Forward and Backward Linear Threshold Ranks

Maria J. Blesa
Computer Science dept.
Universitat Politècnica de Catalunya
Barcelona, Spain
maria.j.blesa@upc.edu

Pau García-Rodríguez
School of Informatics
Universitat Politècnica de Catalunya
Barcelona, Spain
pgarcia@cs.upc.edu

Maria Serna
Computer Science dept and IMTech
Universitat Politècnica de Catalunya
Barcelona, Spain
maria.serna@upc.edu

***Index Terms*—Centrality measures, Spreading/Diffusion models, Linear Threshold Rank**

***Abstract*—*Who are the most important or influencing actors of a social network?* There are many ways to answer this question, but node *centrality* is probably the theoretical concept that better captures it and formalizes it. Traditional centrality measures, like the degree, closeness or betweenness of the actors, are purely based on topological properties of the network and that is not enough. Nowadays social networks are much more than simple topological structures, specially because they are very much involved in the spreading of information and have become a media themselves. Lately, models capturing that diffusion effect are enriching our theoretical knowledge about them. Examples of it are the Independent Cascade model and the Linear Threshold model and, based on them, a new generation of centrality measures are being proposed, e.g. the Independent Cascade Rank and the Linear Threshold Rank (LTR) [28].**

The focus of this research is on analyzing the effect of the initial activation set in the Linear Threshold Rank (LTR). The rank assigns to each actor the size of the set influenced by taking as initial activation set the actor and its immediate neighborhood. In the LTR both predecessors and successors are considered as immediate neighbors. We propose to analyze here the effect on the rank when the initial activation set contains only successors (Forward LTR, FwLTR), or only predecessors (Backward LTR, BwLTR). We perform an experimental analysis in a data set formed by a selection of networks with a variety of parameters. We compare our proposed measures among themselves and with other classical measures. Our results show that LTR and BwLTR behave quite similarly, while FwLTR is different. Furthermore the LTR variations define measures that are different from the other centrality measures considered in the study. To make the comparison, we use the same parameters and statistics used in [28] together with other taken from Social Sciences: the Gini coefficient and a categorization analysis. This second kind of new comparative analysis is of independent interest.

I. Introduction

Facebook, Twitter, LinkedIn, Instagram, Snapchat... In recent years, social media are more and more integrated in our daily lives. Via digital ways we are connected to the rest of the world. We create accounts and connect to our friends or spread our experiences, photos and opinions among followers. Social networks can be represented as a graph with actors as nodes and edges representing the interpersonal ties. The massive increment of social media has allowed the emergence of varied and complex social networks. One of the main questions arising is '*Who are the most important or central persons in the network?*'.

In graph theory and network analysis, indicators of centrality identify the most important vertices within a graph. Applications include identifying the most influential person(s) in a social network, key infrastructure nodes in the Internet or urban networks, and super-spreaders of disease. Centrality concepts were first developed in social network analysis, and many of the terms used to measure centrality reflect their sociological origin [23]. They should not be confused with node influence metrics, which seek to quantify the influence of every node in the network.

Centrality is one of the most studied concepts in social network analysis. It defines the importance of a node from a particular perspective. Furthermore, it might provide relevant analytical information about the graph and its nodes. This started in 1948 when Bavelas published his paper *A Mathematical Model for Group Structures* [3]. Centrality measures determine how structurally relevant an actor is within a social network. Traditional measures are the degree, closeness and betweenness which are related to the topology of the graph. Other well-known centrality measures are the Katz Rank [14] and the PageRank [24].

Two centrality measures have been introduced in an attempt to measure centrality with respect to influence spreading. The Independent Cascade (IC) model is a stochastic model which was initially proposed in the context of marketing [11]. It is based on the assumption that whenever a node is activated, it will do (stochastically) attempt to activate a neighbor. The whole process ends when there are no active nodes with a new chance to spread its influence. Based on the IC model, the Independent Cascade Rank (ICR) centrality measure was defined [16]. The ICR associates to each actor the normalized size of the set that the actor can activate under the IC model.

The Linear Threshold (LT) model is a deterministic model for influence spread based on some ideas of collective behavior [15]. In the LT model the strength of the tie between every pair of actors quantifies the capacity of one to influence the other and, additionally, each actor opposes a resistance to be influenced. A node gets influenced when their active predecessors can exert enough influence to overpass its resistance. Based on the LT model, the Linear Threshold Rank (LTR), was introduced in [28]. The LTR for an actor i can roughly be described as the part of the network which is influenced

Supported by MINECO under grant TIN2017-86727-C2-1-R (GRAMM) and AGAUR under grant 2017-SGR-786 (ALBCOM).

The final publication is available at ACM via http://dx.doi.org/10.1145/3487351.3488355

when initially only actor i and its neighbors (predecessors and successors), *the initial activation set*, are spreading their common opinion.

We analyze here the impact on the LTR of the selection of the initial activation set. In the original paper both predecessors and successors were included in it. However, in real social networks the predecessors and the successors play a different role in spread of influence processes. It seems more natural to think that only the immediate successor of a node could be initially influenced when measuring the influence power of the node. Furthermore, we have the perception that a good set of predecessors would make the role of the successors irrelevant. Under this point of view, we consider two new centrality measures, the Forward Linear Threshold Rank (FwLTR) and the Backward Linear Threshold Rank (BwLTR). The initial set is formed only by the successors of a node i in the former, and only by its predecessors in the latter. Our aim is to asses if the three measures are related among themselves and also, whether the new ones are different or not from the other centrality measures considered in the paper.

We perform experiments on our new defined centrality measures. For those experiments, we selected data sets which are easy to model as an influence graph, with a variety of sizes and properties. We complement the four social networks used in the experiments in [28] with another four social networks. To compare the results of the different centrality measures, we use the Kendall [17] and Spearman [30] coefficients, together with two new methods. We use the Gini coefficient [4], [10], which is mostly known from the field of sociology, as a measure of the inequality of populations with respect to different criteria (e.g., wealth spread). Besides, we apply an inequality study based on the social tables and political arithmetic of Denis Diderot of the 18th century. In this method the actors of the network are divided in three categories based on the outcome of a centrality measure. We observe the migration of nodes across categories when switching to another underlying centrality measure as an additional parameter that might shed light on the intrinsic differences among measures.

Our results show that the FwLTR is in general different from the BwLTR. They only behave almost identical in the almost symmetric network (the `Caida` network) considered in our data set. On the other hand the BwLTR appears to be closely related to the LTR.

The paper is organized as follows. In Section II, we introduce the background on social networks and influence spread together with the definitions of the already known centrality measures. The definition of the LTR and the new variants FwLTR and BwLTR is given in Section II. In Section IV, we describe the social networks used in this study. Details and results of the experiments, together with the parameters of the comparative analysis, are given in Section V. The paper ends with conclusions and a discussion about further research in Section VI.

II. Centrality measures

We introduce some known centrality measures that we will use later to compare our new measures. The reader is referred to [12] for an overview on these and other common centrality measures. In all what follows, we consider a social network as a directed graph, whose nodes are the actors of the network and whose arcs represent the interpersonal ties among these actors. Thus, we represent a social network as a digraph $G = (V, E)$, where V is the set of actors or individuals and E is the set of edges of G. As we will see, sometimes we require a weighted digraph (G, w, θ), where G is a graph and $w : E \to \mathbb{N}$ and $\theta : V \to \mathbb{N}$ are *weight functions* which assigns a non negative weights to every edge and vertex. For a vertex $u \in V$, the output neighborhood is defined $N^+(u) = \{v \mid (u, v) \in E\}$, the input neighborhood as $N^-(u) = \{(v \mid (v, u) \in E\}$, and the total neighborhood as the union of both, i.e., $N(u) = N^+(u) \cup N^-(u)$. Consequently, the outdegree $\delta^+(u) = |N^+(u)|$ refers to the cardinality of the output neighborhood, the indegree $\delta^-(u) = |N^-(u)|$ quantifies the cardinality of the input neighborhood and therefore, $\delta(u) = |N(u)|$ denotes the degree. Given two vertices $u, v \in V$, $d(u, v)$ denotes the total weight of the lightest path between nodes u and v when considering only its edge weights. In social networks, $d(u, v)$ is also often referred to as the *distance* between the nodes. We use $\ell(u, v)$ to refer to the shortest path between u and v in the classical sense, i.e., in terms of the number of edges. Note that $d(u, v) = \ell(u, v)$ when $\forall (i, j) \in E : w_{ij} = 1$.

Centrality measures can be classified in many ways depending on how the concept of centrality is interpreted and computed. In this work, we distinguish two categories depending on whether what strongly influences the calculation of the measure is the topology of the graph, or rather the power to influence other actors.

A. Topology-based measures

The oldest and simplest centrality measure is the *degree* of a node $u \in V$, which is defined as the number of links incident upon it. The degree can be interpreted in terms of the immediate risk of a node to be captured by whatever is flowing through the network (e.g. virus, information).

The *closeness* of a node $u \in V$ is the reciprocal of the sum of the shortest path distances to all other nodes. The closeness takes a measurement based on the distance to all other nodes. Thus, the more central a node is, the closer it is to all other nodes.

The *betweenness* of a node $u \in V$ is the sum of the fraction of all-pairs shortest paths that pass through u. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Thus, a node is more important if it enables short links to other actors in the network. Indirectly, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness. From a reciprocal point of view, it is somehow measuring how many shortest paths would be broken (and thus, how much

disconnected the network would be) if the would is removed. It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network. [8]

These three measures are formally defined respectively as follows:

$$\mathsf{Deg}(u) = \delta(u), \text{ and } \quad \mathsf{Clsn}(u) = \frac{n-1}{\sum_{v \in V \setminus \{u\}} d(u,v)},$$

$$\text{and} \quad \mathsf{Btwn}(u) = \sum_{s,t \in V} \frac{\sigma(s,t|u)}{\sigma(s,t)},$$

where $\sigma(s,t)$ is the number of shortest (s,t)-paths and $\sigma(s,t|u)$ is the number of those paths passing through some node u.

The importance of a node can also derive from properties other than those merely coming from the graph structure. Additional relevance criteria can be introduce in order to provide the property of centrality with different meanings and applications. Within this category, we consider two well-known centrality measures: the Katz centrality and the PageRank.

The *Katz centrality* [14] is a generalization of degree centrality and it can also be viewed as a variant of eigenvector centrality. While degree centrality measures the number of direct neighbors, the Katz centrality measures the number of all nodes that can be connected through a path, while the contributions of distant nodes are penalized. It is based on the idea that an actor is important if it is linked to other important actors or if it is highly linked. The Katz centrality of an actor $u \in V$ is given by

$$\mathsf{Ktz}(u) = \alpha \sum_{v \in V} (a_{v,u})\ \mathsf{Ktz}(v) + \beta,$$

where A is the adjacency matrix representation of the graph (i.e., elements $(a_{i,j}) = 1$ if $(i,j) \in E$, and $(a_{i,j}) = 0$ otherwise), β is a constant which is independent of the network structure and $\alpha \in [0, \lambda_{max}^{-1}]$ is the damping factor, being λ_{max} the largest eigenvalue of A.

Another well-known centrality measure is called *PageRank* [24]. PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. Given a damping factor $\alpha \in (0,1]$, the PageRank of an actor $u \in V$, is given by

$$\mathsf{PgR}(u) = (1-\alpha) + \alpha \sum_{v \in V} \frac{(a_{v,u})\ \mathsf{PgR}(v)}{\delta^+(v)}.$$

B. Influence-based measures

A criteria to consider the importance of a node can also derive from its power to influence the other nodes in the network. The spread of influence is used to model the ways in which actors influence each other through their interactions. Given an influence spread (or diffusion) model describing how an element (e.g. virus, information) spreads through the network and how the nodes and edges of the network participate on that spreading process, the centrality of a node can be determined in order to quantify how important its participation is in the whole spreading procedure. In the initial state, a few nodes $X \subseteq V$ are activated (namely, the *seeds* or *core*), which represent the source of the diffusion. Other nodes will be activated as a response of the influence of their neighbors. Once activated, a node will never be deactivated again. As this procedure evolves in time, it reaches the moment when no new nodes are activated anymore. Conditions other than the natural stagnation of the process (e.g., a fixed certain amount of steps or time, reaching a certain node) can also be used to stop the process. In any case, once the spreading process is finished we can talk about the *influence spread* of the initial core X, i.e., the final amount of nodes influenced by X over time. A key point of a diffusion model will be to stablish the necessary conditions for that influence to activate the nodes. Perhaps the two most prevalent diffusion models in computer science are the *Independent Cascade* model [11] and the *Linear Threshold* model [15] (see also [29]).

The IC model is a stochastic model which was initially proposed in the context of marketing. It is based on the assumption that whenever a node is activated, it will do (stochastically) attempt to activate each actor he targets. Given an activated node $i \in V$, any neighbor $j \in N^+(i)$ will be activated with a probability p_{ij}. When a new actor is activated, the process is repeated for this actor. The whole process ends when there are no active nodes with a new chance to spread its influence.

Given an initial core $X \subseteq V$ and a probability $p \in [0,1]$ (where $\forall (i,j) \in E : p_{ij} = p$), the influence spread of X is denoted by $F'(X,p)$. Based on the Independent Cascade model, in [16] the Independent Cascade Rank of a node $u \in V$ is then defined as

$$\mathsf{ICR}(u,p) = \frac{|F'(u,p)|}{\max_{v \in V}\{|F'(v,p)|\}}.$$

In contrast, the Linear Threshold (LT) model is a deterministic model for influence spread, where the strength of the tie between every pair of actors quantifies the capacity of one to influence the other and, additionally, each actor opposes a resistance to be influenced.

To represent the strength of the interpersonal tie between each par of connected actors, the edges of the graph representing the social network have associated weights. For a given edge $(i,j) \in E$ with weight $w_{ij} > 0$, the interpretation is that actor i can influence j with power w_{ij}. Thus, the greater the weight w_{ij}, the stronger the power of $i \in V$ to influence $j \in V$.[1] Every node $u \in V$ is also considered to have a weight

[1] A variation of the LT model was recently introduced [33], which considered both positive and negative relationships within the network. Those networks are commonly named friend/foe networks, where positive weights correspond to friend relationships and negative weights correspond to a distrust ones. In this paper we only consider networks with positive edge weights.

$\theta_u \geq 0$, which represents its resistance to be influenced. The greater the weight of a node θ_u, the greater the resistance of the node to be influenced. In practical terms, this will mean that it will require the amount of influence coming from its input neighbors to be proportional to its resistance weight in order to be activated. For this reason, θ_u is also sometimes referred to as the *threshold* of node $u \in V$. We name a network (G, w, θ) with such characteristics an *influence graph*.

Based on the LT model, the *Linear Threshold Rank* was recently proposed in [28] as a new centrality measure. This new measure showed to be useful for ranking actors and networks in a distinguishable way. In this paper, we propose new centrality measures related to the Linear Threshold Rank, with special focus on the direction in which the diffusion takes place. For readability and completeness reasons, we postpone the mathematical definition of the Linear Threshold Rank to the next section.

III. The Forward and the Backward Linear Threshold Ranks

Given an influence graph (G, w, θ) and an non-empty core $X \subseteq V$, we consider the following iterative activation process: Let $F_t(X) \subseteq V$ be the set of nodes activated at some iteration t. Initially, at step $t = 0$, only the nodes in X are activated, that is $F_0(X) = X$. At iteration $t + 1$ a node $i \in V$ will be activated if, and only if, $\sum_{j \in F_t(X)} w_{ji} \geq \theta_i$. In other words, a node i is activated when the weights' sum of the activated nodes connected to i is greater or equal to its resistance to be influenced. The value t denotes the current *spread level* of X. The process stops when no additional activation occurs (usual stagnation condition). Therefore, the *spread of influence* of X is defined as

$$F(X) = F_k(X), \text{ where}$$

$$k = \min\{t \in \mathbb{N} | F_t(X) = F_{t+1}(X)\} \leq n.$$

The linear threshold rank of a node $u \in V$ is computed as

$$\mathsf{LTR}(u) = \frac{|F(\{u\} \cup N(u)|}{n}.$$

We introduce here two directed counterparts that take into account the direction of the spread: the *Forward Linear Threshold Rank (FwLTR)* and the *Backward Linear Threshold Rank (BwLTR)*, which are defined as follows. For $u \in V$,

$$\mathsf{FwLTR}(u) = \frac{|F(\{u\} \cup N^+(u))|}{n},$$

$$\mathsf{BwLTR}(u) = \frac{|F(\{u\} \cup N^-(u))|}{n}.$$

Observe that the only difference among the three ranks is in the initial core. As $F(X)$ can be computed in polynomial time for any $X \subseteq V$, $\mathsf{LTR}(u)$ and $\mathsf{FwLTR}(u)$ and $\mathsf{BwLTR}(u)$ can also be computed in polynomial time [28].

IV. The Data Sets

The experiments in [28], where the LTR centrality measure was proposed, are run on four data sets, namely: the **ArXiv** network [21], the **dining-table** partners network [7], the **Dolphin** social network, the **Higgs** network [5]. In addition, we decided to expand the benchmark for out experiments with four additional data sets: the **Caida** network [20], the **Epinions** network [26], the **Human Brain** network [31], [32], and the **Wikipedia** network [18], [19]. In this paper we focus only on directed networks, as for undirected graphs the three measures have the same initial set of activated nodes.

We selected data sets which are easy to model as social networks for influence spread with a variety of sizes and properties. As in [28], large networks are taken from the SNAP's Stanford Large Network Dataset Collection. See Appendix www.cs.upc.edu/~mjblesa/ASONAM. 2021/FB-LTR-appendix.pdf for a brief description of them.

In Table I the characteristics of these datasets are summarized. In their description, besides number of nodes and edges, we use the following structural parameters. The *diameter* of the graph which is the longest shortest path between any pair of nodes. Some of the data sets have an infinitely large diameter due to the fact that these graphs are not (strongly) connected. The *Average Clustering Coefficient* (ACC) is computed by the average of the *local clustering coefficients* of all nodes, where a local clustering coefficient is the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. The *k-core* of a graph is the maximal subgraph such that every vertex has degree at least k. The *main core* (MC) is the core with the largest degree.

In order to get an appropriate representation of each of these social networks as an influence graph (G, w, θ), we need G to be a directed graph representing the relations of the network and we need to associate weight functions w and θ to the components of G. The undirected networks are understood as symmetric directed graphs. Thus, every undirected relation a, b between actors a and b is translated into an arc in both directions, i.e., (a, b) and (b, a). For the unweighted networks, all edges $e \in E$ were assigned weight $w_e = 1$. Due to the nature of the Caida network, a different mapping is considered for it. Its influence graph is constructed considering the following rules: An arc (a, b) with $w_{ab} = 1$ is placed whenever a is provider of b, or b is customer of a. When a and b are peers, we place both arcs (a, b) and (b, a) with $w_{ab} = w_{ba} = 1$. When two actors a and b are siblings, arcs are also placed in both directions but with $w_{ab} = w_{ba} = 2$.

None of the data sets have weights associated to the actors, therefore a node-weight function θ also needs to be defined. We consider θ as in [28]: for each actor $u \in V$, we set $\theta(u) = \lfloor \bar{w}_u/2 \rfloor + 1$, where $\bar{w}_u = \sum_{v \in N^-(u)} w_{vu}$. Thus, the activation value of an actor follows the simple majority rule.

V. Experiments and Results

In order to perform a pairwise comparison among the new and the already known centrality measures, we run

Table I: Summary of the data sets. ACC = Average Clustering Coefficient, MC = size of the main core. When the diameter is ∞, the diameter of the biggest connected component is provided. The symbol † states for the new data sets.

Data set	n	m	Directed?	Edge-weighted?	ACC	Diameter	MC
Caida†	26475	106762	yes	yes	0.2082	17	50
Dining-table	26	52	yes	yes	0.1178	∞ (6)	20
Epinions†	75879	508,837	yes	no	0.1378	14	422
Higgs	256491	328132	yes	yes	0.0156	19	10
Wikipedia†	7115	103689	yes	no	0.1409	7	336

three type of experiments with different indicators: (1) using summary statistics and rank-independent parameters, (2) using statistical correlation by means of the Spearman and Kendall coefficients, and (3) analysing the migration between fixed categories.

Analogously to the experiments in [28], we compare the results obtained by the proposed LT-based centrality measures to those used previously, namely, Deg, Btwn, Clsn, Ktz, PgR and ICR (see Section II). To compute the ICR, PgR and Ktz measures, we use the same parameters as in [28][2]. The measures were programmed in C++ and often ran in parallel. The experiments were run on the RDLAB-UPC computing cluster [1], using `HP ProLiant DL380p server` machines with two `Intel(R) Xeon(R) E5-2660`.

A. Summary statistics

For this set of experiments, the following global rank parameters have been considered: the standard deviation, the number of different values, and the Gini coefficient [4], [10]. The Gini coefficient of a vector $x \in \mathcal{R}^n$ is defined as:

$$\text{Gini}(x) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n\sum_{i=1}^{n}x_i}.$$

While the standard deviation is based on central tendency, i.e. deviation from the mean, the Gini coefficient is a general measurement of dispersion that intends to represent the inequality within a dataset. In contrast to the standard deviation, the Gini coefficient is invariant to scale and bounded in $[0, 1]$ (the closer the value to zero, the more equal the values are). When tested against different natural sparsity properties, the Gini index stands out as the best to measure sparsity [6], [9], [13], [34].

Due to space restrictions, the rank parameters for our data sets can be consulted online We can see that the Gini coefficients of the centrality measures are differing a lot. We can see that the more equal ranks are Clsn and ICR, while the most unequal one is Btwn. The measures on the `Higgs` network show the highest variability with respect to this coefficient (0.081 for ICR to over 0.99 for Btwn and Clsn). For the undirected networks, with few exceptions, the Gini coefficients are below 0.63, independently of the measure. A higher variability appears on the directed networks. FwLTR has smallest Gini coefficient than LTR and BwLTR, in particular for the `Higgs` network. Also, with the exception of `Epinions`, the FwLTR assigns less different values and has the smallest standard deviation. We can conclude that FwLTR assigns few values in an egalitarian way. Observe also that in `Caida`, FwLTR and BwLTR coincide, but they are different from LTR, this is somehow expected as the network is almost symmetric.

For most of the networks, the number of different rank values obtained is clearly largest for Btwn, Clsn, ICR and PgR, than for the other measures.

[2]The ICR measure is computed with probability $p = 0.1$, and the final rank is taken as the average over 100 executions. The PgR measure uses a damping factor $\alpha = 0.85$. The Ktz measure is used with values $\alpha = 0.1$ and $\beta = 1.0$. To make sure the Katz algorithm converges, the tolerance is reduced from the standard six significant digits to only two.

B. Correlation analysis

Let x and y be two lists of n users with x_i and y_i the rankings of the user i in lists x resp. y. The *Spearman's rank correlation coefficient* (ρ) [30] and the *Kendall's rank correlation coefficient* (τ) [17] are defined as follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1))} \quad \text{and} \quad \tau = \frac{n_c - n_d}{0.5n(n-1)},$$

where n_c is the number of concordant pairs (i, j) (i.e. such that either $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$) and n_d is the number of discordant pairs, i.e. those that are not concordant. The values of τ and ρ are in the interval $[-1, 1]$, where 1 means that both lists x and y have the same ranking and -1 means the rankings are the inverses of each other. Where the Spearman coefficient also depends on the relative distances between the rankings, the Kendall coefficient only considers the order of the rankings.

It is common to use statistic correlation for comparing the results of two centrality measures, being the Spearman's ρ and the Kendall's τ the most used coefficients. Results with a small Kendall and Spearman coefficient indicate that the output of the rankings differ. When both coefficients are high, this indicates that the rankings are similar.

Table II draws these coefficients for the networks considered in this article and compares all pairs of centrality measures. We can observe that most of the times the Spearman and Kendall coefficient are very similar. A reason for that might be that the relative distances change without changing the order. When performing a test to check whether two ranks are significantly different we took care also of the significance of the test. The results in color red correspond to test having a p-value greater than 0.05, indicating that there are not significative difference.

We can see a higher correlation among LTR and BwLTR in all the considered networks except in `Higgs`. We can also see a good level of correlation among BwLTR and Clsn (the

effect is even stronger in Wikipedia). The FwLTR seems to be closer to ICR, although the correlation indices are not so high as those among BwLTR and Clsn. Finally, let us observe that in the Caida network the correlation among the measures is different, but we can observe a perfect correlation among FwLTR and BwLTR. We suspect that this might be due to the symmetry of the Caida network. Observe also that in Caida the correlation index with LTR is still high.

Table II: Correlation coefficients for the centrality measures on the networks Caida, Epinions, Higgs and Wikipedia. The Kendall coefficients (τ) are shown in the upper triangular part. The Spearman coefficients (ρ) in the lower triangular part. Results with a p-value bigger than 0.05 are shown in red. For the Katz measure, a – indicates non-convergence.

Caida

ρ \ τ	BwLTR	FwLTR	LTR	Btwn	Clsn	Deg	ICR	Ktz	PgR
BwLTR	1	1	0.896	0.231	0.673	0.364	0.538	–	0.253
FwLTR	1	1	0.896	0.231	0.673	0.364	0.538	–	0.253
LTR	0.982	0.982	1	0.227	0.724	0.362	0.566	–	0.226
Btwn	0.289	0.289	0.284	1	0.232	0.719	0.338	–	0.652
Clsn	0.862	0.862	0.902	0.299	1	0.399	0.625	–	0.193
Deg	0.470	0.470	0.469	0.803	0.510	1	0.555	–	0.812
ICR	0.735	0.735	0.765	0.445	0.816	0.683	1	–	0.326
Ktz	–	–	–	–	–	–	–	–	–
PgR	0.376	0.376	0.339	0.808	0.308	0.926	0.491	–	1

Epinions

ρ \ τ	BwLTR	FwLTR	LTR	Btwn	Clsn	Deg	ICR	Ktz	PgR
BwLTR	1	0.251	0.779	0.525	0.776	0.550	0.165	–	0.595
FwLTR	0.323	1	0.511	0.573	0.250	0.638	0.744	–	0.272
LTR	0.884	0.586	1	0.542	0.651	0.636	0.357	–	0.490
Btwn	0.637	0.691	0.666	1	0.532	0.731	0.530	–	0.615
Clsn	0.916	0.350	0.810	0.653	1	0.555	0.207	–	0.619
Deg	0.659	0.755	0.749	0.834	0.679	1	0.584	–	0.632
ICR	0.240	0.883	0.472	0.661	0.310	0.712	1	–	0.211
Ktz	–	–	–	–	–	–	–	–	–
PgR	0.766	0.389	0.647	0.733	0.798	0.767	0.306	–	1

Higgs

ρ \ τ	BwLTR	FwLTR	LTR	Btwn	Clsn	Deg	ICR	Ktz	PgR
BwLTR	1	-0.270	0.380	0.428	0.803	0.374	-0.214	0.805	0.788
FwLTR	-0.295	1	0.536	0.183	-0.390	0.499	0.660	-0.390	-0.391
LTR	0.402	0.544	1	0.384	0.381	0.923	0.324	0.383	0.370
Btwn	0.445	0.195	0.405	1	0.519	0.407	0.139	0.518	0.508
Clsn	0.818	-0.423	0.405	0.535	1	0.379	-0.286	0.996	0.977
Deg	0.395	0.513	0.936	0.425	0.399	1	0.332	0.381	0.372
ICR	-0.264	0.764	0.382	0.168	-0.350	0.391	1	-0.286	-0.287
Ktz	0.818	-0.423	0.406	0.535	1.000	0.401	-0.351	1	0.979
PgR	0.816	-0.428	0.397	0.528	0.998	0.395	-0.355	0.998	1

Wikipedia

ρ \ τ	BwLTR	FwLTR	LTR	Btwn	Clsn	Deg	ICR	Ktz	PgR
BwLTR	1	-0.004	0.640	0.674	0.936	0.628	-0.016	–	0.937
FwLTR	0.012	1	0.503	0.414	-0.016	0.508	0.830	–	-0.013
LTR	0.729	0.531	1	0.531	0.613	0.983	0.390	–	0.615
Btwn	0.735	0.482	0.615	1	0.659	0.530	0.372	–	0.676
Clsn	0.988	-0.006	0.720	0.725	1	0.604	-0.026	–	0.924
Deg	0.718	0.541	0.998	0.612	0.709	1	0.397	–	0.607
ICR	-0.004	0.932	0.469	0.457	-0.020	0.479	1	–	-0.023
Ktz	–	–	–	–	–	–	–	–	–
PgR	0.988	-0.001	0.721	0.734	0.990	0.711	-0.015	–	1

C. Categorization analysis

In order to further deepen into the comparative analysis of the measures, we apply a new study of inequality to the actors' rankings once the different centrality measures are applied. This time the analysis is done by means of *categorization*, which is a common analysis method in sociology (see, e.g. [25]). For each centrality measure and network, we classify the actors in three *position* categories. For doing so we first sort the actors in decreasing order of rank. Then, following the terminology used in sociology, we define three categories: *Top* (high class), which consists of the top 10 percent actors, *Low* (low class) containing the 50 percent least performing actors, and *Mid* (middle class) that contains the 40 percent of actors in between.

Some actors of a given network might fall into different categories when ranked according to different centrality measures. We will focus our attention on what and how migrations occur when changing the underlying centrality measure. This provides and additional insight on the nature of the differences. We depict those migrations by showing one category, in terms of percentages of the categories of the other measures, it is composed of. Those percentages provide the migration levels from the three categories in one measure to the considered category in the target measure. A drawing of all the migrations from BwLTR, FwLTR, and LTR to and from the remaining measures can be found online.

We can observe that the majority of migrations occur between adjacent classes and in percentages that are not very high. The differences in percentages when the Top class is involved, with respect to those of Mid and Low, follow naturally form the big difference in their sizes. The percentages in migrations among the Mid and Low categories are more similar, of course the corresponding sizes are also not far away. Direct migrations between categories Top and Low, or vice-versa, with high percentages do not appear in many networks. For Epinions almost all the migrations are below 35%. The highest percentages are among classes Mid and Low and migrations from Low to Top pretty low (see Figure 1).

Furthermore, migrations with a high percentage of actors only occur in very few networks and among few measures. We observe this behavior is Higgs (one of the largest in the collection). The extreme behavior apppears only when going from BwLTR to FwLTR, and from BwLTR to ICR and in smaller percentages (58% and 65%). The fact that the inequality coefficient gini for FwLTR and ICR is lower than that of BwLTR, and that both ICR and FwLTR have little correlation with BwLTR in this network could explain this big migration of actors from the superior to the inferior category. Also, we can observe that the lower is the Gini coefficient (the gini of ICR is very close to 0) more actors seem to move among extreme categories.

VI. Conclusions and Future Work

The starting point of this paper was the Linear Threshold Rank (LTR) introduced in [28]. This is a measure to determine the importance of nodes within a social network based on the Linear Threshold model. To give a better representation of social networks, we propose to use the definition with only successors or only predecessors in the initial activation set, instead of considering both (as done in [28]). This decision was also reinforced by the fact that we observed that considering both sets of neighbours might often result in improper height of the rankings. Our results show that the initial activation of a node and its predecessors (BwLTR), provides a ranking similar to the one produced when activating also its successors (LTR). This suggests that the predecessors are somehow shadowing

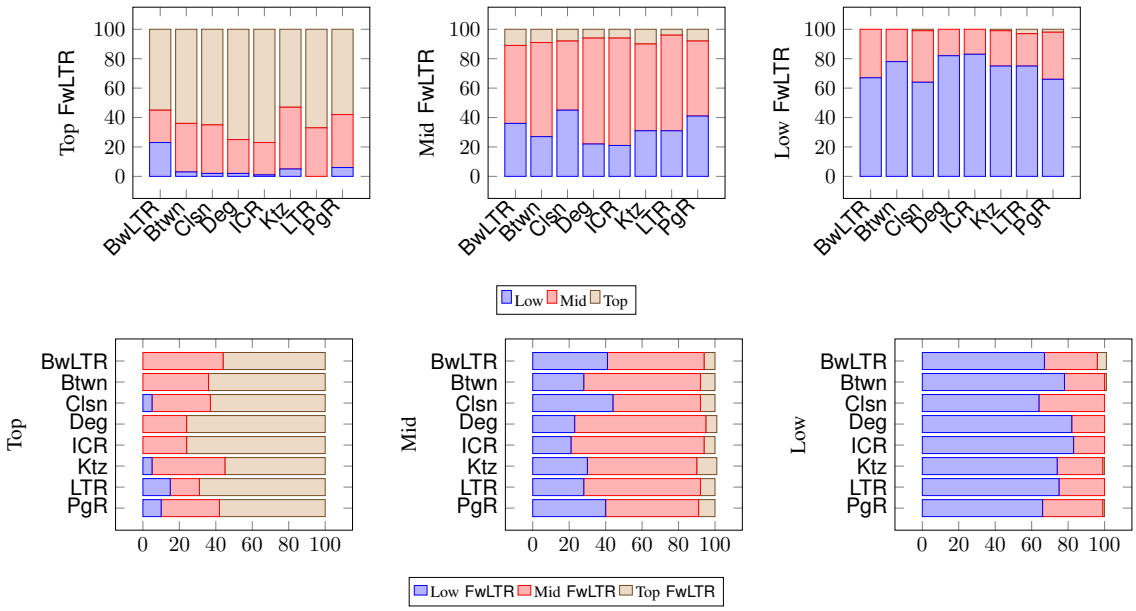

Figure 1: `Epinions` network: Migrations for the position classification for the FwLTR.

the real influence of the node, which is better captured by the FwLTR.

We decided to contrast the different centrality measures among them. For that aim, we selected a group of eight networks as dataset for our experiments. Those networks represent different types of social networks with different characteristics. We applied different methods to analyse the results of the different centrality measures on the dataset. In addition to the classical statistical measures, we applied two new methods coming from other areas of science. First we applied the Gini coefficient, which is mostly known from the field of sociology as a measure of the inequality among populations. We saw that it is hard to directly use the Gini coefficient of a centrality measure as an indicator of the social network, since those values are varying per centrality measure and no fixed trends could be found. Second, we applied an inequality study based on the social tables and political arithmetic of Denis Diderot of the 18th century. For the application of this method, we divided the actors three categories based on the outcome of a centrality measure. That provides us with a 3-stages classification of the actors according to their power of influence. We payed special attention to the migration of nodes within categories when switching to another underlying centrality measure. This method seems to be a promising tool in the analysis of different centrality measures since, e.g., it provides us with information on how robust is the category of top-influencing actors when considering different centrality measures.

For further research, it sounds specially interesting to observe centrality measures for influence graphs where a positive and a negative seed try to spread their (opposite) influence. This can model for example political campaigns where two candidates try to reach as many votes as possible, a referendum or another yes or no voting. How can the LTR be defined for such models, how to choose the parameters? Probably initial activation sets for both opponents need to be defined. The idea of spreading opposing opinions can even be generalized to k parties trying to spread their different opinions. How can such influence spread be modelled? And how can we adapt the centrality measures to this model? Here we come close to the field of game theory, where *influence games* are defined [22]. A coalition of actors $X \subseteq N$ is defined as a winning coalition if this coalition activates more actors than a given threshold (so probably X could be used as the initial activation set). More types of influence games can be developed and investigated. Besides trust relationships in social networks, there exist also models where distrust relationships are included. Another idea for further research is to investigate how to model them and how do centrality measure behave on them.

One of the drawbacks of our study is that we could not find datasets where thresholds were included. In fact, we could not find data reporting the process of activation in a network. For research on influence spread it would be very useful to be able to have real datasets with this kind of information. In our experiments, we fixed the thresholds for the resistance to be the single majority criteria. Other thresholds were considered in [27]. It would of interest to analyze what determines the best threshold for a network.

In the LT model, the type of information which is spread is not taken into account. The model behaves independent of the message which is spread. Regard for example a retweeting Twitter network. The threshold to retweet something about a terrorist attack nearby is probably lower than the threshold for retweeting a message about the weather, because of the

urgency and emotional load of the information. It might be interesting to adapt the behaviour of the influence spread model to the type of information spread. One might also think on extending the LT model in order to incorporate the devaluation of the opinions over time, or enriching it with the introduction of some adversarial element on the diffusion process that maliciously modifies the final influence.

Finally, It might be interesting to design a model for influence spread which is a combination of the two mostly used models: the deterministic Linear Threshold model and the stochastic Independent Cascade model. The idea is to combine the best of both models. One possible way is that each actor i is certainly activated if the incoming influence exceeds its threshold $f(i)$, but, when this is not the case, there is still a probability that actor i is activated. This probability can be made larger when the incoming influence is closer to the threshold. The probability that an actor i is activated in iteration t can, for example, be defined as follows:

$$P(i \in F_t(X)) = \exp(-\max\{f(i) - \sum_{j \in F_{t-1}(X)} w_{j,i},\ 0\}/c).$$

where $F_t(X)$ is the spread level of initial activation set X in time step t, $w_{j,i}$ is the weight of the edge (j, i) and c is some parameter which determines the decrease of the probability when the incoming influence is lower than the threshold. The larger c, the bigger the probability that an actor is activated although the threshold is not exceeded. The idea for this influence model is based on *Simulated Annealing* [2], which belongs to a class of local search algorithms that are known as *threshold algorithms*. In these algorithms parameter c is often called a *cooling parameter*.

Any future work should also consider a larger dataset for experiments, to provide us with additional information about what centrality measure better identifies the important spreading nodes for every type of network.

ACKNOWLEDGEMENTS

We thank Eline van Hove for the preliminary results on this topic while visiting the Computer Science Department at UPC.

REFERENCES

[1] Research & Development Lab (**rd**lab). Universitat Politècnica de Catalunya - BarcelonaTech. rdlab.cs.upc.edu. Accessed: 2020-06.

[2] E.H.L. Aarts,H.M. Korst and P.J.M. van Laarhoven. Simulated annealing. *Local Search in Combinatorial Optimization*, 91–120, 1997.

[3] A. Bavelas. A Mathematical Model for Group Structures. *Human Organization: Summer*, 7(3):16–30, 1948.

[4] L. Ceriani and P. Verme. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10(3):421–443, 2012.

[5] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. The anatomy of a scientific rumor, Higgs data set. *Scientific Reports 3*, 2980, 2013.

[6] F.G. De Maio. Income inequality measures. *Journal of Epidemiology & Community Health*, 61(10):849–852, 2007.

[7] W. De Nooy, A. Mrvar, and V. Batagelj. Exploratory social network analysis with pajek. *Cambridge University Press Chapter 1*, 2004.

[8] L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[9] David E. A. Giles. Calculating a standard error for the Gini coefficient: Some further results. *Oxford Bulletin of Economics and Statistics*, 66(3):425–433, Wiley 2004.

[10] C. Gini. *Variabilitàe Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, 1912.

[11] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Technical Report, Academy of Marketing Science Review*, 2001.

[12] D. Gómez, J.R. Figueira, and A. Eusébio. Modeling centrality measures in social network analysis using bi-criteria network flow optimization problems. *European Journal of Operations Research*, 226:354–365, 2013.

[13] N. Hurley and S. Rickard. Comparing measures of sparsity. ArXiv:0811.4706v2 [cs.IT] 27 Apr 2009.

[14] L. Katz. A new status index derived from sociometric analysis. *Psychometrik 18*, 39–43, 1953.

[15] D. Kempe, J.M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. *Theory of Computing*, 11:105–147, 2015.

[16] D. Kempe, J.M. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. *Intl. Colloquium on Automata, Languages and Programming, ICALP 2005*, Lecture Notes in Computer Science, vol. 3580, 1127–1138, 2005.

[17] M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[18] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. 2010.

[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. 2010.

[20] J. Leskovec, J. Kleinberg, and C. Faloutos. Graphs over time: Densification laws, shrinking diameters and possible explanations. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2005*, 177–187, 2005.

[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[22] X. Molinero, F. Riquelme and M. Serna. Cooperation through social influence. *European Journal of Operational Research*, 242:960–974, 2014.

[23] M.E.J. Newman. Networks: An Introduction. *Oxford University Press*, 2010.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, technical report. *Stanford Digital Library*, 1999.

[25] T. Piketty. *Capital in the Twenty-First Century*. Belknap Press: An Imprint of Harvard University Press, 2015.

[26] M. Richardson, R. Agrawal, and P. Domingos. Trust Management for the Semantic Web. *Intl. Semantic Web Conference (ISWC)*. Lecture Notes in Computer Science, vol 2870, 351–368. Springer. 2003.

[27] F. Riquelme, P. Gonzalez-Cantergiani, X. Molinero, and M. Serna. The neighborhood role in the linear threshold rank on social networks. *Physica A*, vol. 528, 21430:1–10, 2019.

[28] F. Riquelme, P. Gonzalez-Cantergiani, X. Molinero, and M. Serna. Centrality measures in social networks based on linear threshold model. *Knowledge-Based Systems*, 40:92–102, 2017.

[29] P. Shakarian, A. Bhatnagar, A. Aleali, E. Shaabani, and R. Guo. *The Independent Cascade and Linear Threshold Models*. Diffusion in Social Networks, Chp. 4. SpringerBriefs in Computer Science. Springer, 2015.

[30] C. Spearman. The proof and measurement of association between two things. *AM. J. Psychol*, 15:88–103, 1904.

[31] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz. The budapest reference connectome server v2.0. *Neuroscience Letters*, 595:60–62, 2015.

[32] B. Szalkai, C. Kerepesi, B. Varga, and V. Grolmusz. Parameterizable consensus connectomes from the human connectome project: The budapest reference connectome server v3.0. *Cognitive Neurodynamics*, 2016.

[33] X. Weng, Z. Liu, and Z. Li. An efficient influence maximization algorithm considering both positive and negative relationships. *2016 IEEE Trustcom/BigDataSE/ISPA*, 1931–1936, Tianjin, 2016.

[34] S. Yitzhaki. Gini's mean difference: a superior measure of variability for non-normal distributions. *METRON - International Journal of Statistics*, LXI(2):285–316, 2003.