are listed in a table and summarized below. The influential network

characteristics and changes in the effects of the characteristics from standalone to competitive results are noted here, but the underlying reasons will be discussed in Section VII. In the table, positive direction of correlation represents that as the value of network characteristic increases, the performance of the heuristic tends to also increase. If the direction is negative, the performance of the heuristic decreases with an increase in the network characteristic. Since the opponent type is a categorical variable, direction of correlation is not indicated for it.

Table 6.6. Significant Effects on the Competitive Performances of the Heuristics.

	Significant Effects in Order	**Direction of correlation**
BC	Average degree (AD)	Negative
	Average Clustering Coefficient (ACC)	Negative
	Opponent type	None
DC	Average degree	Negative
	Average Clustering Coefficient	Negative
	Normalized Average Path Length (NAPL)	Negative
	Opponent type	None
CC	Opponent type	None
	Normalized Degree Variance (NDV)	Positive
	Average Clustering Coefficient	Negative
	Average degree	Positive
EC	Normalized Degree Variance (NDV)	Positive
	Average Degree	Positive
	Normalized Average Path Length	Negative
	Average Clustering Coefficient	Negative

It is found that AD and ACC are the most important characteristics on the competitive performance of the heuristics *BC* and *DC*. Both of the heuristics accomplish higher influence spreads on the slightly clustered networks with lower AD. Besides that, the results show that

the AD becomes more important than the ACC for the performance of *BC* and *DC* when there is another influence propagating in the network. On the contrary, in the standalone experiments, ACC has been observed on the most influential characteristic on the performances of the heuristics.

Considering the heuristics *BC* and *DC*, it is observed that the opponent type does not affect the performance of the heuristics *BC* and *DC* as much as AD and ACC. Hence, it can be said that the heuristics *BC* and *DC* are less sensitive to the opponent type compared to their sensitivity to the network characteristics. Only exception is the case that they compete against each other. When the two heuristics compete against each other on a network with low AD and ACC, the heuristic *DC* outperforms the heuristic *BC*.

The NAPL is found as another influential characteristic on the competitive performance of *DC*. If the heuristic *DC* is used on a highly clustered network with high AD, a decrease in the NAPL is advantageous for the performance of *DC*.

The opponent type is found as the most significant characteristic on competitive performance of the heuristic *CC*. Considering the all the four heuristics, the heuristic *CC* is the only one that the effect of the opponent type on its competitive performance is observed as more important than the effect of the network characteristics. When *CC* competes against *BC* or *DC* can spread eleven points less on average compared to the case that it competes against *EC* (see Figure 6.7).

Additionally, it is found that the heuristic *CC* competes better on the networks with high NDV, high average degree and low ACC regardless of the opponent type.

The NDV is observed as the most significant characteristic on the competitive performance of the heuristic *EC*. If the heuristic *EC* competes on a highly variant network in terms of node degrees, it can reach 15% more influence on average compared to other networks (see Figure 6.8). In addition to this, it seen that *EC* competes better on the networks with NDV, a decrease in the NAPL and the ACC and an increase in the average degree are advantageous for the performance of it.

The opponent type is not found as an important characteristic on the competitive performance of *EC*. The heuristic *EC* is significantly affected by the presence of another competing influence on the network regardless of the opponent type.

7. DISCUSSION

In this thesis, firstly, the four centrality-based heuristics are compared with *Random (R)* seed selection heuristic in terms of their standalone performances. The results indicate that all the four centrality-based heuristics outperform the performance of the heuristic *R.* This result contradicts the claim of Zhao *et al.* (2017) that the performance of the heuristic *EC* is worse than *R.* The authors have tested the standalone performances of the heuristics *EC* and *R* on the two real-world datasets under a diffusion model inspired by the game theory perspective (Zhao *et al.*, 2017). However, in this thesis, the standalone performances of the heuristic *EC* and *R* are tested on the thirteen real-world datasets and under an extension of Linear Threshold Model. The results show that the heuristic *EC* outperforms *R* on the eight networks among all the datasets. Thus, this study shows that the performances of the seed selection heuristics are very dependent on the diffusion model and the network topology.

Table 7.1. Ranks of the Four Centrality-Based Heuristics According to Standalone Performances.

	N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**
BC	1	1	1	1	1	2	1	1	1	2	2	1	1
DC	2	2	2	3	2	1	2	2	2	1	1	2	2
CC	3	3	3	2	3	3	3	3	4	3	3	4	3
EC	4	4	4	4	4	4	4	4	3	4	4	3	4

Considering the average influence spreads in the standalone experiments, Table 7.1 is created. In the table, "1" represents the most successful heuristic on the network of interest in terms of average standalone performance, while "4" represents the heuristic with the worst average standalone performance. It is seen from the table that there is a general ranking of the four centrality-based heuristics: *BC>DC>CC>EC.* Although there is a general ranking, the results also indicate that the most successful heuristic changes from one network to another network. For example, Table 7.1 shows that the best performance on the three networks among all the datasets belongs to the heuristic *DC* instead of *BC.* Additionally, the heuristic *CC* outperforms the heuristic *DC* on the network N4 (*Ca-GrQc*), and the heuristic

EC is better than the heuristic *CC* on the networks N9 (*ia-fb-messages*) and N12 (*rt-voteonedirection*).

After completion of the competitive experiments, the performances of the four heuristics on the thirteen different datasets are compared. For the comparison, the average influence spreads (average of 40 replication) of a heuristic against another heuristic on a dataset are listed as in Figure 7.1. Then, overall average of each row for every network is calculated. For example, the average influence spread of the heuristic *BC* against its three opponents on *Ego-Facebook* is calculated as 31.6% by taking the averages of 27.3%, 30.3% and 37.2% (see the first row of *Ego-Facebook* in Figure 7.1). According to these overall averages, the competitive performances of the four heuristics are ranked for each network dataset in Table 7.2. In the table, "1" represents again the most successful heuristic on the network of interest in terms of average competitive performance, while "4" represents the heuristic with the worst performance. Referring Table 7.2, it is seen that there is also a general ranking of the four centrality-based heuristics in terms of their competitive results: *DC>BC>CC>EC.* Therefore, it is concluded that the performance ranking of the heuristics differ in the standalone and the competitive experiments, which shows that performance of a heuristic is sensitive to whether there is an opponent or not.

Table 7.2. Ranks of the Four Centrality-Based Heuristics According to Competitive Performances.

	N1	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**
BC	1	2	2	1	2	2	1	1	2	2	2	2	1
DC	2	1	1	3	1	1	2	2	1	1	1	1	2
CC	3	3	3	2	3	3	3	3	4	3	4	4	3
EC	4	4	4	4	4	4	4	4	3	4	3	3	4

Although the heuristic *DC* outperforms the heuristic *BC* according to the competitive performance ranking, there are only four networks (see *Ca-Netsci, Ego-Twitter, rt-voteonedirection* and *socfb-Hamilton* results in Figure 7.1) that the heuristic *DC* accomplishes higher influence spread than *BC* when they compete on the same network. The reason for this contradiction is that the heuristic *DC* competes better than *BC* against the heuristic *CC* and the heuristic *EC.* Thus, it is concluded that using *DC* against *CC* and *EC* is

more strategical than using *BC* against *CC* and *EC*. This finding contributes a clear understanding of the fact that the performance of a heuristic can differentiate according to the opponent type.

The findings so far highlight that the performances of the four centrality-based heuristics differ from one network to another network in both standalone and competitive experiments. Apart from the main aim of this thesis, one of the secondary objectives is set on clarifying whether the importance of network characteristics effects differ from the standalone case to the competitive case. Therefore, the effects of the characteristics are analyzed firstly for the standalone performances of the heuristics.

The results with the standalone performances show that the three network characteristics among all the five characteristic are influential on the performances of the heuristics: The ACC (average clustering coefficient), AD (the average degree), and the NAPL (normalized average path length). The most influential one is found as the ACC for each centrality-based heuristic. The second most influential characteristic is observed as the AD for the heuristics *BC* and *DC* while it is observed as the NAPL for the heuristics *CC* and *EC*. The NDV (normalized degree variance) and the D (density) are observed as the insignificant characteristics for the standalone performances of the heuristics.

Considering the standalone case, the results indicate that all the heuristics perform better on the slightly clustered networks. This finding is line with the study of Hussain *et al.* (2013). The authors concluded that the heuristics *BC, DC, CC* and *EC* perform better on scale-free and hybrid networks of small-world and random networks under the Linear Threshold Model (Hussain *et al.*, 2013), which are classified as the slightly clustered networks due to their low ACC. Besides, in order to obtain the best performances from the heuristic *BC* and *DC*, the results reveal that they should be used on the slightly clustered networks with low AD. Contrary to this, an increase in AD is found as advantageous for the performances of *CC* and *EC*. Peres (2014) showed that the AD has a positive and significant effect on the information diffusion process. The author also concluded that the ratio between the average degree of the top most 10% most connected nodes has also a positive and significant impact on the information diffusion regardless of the seed selection approach (Peres, 2014). Meaning that, since this ratio decreases probably when the AD of the network

increases, the negative impact of the AD arises for the heuristic *DC.* The heuristic *DC* finds the most connected nodes in the network. When the AD of the network is also high, selecting the seeds according to their degree centralities becomes ineffective. Besides, Valente *et al.* (2008) studied the correlation between the betweenness and degree centralities of the nodes. They concluded that betweenness and degree centralities of the nodes are highly correlated (Valente *et al.*, 2008). According to this correlation, it can be interpreted that *BC* also selects the most connected nodes for initial seed as well. the Thus, an increase in AD has also a negative effect on *BC* by reducing the AD of social hubs (top most connected nodes) in the network.

Additionally, the thesis confirms that all the heuristics perform better with a decrease in the NAPL (normalized average path length). The NAPL represents the required steps on average to reach a node in information diffusion (Chen *et al.*, 2008). Therefore, it might be expected that a decrease in the NAPL has a positive effect on all the four heuristics.

Referring the competitive experiment results, it is concluded that the presence of another influence on the network dramatically reduces the average influence spread of all the four heuristics. For each heuristic, the percentage of loss is calculated based on the proportion of average influence spread percentages' difference between standalone and competitive experiments to average influence spread percentage in standalone experiment. The heuristic *DC* is observed as the best heuristic in terms of losing the least influence from its standalone performance in the competitive environment (-35.7%). The heuristic *DC* is followed by *BC* (-37.9%), *CC* (-54.3%) and *EC* (-58.9%) respectively. Meaning that, the heuristic *DC* is the most robust one to the effect of the presence of another influence on the network while the heuristic *EC* is the most sensitive heuristic.

According to Table 7.3, the only thing that changes from the standalone case to competitive environment is not the average influence spreads of the heuristics. Besides that, the effects of the network characteristics on the heuristics' performances change, which can be grouped as: Firstly, the importance ranking of the effective characteristics differentiate. Secondly, a significant characteristic in standalone case becomes insignificant. Thirdly, the effect of an insignificant characteristic in the standalone environment arises in the competitive environment. Thus, this thesis reveals that the effect of network characteristics

on the influence maximization should be investigated by considering the reality that there are more than one influences spreading on the networks unlike the existing literature (e.g. Barthelemy *et al.*, 2004; Hussain *et al.*, 2013; Peres, 2014; Liu and Hong, 2018).

N2: Hamsterster (%)

	BC	***DC***	***CC***	***EC***
BC		25.8	27.4	30.2
DC	22.6		28.1	33.0
CC	18.6	15.0		27.6
EC	16.9	12.7	15.7	

N3: Email-Eu-Core (%)

	BC	***DC***	***CC***	***EC***
BC		25.0	25.6	27.1
DC	23.9		27.4	32.6
CC	22.2	18.3		30.7
EC	21.2	15.0	14.4	

N4:Ca-GrQc (%)

	BC	***DC***	***CC***	***EC***
BC		30.5	28.6	31.6
DC	9.6		13.5	20.8
CC	12.7	21.3		23.2
EC	7.5	9.0	8.0	

N5:Wiki-Vote (%)

	BC	***DC***	***CC***	***EC***
BC		38.9	39.2	46.04
DC	33.6		48.4	49.4
CC	26.4	21.5		36.2
EC	27.5	25.7	28.7	

N6:Ca-Netsci (%)

	BC	***DC***	***CC***	***EC***
BC		23.7	30.9	32.9
DC	29.0		32.2	36.1
CC	14.7	17.4		22.7
EC	10.3	9.2	11.7	

N7:Ca-Hepth (%)

	BC	***DC***	***CC***	***EC***
BC		25.7	30.6	32.2
DC	19.8		26.8	30.7
CC	14.4	14.6		23.2
EC	10.1	9.9	10.8	

N9: ia-fb messages (%)

	BC	***DC***	***CC***	***EC***
BC		28.8	33.8	33.0
DC	27.6		35.1	34.9
CC	21.5	20.6		25.2
EC	22.2	20.9	28.4	

N10: Ego-Twitter (%)

	BC	***DC***	***CC***	***EC***
BC		27.2	80.1	87.6
DC	71.2		84.5	94.0
CC	14.3	14.3		46.7
EC	6.6	5.3	13.7	

N11: rt-voteonedirection (%)

	BC	***DC***	***CC***	***EC***
BC		41.0	73.4	67.8
DC	55.5		76.1	63.5
CC	21.0	20.6		30.9
EC	25.8	33.2	36.0	

N12: socfb-Hamilton (%)

	BC	***DC***	***CC***	***EC***
BC		17.3	20.7	19.2
DC	26.2		23.3	24.1
CC	20.0	14.1		16.1
EC	23.2	13.3	19.7	

N13: Email-Univ (%)

	BC	***DC***	***CC***	***EC***
BC		27.8	30.5	35.2
DC	25.1		29.2	35.5
CC	18.6	18.3		29.9
EC	16.7	15.5	17.5	

Figure 7.1. Average Influence Spreads of the Heuristics Against Each Other on the Network Datasets.

Table 7.3. Summary Table of the Ranked Significant Effects on the Standalone and Competitive Performances of the Heuristics.

	Characteristics with Significant Effects	
	Standalone Case	**Competitive Case**
BC	Average Clustering Coefficient Average Degree Normalized Average Path Length	Average Degree Average Clustering Coefficient Opponent type
DC	Average Clustering Coefficient Average Degree Normalized Average Path Length	Average Degree Average Clustering Coefficient Normalized Average Path Length Opponent type
CC	Average Clustering Coefficient Normalized Average Path Length Average Degree	Opponent type Normalized Degree Variance Average Clustering Coefficient Average degree
EC	Average Clustering Coefficient Normalized Average Path Length Average Degree	Normalized Degree Variance Average Degree Normalized Average Path Length Average Clustering Coefficient

To evaluate the differentiated network characteristic effects and performance results between the standalone and the competitive environments, the heuristics *BC* and *DC* are firstly evaluated. Although *BC* and *DC* lose dramatically from their standalone performances in a competitive environment, the opponent type is found as less influential than network characteristics. The most influential characteristic on the competitive performance of *BC* and *DC* is observed as the AD. In Figures 7.2, 7.3, 7.4 and 7.5, influence spread performances of the heuristics in standalone and competitive experiments are indicated consecutively in the box plots to compare them easily. In the box plots above standalone performances are shown, and in the box plots below competitive performances are shown. The box plots represent the distribution of the final influence spreads in replications. The mark in the middle of each box plot represents the average influence spread of 40 replications in the network of interest. When Figure 7.2 and 7.3 are examined, it is seen that the largest ranges

belong to the competitive performance box plots of N10 (*Ego-Twitter*) and N11 (*rt-voteonedirection*). These networks have the lowest average degrees among the thirteen datasets. Besides that, the skewness pattern of the N10 and N11 competitive box plots of *DC* is seen as symmetrical while the pattern for *BC*'s box plots is asymmetrical. Meaning that, the heuristic *BC'* s performance fluctuates with different opponent types on the networks with low AD while *DC* performs steadily against its opponents.

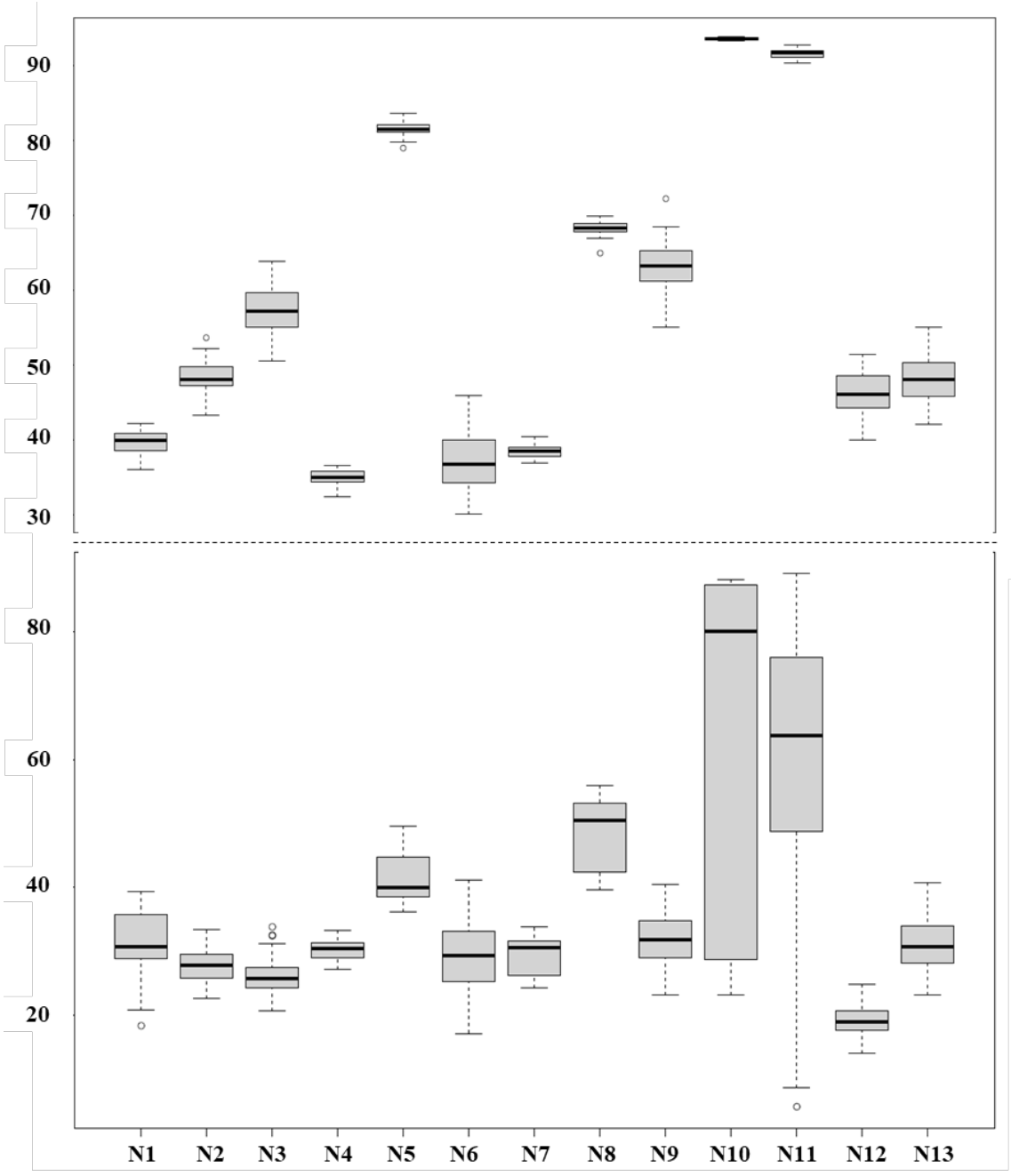

Figure 7.2. Box Plots of the Standalone & Competitive Performances of the Heuristic *BC.*

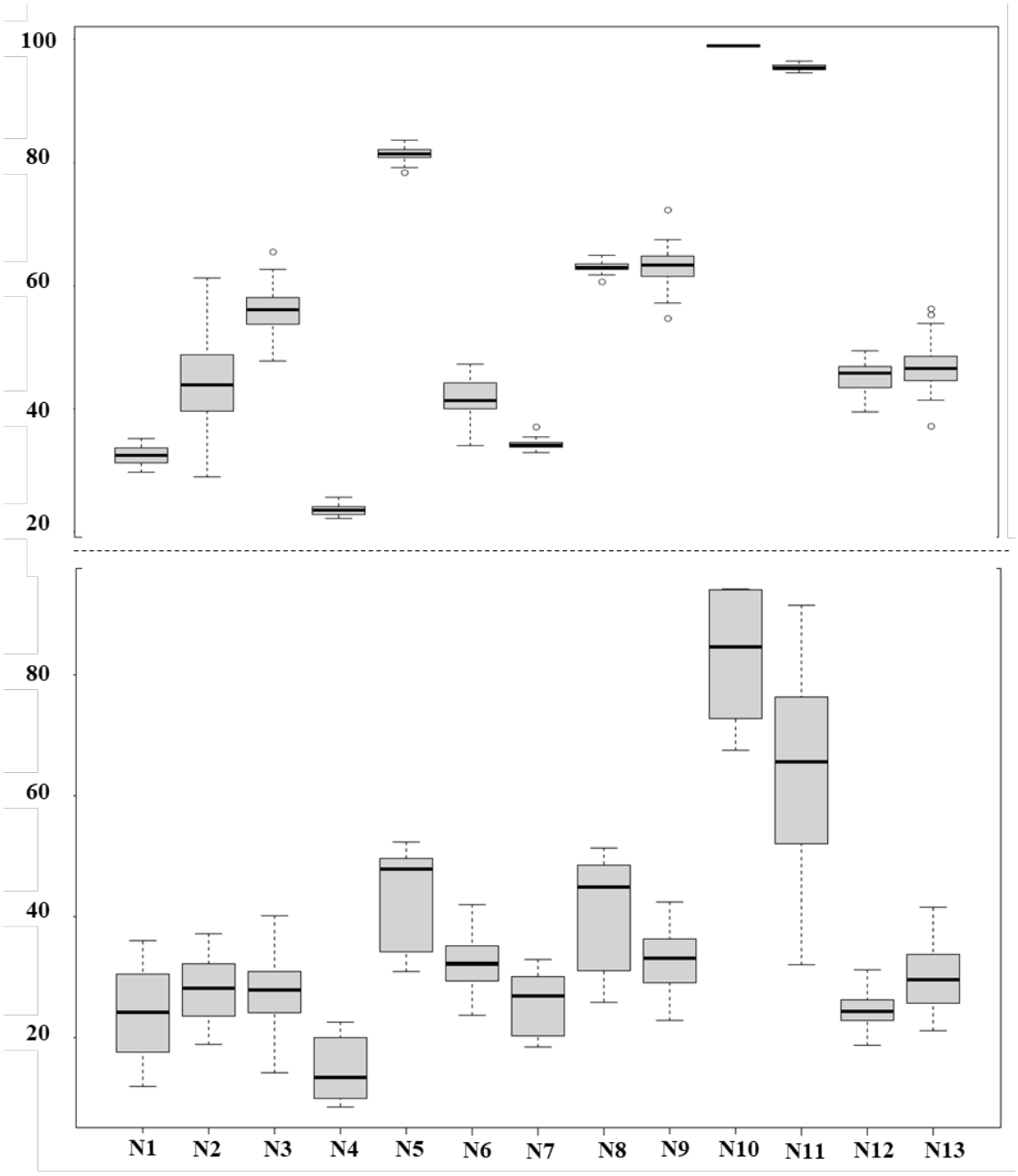

Figure 7.3. Box Plots of the Standalone & Competitive Performances of the Heuristic *DC*.

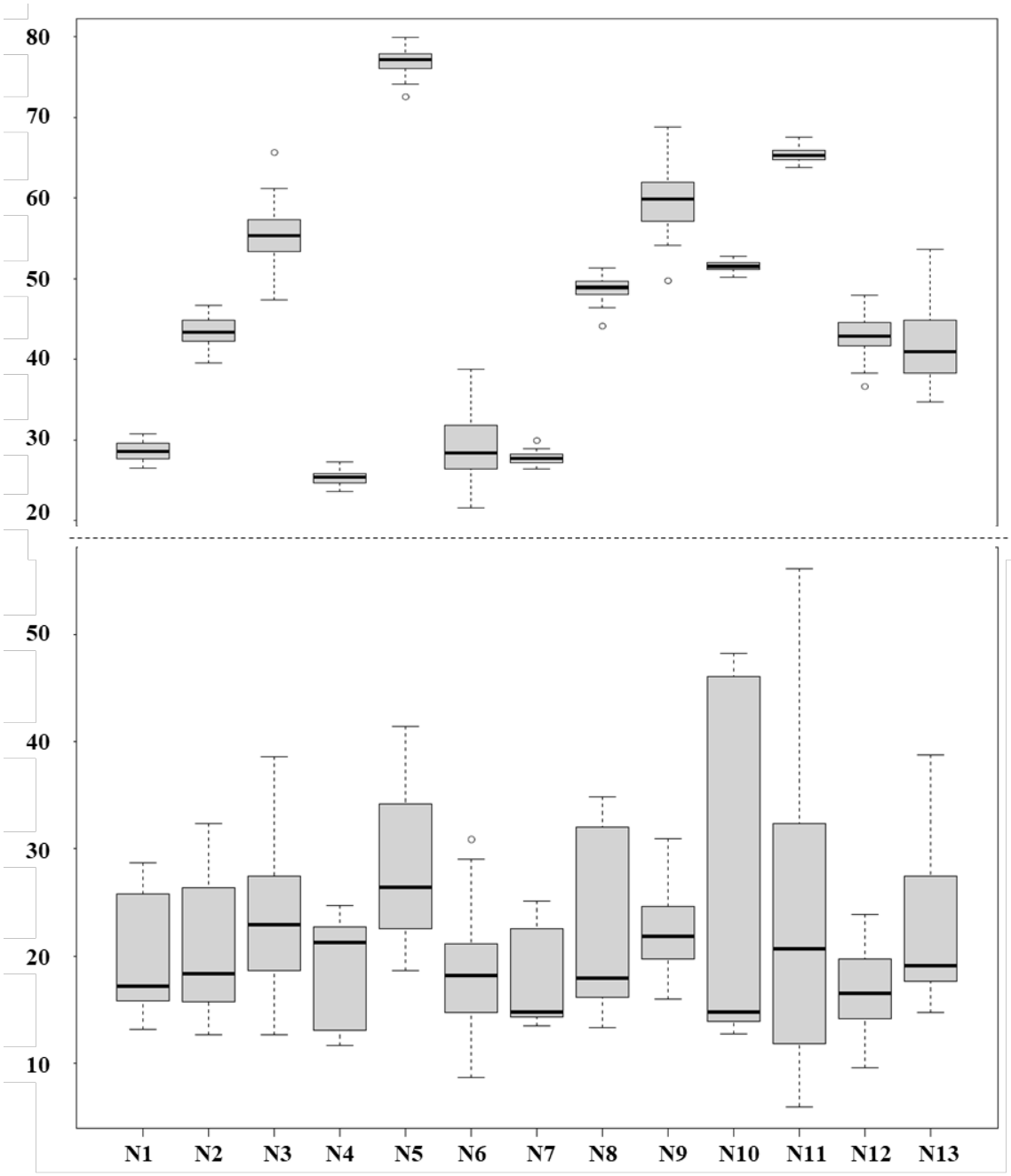

Figure 7.4. Box Plots of the Standalone & Competitive Performances of the Heuristic *CC*.

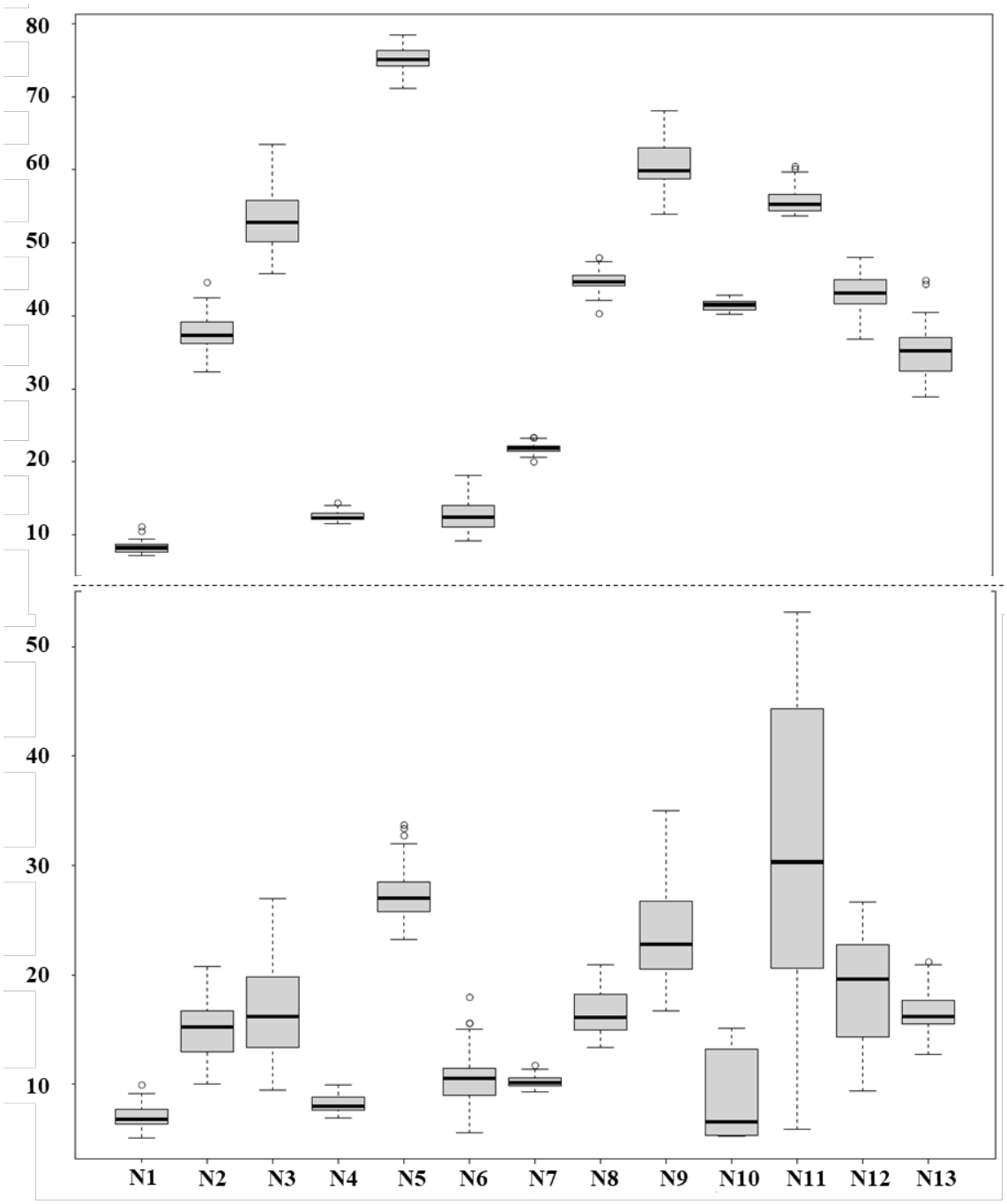

Figure 7.5. Box Plots of the Standalone & Competitive Performances of the Heuristic *EC*.

The results also demonstrate that the effect of NAPL on the performance of *BC* disappears in the competitive environment while it is still important for the performance of *DC*. The betweenness centrality is a path-based centrality metric representing the ratio of the shortest paths through the node of interest among all the shortest paths in the network (Lü *et al.*, 2016). Therefore, using *BC* is even more strategical in the networks with higher NAPL.

Although a decrease in NAPL promotes higher information diffusion regardless of the seed selection heuristic, changes in the NAPL become insignificant for the performance of *BC* against its opponents due to its nature controlling easily the path lengths in a network. Besides that, the NAPL is still significant for *DC*, since *BC* limits *DC's* propagation on the networks with higher NAPL (see Figure 6.6).

Considering the changes in the significant effects on the standalone and competitive performance of *CC*, the heuristic *CC* is found as the only heuristic affected more by the opponent type than the network characteristics. When Figure 7.4 examined, it is seen that the ranges of the competitive performance box plots become quite larger compared to the standalone performance box plots regardless of the datasets. In Figures 7.2, 7.3 and 7.5, it is not possible to see such a change in the ranges of the box plots for each dataset. Meaning that, the heuristic *CC* is very sensitive to the seed selection strategy of the opponent regardless of the network characteristics.

As it is stated in Section 6.2.4, the opponent type is not found as a significant characteristic on the competitive performance of the heuristic *EC.* Although *EC* is the one that loses the most from its standalone performance due to the existence of another influence in the network, the amount of the loss does not change depending on the opponent type. When Figure 7.5 is compared with the Figure 7.2, Figure 7.3 and Figure 7.4, it can be easily seen that the least difference between the ranges of standalone performance box plots and the competitive box plots belongs to the heuristic *EC.* This observation also supports the finding that *EC's* performance against its opponents is not affected by the opponent type.

Considering the results with the both heuristics *CC* and *EC*, the effect of the NDV (normalized degree variance) becomes the most significant network characteristic on these two heuristics' competitive performances. In the standalone performances of all heuristics, it has been observed that the NDV has no significant effect on the performance of any heuristic. When Figure 7.4 and Figure 7.5 are examined, the best competitive performances of both *CC* and *EC* are seen on N5 (*Wiki-Vote*) and N11 (*rt-voteonedirection*) which have the highest NDV among all the datasets. In addition to *CC* and *EC*, according to Figure 7.2. and 7.3, the datasets N5 and N11 are also among the networks where *BC* and *DC* perform

better. Moreover, the performances of *BC* and *DC* outperform the performances of *CC* and *EC* on these networks.

High NDV of a network indicates that there are many social hubs in the network. Since *CC* and *EC* propagate limitedly against *BC* and *DC,* only social hubs promote better performances of them against the superiorities of *BC* and *DC.* Thus, the NDV becomes the most influential characteristic on the competitive performance of *CC* and *EC*. Kim *et al.* (2017) highlighted that social hubs foster information diffusion regardless of seeding approach, but the roles of social hubs in the network are also important. The authors concluded that social hubs playing a bridging role in the network are more influential than the hubs playing crucial position role. Social hubs with bridging roles are found based on their betweenness centralities, while hubs with crucial positions are found based on their other centrality metrics such as eigenvector and closeness centrality (Kim *et al.*, 2017). Besides that, degree centralities and betweenness centralities of the nodes in a network are generally correlated (Valente *et al.*, 2008). Meaning that, social hubs with bridging roles may also have high degree centralities. This fact explains why *BC* and *DC* outperform *CC* and *EC* on highly variant networks in terms of degrees despite the fact that the NDV has positive and important effect on the performance of *CC* and *EC*.

Apart from the NDV, the AD and the ACC have found as influential on the competitive performances of *CC* and *EC* as in their standalone performance results. An increase in AD and a decrease in ACC result in higher influence spreads for both *CC* and *EC.* As a difference between standalone and competitive results, the effect of NAPL has disappeared for *CC* while a decrease in NAPL has still positive effect on *EC* in the competitive results. The reason behind this can be explained as follows: Closeness centrality is a path-based centrality metric, which measures the average distance of a node to all other nodes in a network (Cohen *et al*., 2014). Moreover, a decrease in NAPL has a positive effect on the performance of *EC* and other heuristics as it is stated before. To sum up, *CC* is more robust to changes in the NAPL of the networks due to its nature, while a change in NAPL means that losing one of the limited advantages against its opponents for *EC*.

Lastly, D (the density) has not showed a significant effect on the standalone or the competitive performance of any heuristic. Katona *et al.* (2011) highlighted that the D of

relationships among active nodes of potential inactive nodes may affect positively the ultimate total number of the active nodes (Katona *et al.*, 2011). Considering the same perspective, it can be concluded that if the ties between inactive nodes denser than the ties between the inactive nodes and initially activated seeds, D may affect negatively the information diffusion. Thus, D creates negative and positive effects on information diffusion simultaneously. Therefore, the results indicate that global D has not a significant effect on the performances of the heuristics, and the effect of density should be investigated through the densities of clusters, which include initial seeds.

8. CONCLUSION

This study focuses on the problem of varying performances of seed selection approaches depending on the network topological characteristics. Social networks enable people to disseminate a variety of information to masses easily, but under the budget and time constraints. Therefore, many seed selection approaches have been proposed to find a small subset of nodes so that the influence spread could be maximized. More recently, it has been revealed that the performances of seed selection approaches in a given network are shaped and affected by the network characteristics. Many studies have explored the effects of network characteristics or network types on the information diffusion processes and on the performances of seed selection approaches in a non-competitive environment. However, none of these studies considered the fact that there are more than one influences spreading simultaneously over a network in the real world and the performances of the seed selection approaches are also affected by the competition between them. Motivated by this gap in the literature, this thesis aims to investigate the direct impacts of the network characteristics on the competitive performances of seed selection approaches.

To be more precise, a simulation-based study for the effects of network characteristics on the competitive influence maximization is conducted under the circumstance that there are two opposing information spreading over the network. For that purpose, a two-stage analysis is carried out. Firstly, the standalone performances of the approaches (performance in non-competitive environment) are simulated, and the effects of the network characteristics on their performances are observed. In the second stage, the approaches are compared in terms of their performances against each other, and the effects of network characteristics on their competitive performances are investigated. Whether there is a difference between the effects of network characteristics on the standalone and the competitive performances of the approaches or not is also interpreted.

In this regard, the direct effects of the following network characteristics are investigated: Average clustering coefficient (ACC), average degree (AD), normalized average path length (NAPL), normalized degree variance (NDV) and density (D). Furthermore, the most popular four centrality-based heuristics namely *betweenness (BC),*

degree (DC), *closeness (CC)* and *eigenvector centrality (EC)* are involved in the scope of the thesis.

The experiments are conducted on 13 real-world network datasets, which differ in terms of their network characteristics and categories. There are three Facebook, two Twitter, three collaborations, one citation, one email communication, one Wikipedia and one online social friendship network in the datasets. To conduct the experiments, the Linear Threshold Model is extended to a competitive version according to the nature of the problem of the thesis. The same uniformly distributed thresholds and weights in the interval [0,1] are assigned to nodes and edges for the two types of the opponents, i.e. thresholds and weights do not depend on the opponent type. Thus, the identical conditions are created to compare the heuristics performances against each other.

To simulate the information diffusion, a simulation model is created through Python programming language. Before the experiments, the model is tested in terms of consistency between the conceptual design and the execution. After ensuring the accuracy of the model, the simulation runs are firstly conducted with replications for the standalone experiments. Since whole the heuristics performed better performance than *R*, all of them are included in the competitive experiments. For each heuristic combination (*BC* vs *DC*, *BC* vs *CC*, *BC* vs *EC*, *DC* vs *CC*, *DC* vs *EC*, *EC* vs *CC*), the results of the replications are averaged for each heuristic in every heuristic combination. Later, the results and the dataset statistics are combined through the regression trees to interpret the effects of network characteristics on the heuristics' performances easily.

The results indicated that there is general ranking of the heuristics in terms of their standalone performances, which is as follows: *BC>DC>CC>EC>R.* Besides, the heuristics *BC* and *DC* performed better than *R* on all of the datasets. Contrary to the existing literature conducted on a few network dataset and with a different diffusion model, the heuristic *EC* also performed better than *R* in this study. This contradiction highlights that the performances of the heuristics are affected by the diffusion model and network characteristics.

Referring the regression trees created for the standalone performances of the heuristics, the study reveals that the ACC is the most significant network characteristic for the standalone performances of all the heuristics, and the heuristics perform better on the slightly clustered networks. The other influential characteristics are found as the AD and the NAPL. The best standalone performances of *BC* and *DC* are observed on the slightly clustered networks with low AD and NAPL. Besides, *CC* and *DC* reach their highest influence spreads on the slightly clustered networks with high AD and low NAPL.

According to the average influence spreads in the competitive experiments, the performance ranking of the heuristics changes from non-competitive to competitive environment as follows: *DC*>*BC*>*CC*>*EC*. In contrast with the general ranking, the results demonstrate that *BC* outperforms *DC* on most of the networks except the ones with very low AD. However, interestingly, using the heuristic *DC* against the heuristics *CC* and *EC* is more strategical than using *BC*. Apart from the heuristics' performance superiorities over each other, the presence of an opponent in the network results in a dramatic influence loss from the performances of the heuristics. In addition to this, the heuristic *DC* is found as the most robust one to the existence of an opponent in the network, and it is followed by *BC, CC* and *EC* respectively.

Comparing the effects of network characteristics between standalone and competitive results, it is observed that the importance ranking of the significant characteristics changes. For instance, the most influential characteristic on the performance of *BC* and *DC* heuristics becomes the AD characteristic instead of ACC. Besides, it is observed that *DC* outperforms *BC* on the slightly clustered networks with a very low AD, although a decrease in AD is advantageous for both of the heuristics. The effect of the NAPL disappears for the heuristics *BC* and *CC* while it is still important for *DC* and *EC* because the path-based nature of betweenness and closeness metrics results in robustness to changes in the NAPL characteristic of a network. Moreover, the effect of NDV arises on the heuristics *CC* and *EC.* Although *CC* and *EC* are still behind *BC* and *DC*, they perform better on the variant networks in terms of node degrees. Apart from the changes in the effects of characteristics, the D has not showed a significant effect on any heuristic's performance in both standalone and competitive experiments. The reason behind this is probably the fact

that the D creates positive and negative effects on the information diffusion at the same time due to different effects of clusters' densities including initial seeds.

By analyzing also the effect of the opponent type on the performances of the heuristics, this thesis has shown that only the heuristic *CC* is affected more by the opponent type than the network characteristics. In other words, the performance of *CC* fluctuates quietly as type of the opponent changes. Apart from *CC,* influence loss of other heuristics in competitive environment does not depend on the type of opponent.

To sum up, combining the results, this study has shown that all the four centrality-based heuristics perform well on the slightly clustered networks with lower NAPL when they are alone in the network. However, an increase in AD on such networks is advantageous for the heuristic *CC* and *DC* while it is disadvantageous for *BC* and *DC*. In line with the objectives of the study, the research clearly illustrates that the performances of the heuristics and the importance of the network characteristics' effects are sensitive to the presence of an opponent in the network and to the type of it. Considering the four heuristics, it is concluded that all of them lose from their performances dramatically in competing against each other. Moreover, while *BC* outperforms *DC,* the best competitive results against *CC* and *EC* are obtained from the heuristic *DC.* Unlike the standalone performances, the most influential characteristic is AD for the competition between *BC* and *DC,* and the most supporting and influential factor is an increase in the NDV for the competition of *CC* or *EC* against their three opponents.

Overall, this thesis sheds light on how the network characteristics affect the performances of the centrality-based seed selection heuristics against each other with an integrated understanding in social network analysis and competitive influence maximization. In the scope defined, 13 real-world datasets and an extension of Linear Threshold Model are used, and it is assumed that there are two opposing information spreading over a network. Therefore, there are several future directions that would be necessary to evaluate how these findings can be extended and generalized in an effective manner. Firstly, it would be beneficial to enlarge the number and the size of the network datasets in future studies. Secondly, rather than using an extension of Linear Threshold Model, a similar study can be conducted with different competitive information diffusion models. Finally, conducting a

similar study with more than two opposing information spreading over a network may be beneficial.

REFERENCES

Albert, R. and A. L. Barabási, 2002, "Statistical mechanics of complex networks", *Reviews of Modern Physics*, Vol. 74, No. 1, pp. 47-97.

Barthelemy, M., 2004, "Betweenness centrality in large complex networks", *The European Physical Journal B*, Vol. 38, No. 2, pp. 163-168.

Barthelemy, M., A. Barrat, R. Pastor-Satorras, and A. Vespignani, 2004, "Velocity and hierarchical spread of epidemic outbreaks in scale-free networks" *Physical Review Letters*, Vol. 92, No. 17, pp. 1-4.

Belsley, D. A., E. Kuh and R. E. Welsch, 2005, *Regression diagnostics: Identifying influential data and sources of collinearity*, Vol. 571, John Wiley & Sons, New York.

Berger, E., 1999, *Dynamic Monopolies of Constant Size*, M.S. Thesis, the Technion Israel.

Bharathi, S., D. Kempe and M. Salek, 2007, "Competitive influence maximization in social networks", *International workshop on web and internet economics*, pp. 306-311, Springer, Berlin, Heidelberg.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez and D. U. Hwang, 2006, "Complex networks: Structure and Dynamics", *Physics reports*, Vol. 424, No.4, pp. 175-308.

Bonacich, P., 1987, "Power and centrality: A family of measures", *American journal of sociology*, Vol. 92, No. 5, pp. 1170-1182.

Borgatti, S. P., and M. G. Everett, 2006, "A graph-theoretic perspective on centrality", *Social networks*, Vol. 28, No. 4, pp. 466-484.

Borgs, C., M. Brautbar, J. Chayes and B. Lucier, 2014, “Maximizing social influence in nearly optimal time”, *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pp. 946-957.

Borodin, A., Y. Filmus and J. Oren, 2010, “Threshold models for competitive influence in social networks”, In *International workshop on internet and network economics*, pp. 539-550, Springer, Berlin, Heidelberg.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone, 1993, *Classification and Regression Trees*, Chapman & Hall, London.

Budak, C., D. Agrawal and A. El Abbadi, 2011, “Limiting the spread of misinformation in social networks.”, *Proceedings of the 20th international conference on World wide web*, pp. 665-674.

Cancho, R. F., R. V. Solé, and R. Köhler, 2004, “Patterns in syntactic dependency networks”, *Physical Review E*, Vol. 69, No. 5, pp. 1-8.

Carnes, T., C. Nagarajan, S. M. Wild and A. Van Zuylen, 2007, “Maximizing influence in a competitive social network: a follower's perspective”, *Proceedings of the ninth international conference on Electronic commerce*, pp. 351-360.

Chechik, S., D. H. Larkin, L. Roditty, G. Schoenebeck, R. E. Tarjan and V. V. Williams, 2014, “Better approximation algorithms for the graph diameter.”, *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1041-1052.

Chen, D., L. Lü, M. S. Shang, Y. C. Zhang and T. Zhou, 2012, “Identifying influential nodes in complex networks”, *Physica a: Statistical mechanics and its applications*, Vol. 391, No. 4, pp. 1777-1787.

Chen, F., Z. Chen, X. Wang and Z. Yuan, 2008, “The average path length of scale free networks”, *Communications in Nonlinear Science and numerical simulation*, Vol.13 No.7, pp. 1405-1410.

Chen, W., Y. Wang and S. Yang, 2009, "Efficient influence maximization in social networks", *Proceedings of the 15th Association for Computing Machinery international conference on knowledge discovery and data mining*, pp.199-208.

Cheng, S., H. Shen, J. Huang, G. Zhang and X. Cheng, 2013, "Static greedy: solving the scalability-accuracy dilemma in influence maximization", *Proceedings of the 22nd Association for Computing Machinery international conference on Information & Knowledge Management*, pp. 509-518.

Cohen, E., D. Delling, T. Pajor and R. F. Werneck, 2014, "Computing classic closeness centrality, at scale", *Proceedings of the second Association for Computing Machinery conference on Online social networks*, pp. 37-50.

Datareportal, 2020, "Digital 2020: Global Digital Overview", No. 10, https://datareportal.com/library

Dorogovtsev, S. N. and J. F. Mendes, 2002, "Evolution of networks", *Advances in physics*, Vol. 51, No. 4, pp. 1079-1187.

Erkol, Ş. and G. Yücel, 2017, "Influence maximization based on partial network structure information: A comparative analysis on seed selection heuristics", *International Journal of Modern Physics C*, Vol. 28, No. 10, p. 1750122.

Estevez, P. A., P. Vera and K. Saito, 2007, "Selecting the most influential nodes in social networks", *International Joint Conference on Neural Networks*, pp. 2397-2402.

Facebook, 2020, "Quarterly Report 2020 - March", No. https://investor.fb.com/investor-news/.

Ghalmane, Z., M. El Hassouni and H. Cherifi, 2018, "Betweenness centrality for networks with non-overlapping community structure", *Institute of Electrical and Electronics Engineer workshop on complexity in engineering*, pp.1-5.

Goldenberg, J., B. Libai and E. Muller, 2001, "Talk of the network: A complex systems look at the underlying process of word-of-mouth", *Marketing letters*, Vol. 12, No. 3, pp. 211-223.

Goyal, A., W. Lu and L. V. Lakshmanan, 2011, "Simpath: An efficient algorithm for influence maximization under the linear threshold model", *11th international conference on data mining*, pp. 211-220, IEEE.

Granovetter, M., 1978, "Threshold models of collective behavior", *American journal of sociology*, Vol. 83, No. 6, pp. 1420-1443.

Harrigan, N., P. Achananuparp and E. P. Lim, 2012, "Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news", *Social Networks*, Vol. 34, No. 4, pp. 470-480.

Hussain, O. A., Z. Anwar, S. Saleem and F. Zaidi, 2013, "Empirical analysis of seed selection criterion in influence mining for different classes of networks", *International Conference on Cloud and Green Computing*, pp. 348-353, IEEE.

Jahanpour, E. and X. Chen, 2013, "Analysis of complex network performance and heuristic node removal strategies", *Communications in Nonlinear Science and Numerical Simulation*, Vol 18, No. 12, pp. 3458-3468.

Jiang, C., Y. Chen and K. R. Liu, K., 2014, "Evolutionary dynamics of information diffusion over social networks", *Transactions on Signal Processing*, Vol. 62, No. 17, pp. 4573-4586.

Jiang, Q., G. Song, C. Gao, Y. Wang, W. Si, and K. Xie, 2011, "Simulated annealing based influence maximization in social networks", In *25th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*., pp. 127-132.

Katona, Z., P. P. Zubcsek and M. Sarvary, 2011, "Network effects and personal influences: The diffusion of an online social network", *Journal of marketing research,* Vol. 48, No. 3, pp. 425-443.

Kempe, D., J. Kleinberg and É. Tardos, 2003, "Maximizing the spread of influence through a social network", *Proceedings of the 9th Association for Computing Machinery international conference on Knowledge discovery and data mining*, pp. 137-146.

Kemper, A., 2009, *Valuation of network effects in software markets: A complex networks approach*, Springer Science and Business Media.

Kim, J., O. Kwon and D. H. Lee, 2017, *Social influence of hubs in information cascade processes*, Management Decision.

Kimura, M., K. Saito, R. Nakano and H. Motoda, 2010, "Extracting influential nodes on a social network for information diffusion", *Data Mining and Knowledge Discovery,* Vol. 20, No. 1, pp. 70-97.

Kitsak, M., L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, 2010, "Identification of influential spreaders in complex networks", *Nature Physics,* Vol. 6, No. 11, pp. 888-893.

Leskovec, J. and J. J. Mcauley, 2012, "Learning to discover social circles in ego networks", *In Advances in neural information processing systems*, pp. 539-547.

Leskovec, J., D. Huttenlocher and J. Kleinberg, 2010, "Predicting positive and negative links in online social networks", *Proceedings of the 19th international conference on World wide web,* pp. 641-650.

Leskovec, J., J. Kleinberg and C. Faloutsos, 2007, "Graph evolution: Densification and shrinking diameters", *Association for Computing Machinery transactions on Knowledge Discovery from Data*, Vol. 1, No. 1, pp.1-41.

Leskovec, J., A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen and N. Glance, 2007, "Cost-effective outbreak detection in networks", *Proceedings of the 13th Association for Computing Machinery international conference on Knowledge discovery and data mining*, pp. 420-429.

Liu, W., K. Yue, H. Wu, J. Li, D. Liu, and D. Tang, 2016, "Containment of competitive influence spread in social networks", *Knowledge-Based Systems*, Vol. 109, pp. 266-275.

Liu, Q. and T. Hong, 2018, "Sequential seeding for spreading in complex networks: Influence of the network topology", *Physica A: Statistical Mechanics and its Applications*, Vol. 508, pp. 10-17.

Lü, L., D. Chen, X. L. Ren, Q.M. Zhang, Y. C. Zhang, and T. Zhou, 2016, "Vital nodes identification in complex networks", *Physics Reports*, Vol. 650, pp. 1-63.

Newman, M. E., 2003, "The structure and function of complex networks", *Society for Industrial and Applied Mathematics review*, Vol. 45, No. 2, pp. 167-256.

Nguyen, H. T., M. T. Thai and T. N. Dinh, 2016, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks", *Proceedings of the 2016 International Conference on Management of Data*, pp. 695-710.

Peres, R., 2014, "The impact of network characteristics on the diffusion of innovations", *Physica A: Statistical Mechanics and Its Applications*, Vol. 402, pp. 330-343.

Ryan A. R. and K. A. Nesreen, 2015," The Network Data Repository with Interactive Graph Analytics and Visualization.", In *29th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pp. 4292-4293.

Stonedahl, F., W. Rand and U. Wilensky, 2010, "Evolving viral marketing strategies", *In Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pp. 1195-1202.

Sumith, N., B. Annappa and S. Bhattacharya, 2018, “Influence maximization in large social networks: Heuristics, models and parameters”, *Future Generation Computer Systems,* Vol. 89, pp. 777-790.

Tang, Y., X. Xiao and Y. Shi, 2014, “Influence maximization: Near-optimal time complexity meets practical efficiency”, *Proceedings of the 2014 Association for Computing Machinery International Conference on Management of Data,* pp. 75-86.

Tang, J., X. Tang, X. Xiao, X. and J. Yuan, 2018, “Online processing algorithms for influence maximization”, *Proceedings of the 2018 International Conference on Management of Data*, pp. 991-1005.

Twitter, 2020, “Quarterly Report 2020 - March”, https://investor.twitterinc.com/financial-information/quarterly-results.

Valente, T. W., K. Coronges, C. Lakon and E. Costenbader, 2008, “How correlated are network centrality measures?”, *Connections,* Vol. 28, No. 1, pp. 16-26.

Wang, X. F. and G. Chen, 2003, “Complex networks: small-world, scale-free and beyond”, *IEEE circuits and systems magazine,* Vol. 3, No. 1, pp. 6-20.

Wasserman, S. and K. Faust, 1994, *Social network analysis: Methods and applications*, Vol. 8, Cambridge university press.

Watts, D. J. and S. H. Strogatz, 1998, “Collective dynamics of ‘small-world’ networks", *Nature*, Vol. 393, No. 6684, pp. 440-442.

Wu, H., W. Liu, K. Yue, W. Huang and K. Yang, 2015, “Maximizing the spread of competitive influence in a social network oriented to viral marketing”, In *International conference on web-age information management*, pp. 516-519, Springer, Cham.

Yen, C. C., M. Y. Yeh and M. S. Chen, 2013, December, “An efficient approach to updating closeness centrality and average path length in dynamic networks”, In *13th International Conference on Data Mining*, pp. 867-876, IEEE.

Zhang, X., J. Zhu, Q. Wang and H. Zhao, 2013, “Identifying influential nodes in complex networks with community structure”, *Knowledge-Based Systems*, Vol. 42, pp. 74-84.

Zhao, J., Q. Liu, L. Wang and X. Wang, 2017,” Competitive seeds-selection in complex networks”, *Physica A: Statistical Mechanics and its Applications*, Vol. 467, pp. 240-248.

Zhuang, Y. B., J. J. Chen and Z. H. Li, 2017, “Modeling the cooperative and competitive contagions in online social networks”, *Physica A: Statistical Mechanics and its Applications*, Vol. 484, pp. 141-151.

Zimmermann, M. G., V. M. Eguíluz and M. San Miguel, 2004, “Coevolution of dynamical states and interactions in dynamic networks”, *Physical Review E*, Vol. 69, No. 6, p. 065102.

APPENDIX A: REGRESSION TREE EXAMPLE

The study of Belsley *et al.* (2005) which investigates whether there is an effect of air pollution concentration on housing values in Boston is explained in detail to give information about how to read regression trees. The data includes mean value of houses in dollars and thirteen explanatory variables listed in the table below:

Table A.0.1. Response and Explanatory Variables (Belsley et al., 2005)

y	Mean value of houses in thousands of dollars (MV)
x_1	Crime rate (CRIM)
x_2	Percent land zoned for lots (ZN)
x_3	Percent nonretail business (INDUS)
x_4	1 if on Charles River, 0 otherwise (CHAS)
x_5	Nitrogen oxide concentration, pphm (NOX)
x_6	Average number of rooms (RM)
x_7	Percent built before 1940 (AGE)
x_8	Weighted distance to employment centers (DIS)
x_9	Accessibility to radial highways (RAD)
x_{10}	Tax rate (TAX)
x_{11}	Pupil/teacher ratio (P/T)
x_{12}	Percent black (B)
x_{13}	Percent lower-status population (LSTAT)

Each frame in Figure A.1. is named as node. The number on the top in the frames represents the mean value of the response variable which is the housing value in this example and the below percentage value represents what percentage of the data falls into that node. The first frame, the root node, is the mean value of the houses in Boston with 22.5 thousands of dollars.

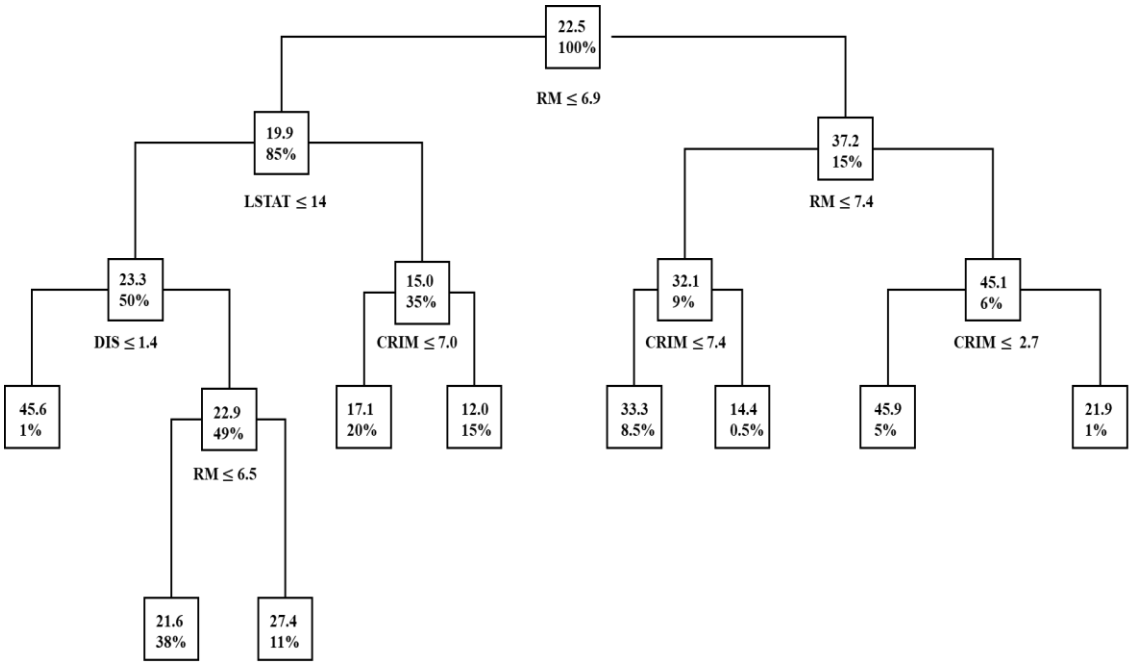

Figure A.0.1. Sample Regression Tree Chart (Belsley et al., 2005).

The rules that split the nodes are written below each frame. The left branch under each frame contains the data of which the rule of inequality is valid for and the right branch contains the data of the opposite. For example, the root node is partitioned into two nodes according to the average number of rooms in the houses. It represents that 85% of the houses in Boston have less than 6.9 rooms with the 19.9 thousand dollars value on average and the 15% of the houses have more than 6.9 rooms with the average value of 37.2 thousand dollars. Then, the left branch is split according to the percentage of lower-status population around the house. The 50% of the data goes left, and then it is split again into two. The first terminal node is generated where DIS <= 1.4. According to that terminal node, only 1% of the houses in Boston have less than 6.9 rooms, maximum 14% of the population around their neighborhood comes from lower-status and the weighted distance to business centers is less than 1.4 km. The average value of these houses is pretty high with 45.6 thousand dollars.

The right branch under the root node splits on number of rooms (RM) again. According to this split, 9% of the houses in Boston have less than 7.4 rooms with 32.1 thousands dollars average value while 6% have more than 7.4 rooms with 45.1 thousand dollars average. Then, both of the nodes are split based on CRIM explanatory variable. It can be clearly seen from the terminal nodes; crime rate significantly changes the values of

houses. As a result, the nodes are followed down with this approach and the other branches can be interpreted in the same way.