Measuring Competence: Improvements to Determine the Degree of Opinion Leadership in Social Networks

Michael Spranger*†, Kai-Jannis Hanke†, Florian Heinke† and Dirk Labudde†‡

†University of Applied Sciences Mittweida
Forensic Science Investigation Lab (FoSIL), Germany
Email: *name.surname*@hs-mittweida.de
‡Fraunhofer
Cyber Security
Darmstadt, Germany
Email: labudde@hs-mittweida.de

***Abstract*—In recent years, the automated, efficient and sensitive monitoring of social networks has become increasingly important for the criminal investigation process and crime prevention. Previously, we have shown that the detection of opinion leaders is of great interest in forensic applications to gather important information. In the current work, it is argued that state of the art methods, determining the relative degree to which an opinion leader exerts influence over the network, have weaknesses if networks exhibit a star-like social graph topology, whereas these topologies result from the interaction of users with similar interests. This is typically the case in networks of political organizations. In these cases, the underlying topologies are highly focused on one (or only a few) central actor(s) and lead to less meaningful results by classic measures of node centrality commonly used to ascertain the degree of leadership. With the help of data collected from the Facebook and Twitter network of a German political party, these aspects are examined and a quantitative indicator for describing star-like network topologies is introduced and discussed. This measure can be of great value in assessing the applicability of established leader detection methods. Finally, two variations of a new measure– the CompetenceRank – which is based on the LeaderRAnk score and aims to address the discussed problems in cases with and without additional network data such as *likes* and *shares*, are proposed.**

Keywords–Forensic; Opinion Leader; Graph Theory.

I. Introduction

The detection of opinion leaders in online social networks has been discussed extensively over the past few years. While the term "detection" is generally associated with a binary decision, here – in accordance with other papers in this domain – it is used to refer to the determination of the degree of leadership. The scope of application is manifold and reaches from determining influencers and brand ambassadors up to finding those who influence the political opinion of a group of people. Especially the last application can be of interest to law enforcement and intelligence agencies. In [1] it was shown that in some situations previous approaches based on the work by Katz [2], who focused on networks in the offline world, do not capture the core of the problem and as a result lead to an inaccurate assessment of opinion leadership.

Measures for opinion leadership on social networks tend to focus on a single aspect: network contribution. However, it becomes clear that only evaluating network contribution such as posting content, commenting it or replying to it does not capture the full range of interactions social media platforms have to offer. Besides network contribution or content generation in the ordinary sense we also find a secondary form of participation, which solely relies on existing content. Virtually nodding in agreement by clicking *like* or extending the reach of a given post by sharing it, is not creating new content in a given network. However, measures reflecting such activities exist on most social media platforms and play a substantial role in determining ones reach and authority. These secondary measures do not only shape how people interact but also influence who rises to the position of an opinion leader.

This section shall give a brief introduction to the field in which situations may occur, in which the LeaderRank leads to inappropriate results. Furthermore, it will give an overview of topology-based approaches and it finishes with the scope and structure of the paper.

A. General Motivation

Analyzing social networks has become an important tool for investigators, intelligence services and decision makers of police services. The information gained this way can be used to solve crimes by searching for digital evidence that relates to the crime in the real world. Additionally, methods of predictive policing can help to organize police missions as was shown in [3]–[5]. The detection of opinion leaders in social networks is an important task for different reasons. On the one hand, owners of influential profiles are often also influential in the offline world. Knowing these people helps to determine the direction of an investigation or more concretely to target persons of interest. On the other hand, as was suggested in previous work [5], it might be of interest to contact these profiles by means of chatbots to gain access into closed groups in an effort to gather important information for intelligence services. Intuitively, opinion leaders, when considered as nodes with high structural importance, can be detected with the help of centrality measures. However, different kinds of influence in a network have to be distinguished. Nodes can have a great influence as corresponding actors are able to spread information fast and widely in a network, or they can have a great influence because they write something of importance that attracts many other users in the network to respond.

B. Leader Detection by means of Network Centrality Measures

In the literature, one can mainly find centrality measures for the former type of influence. For example, highly active profiles can be recognized using degree centrality, meaning, the relative number of outgoing edges of a node. These profiles are represented by nodes with a high degree centrality and are especially useful to spread information in a network due to their high interconnectedness. In this context, the closeness centrality – the inverse of the mean of the shortest path of a node to any other node in the network – is even more effective. It describes the efficiency of the dissemination of information of a certain node.

Furthermore, the betweenness centrality of a certain node, which is defined as the number of shortest paths between two nodes that cross this node, describes the importance of this node for the dissemination of information in a network. Therefore, the higher the betweenness centrality of a node, the greater its importance for the exchange of information in a network.

Moreover, the eigenvector centrality of a node is defined as the principal eigenvector of the adjacency matrix of a network. In contrast to the measures discussed beforehand, PageRank [6], as one of the best measures of node centrality, does not only consider the centrality of the node itself, yet also of its neighboring nodes.

As part of the opinion leader detection research, LeaderRank [7] was introduced as a further development of PageRank in order to find nodes that spread information further and faster. However, all of these centrality measures consider nodes that are involved in the dissemination of information mainly based on their activity. For the purpose of the intended usage, users who achieve high impact through what they have written are of much greater interest. Thus, similar to the citation of papers and books and its impact on the author's reputation, the importance of a node has to be higher when it reaches a high number of references and citations with low activity.

Especially social media platforms provide comparable metrics, such as *likes* and *shares* that partially reflect the author's reputation and credibility. Hence, it is imperative to consider respective measures of acceptance, expertise and authority when determining opinion leaders in any digital social network.

Interestingly, Li et al. considered the so-called node spreadability as the ground truth for quantifying node importance in a subsequent study [8]. Subsequently, node spreadability is based on a straightforward Susceptible-Infected-Removed (SIR) infection model from which the expected number of infected nodes upon initially infecting the node in question is estimated. However, this expected number can only be estimated from simulation, which, furthermore, is dependent on the parameterization of the SIR model. In this respect, all centrality measures can be considered as heuristic approximations of node spreadability.

C. Scope and Structure of the Paper

In this work, we discuss problems that can arise when aiming to detect opinion leaders in social networks yielding highly central topologies similar to star graphs. Examples for such networks are especially group pages on Facebook or vk.com where user interactions and activities are mostly triggered by and focused on posts made by the page owner. In such cases, the page owner – a trivial leader in the sense of centrality measures discussed above – acts as a score aggregator and can thus lead to distorted scoring, which can eventually be adverse in the context of opinion leader detection. In this case, classic centrality measures can be considered inappropriate. Based on interactions of users of the Facebook page of the German political pary "DIE LINKE" tracked for five consecutive months (January - May 2017), this problem is illustrated. We further introduce the LeaderRank skewness as a quantitative measure of aggregator-induced distorted LeaderRank scoring, which in experiments show to be superior to network entropy with respect to expressiveness. Additionally, a simple modified LeaderRank score, to which we refer to as CompetenceRank, is introduced. It is proposed to be more suitable for opinion leader detection in such networks, especially, if additional data for *likes* and *shares* are not available.

For such cases in which these data is available an improved version of the CompetenceRank is proposed and evaluated using the Twitter network of "DIE LINKE". The corresponding data set contains not only tweets, comments and replies from the entire year 2018, it also incorporates the accompanying *like* and *retweet* counts for each tweet, comment and reply. In politically motivated networks, as the one analyzed in this paper, the improved CompetenceRank shows a substantial increase in performance compared to the LeaderRank and the simple CompetenceRank.

The paper is structured as follows: in Section II, a brief literature overview on the topic of opinion leader detection is given, followed by a summary of the LeaderRank algorithm. In Section III two shortcomings of the LeaderRank are discussed: firstly, the skewness of the rank distribution in star-shaped network topologies and, secondly, that not all available data of social media platforms are taken into account. Subsequently, in the same section the deduction and definition of the normalized LeaderRank skewness as a metric for an approximation of a star-shaped topology is discussed and compared with the normalized graph entropy. In Section IV three datasets are introduced, which were used to evaluate these metrics, two of which were also used to develop solutions for the aforementioned problems as proposed in Section V by introducing the CompetenceRank for taking authority into account as well as an improvement for cases in which additional data is available. Subsequently, Section VI contains an evaluation of both CompetenceRank versions using the Twitter network. Finally, a conclusion as well as an overview of future work is given in Section VII.

II. Detection of Opinion Leaders

Opinion leaders in the context of the intended analysis of social networks are individuals, who exert a significant amount of influence on the opinion and sentiment of other users of the network through their actions or by what they are communicating. In social sciences the term "opinion leader" was introduced before 1957 by Katz and Lazarsfeld's research on diffusion theory [2]. Their proposed two-step flow model retains validity in the digital age, especially in the context of social media.

Katz et al. assume that information disseminated in a social network is received, strengthened and enriched by opinion

leaders in their social environment. Each individual is influenced in his opinion by a variety of heterogeneous opinion leaders. This signifies that the opinion of an individual is mostly formed by its social environment. In 1962, Rogers referenced these ideas and defined opinion leader as follows:

> "Opinion leadership is the degree to which an individual is able to influence informally other individuals' attitudes or overt behavior in a desired way with relative frequency." [9, p. 331]

For the present study, one important question to answer is what influence means, or rather how to identify an opinion leader or how the influencer can be distinguished from those being influenced. Katz defined the following features [2]:

1) personification of certain values,
2) competence,
3) strategic social location.

One approach to identify opinion-leaders is to extract and analyze the content of nodes and edges of networks to mine leadership features. For instance, the sentiment of communication pieces can be analyzed to detect the influence of their authors, as shown by Huang et. al., who aim to detect the most influential comments in a network this way [10]. Another strategy is to perform topic mining to categorize content and detect opinion leaders for each topic individually, as opinion leadership is context-dependent [2] [11]. For this purpose, Latent Dirichlet Allocation (LDA) [12] can be used, as seen in the work of [13]. Furthermore, Aleahmad et. al. achieved good results with OLFinder by utilizing both topic mining methods and centrality measures [14]. Additionally, Chen et al. proposed D_OLMiner, which derives opinion leaders from dynamic social networks [15].

Another novel approach, the firefly algorithm, a meta-heuristic optimization algorithm that can deal with especially large networks, is based on the behavior of fireflies and is used by Jain et. al. to determine local and global opinion leaders [16].

For this study, we considered the implementation of content-based methods problematic, as texts in social networks mostly lack correct spelling and formal structure which impairs such methods' performance. Additionally, leaders can be identified by analyzing the flow of information in a network. By monitoring how the interaction of actors evolves over time, one can identify patterns and individuals of significance within them. To achieve this, some model of information propagation is required, such as Markov processes employed by [17] and the probabilistic models proposed by [18]. These interaction-based methods consider both topological features and their dynamics over time. DDOL is a recent, dynamic approach by Queslati et. al. that focuses on social signals (shares, comments, likes) and terms that are frequently encountered in the expression of opinions. DDOL does not include centrality measures and has a slightly lower precision than PageRank but contrary to PageRank it works on dynamic networks and a has a lower computational complexity [19].

Parts of this study use methods that are solely based on a network's topology, therefore, considering features, such as node degree, neighborhood distances and clusters, to identify opinion leaders. One implementation for the former is the calculation of node centrality. The underlying assumption is that the more influence an individual gains, the more central it is in the network. Which centrality measure is most suitable is dependent on the application domain. We judged eigenvector centrality to be most adequate. One of the most popular algorithms is Google's PageRank algorithm [6]. The application of PageRank for the purposes of opinion leader detection has seen merely moderate success [20] [21].

With LeaderRank scores, Lü et al. advocate further development and optimization of this algorithm for social networks, and have achieved surprisingly good results [7]. Herein, users are considered as vertices and directed edges as relationships between opinion leaders and users. All users are also bidirectionally connected to a ground vertex, which ensures connectivity as well as score convergence. In short, the algorithm is an iterative multiplication of a vector comprised by per-vertex scores $s_i(t)$ at iteration step t with a weighted adjacency matrix until convergence is achieved according to some convergence criteria. Initially, at iteration step t_0, all vertex scores are set to $s(0) = 1$, except for the ground vertex score which is initialized as $s_g(0) = 0$. Equation (1) describes the LeaderRank algorithm as a model of probability flow through the network, where $s_i(t)$ indicates the score of a vertex i at iteration step t.

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{a_{ji}}{e_{v_j}^{out}} s_j(t) \tag{1}$$

Depending on whether or not there exists a directed edge from vertex j to the vertex i, the value 1 respectively 0 is assigned to a_{ji}. $e_{v_j}^{out}$ describes the number of outgoing edges of a vertex j. The update rule given in Equation (1) can be rewritten as a matrix-vector product:

$$\mathbf{s}(t+1) = \tilde{\mathbf{A}}\mathbf{s}(t), \tag{2}$$

where $\mathbf{s}(t)$ corresponds to the vector of the $N+1$ vertex scores at iteration step t, and $\tilde{\mathbf{A}}$ is the weighted adjacency matrix of size $(N+1) \times (N+1)$ with

$$\tilde{A}_{ji} = \frac{a_{ji}}{e_{v_j}^{out}}. \tag{3}$$

The final score is obtained as the score of the respective vertex at the convergence step t_c and the obtained ground vertex score, as shown in (4). At t_c, equilibration of LeaderRank scores towards a steady state is observed.

$$\mathrm{S}_i = s_i(t_c) + \frac{s_g(t_c)}{N} \tag{4}$$

Furthermore, note that

$$\sum_{i=1}^{N} S_i = \sum_{i=1}^{N} s_i(t) = N. \tag{5}$$

The advantage of this algorithm compared to PageRank is that the convergence is faster and, above all, that vertices that spread information faster and further can be found. In later work, for example, by introducing a weighting factor, as in [8] or [22], susceptibility to noisy data has been further reduced and the ability to find influential distributors (hubs) of information has been added.

III. Issues with LeaderRank

The LeaderRank algorithm can be understood as a reversion of a discrete model of diffusion. In that sense, the initialization $s_i(0) = 1$ at t_0 can be interpreted as assigning a uniform concentration distribution of some virtual compound that, in the processes, is re-distributed according to the model. In that respect, central actors showing the highest activity in star-like networks can induce score aggregation and migration towards their central nodes as well as their adjacent nodes, whereas nodes in the 'peripheral region' of the network become inadequately represented by their scores. Therefore, one can hypothesize that ranked lists obtained from LeaderRank scores can not be considered meaningful if a given network in question exhibits a star-like topology.

Another problem of LeaderRank comes into existence when considering means of communication that differ from traditional ones in person dialogues. Most social media platforms utilize *likes*, *shares*, *dislikes* and the concept of building a follower base. The amount of, for example, *likes* that a post receives or the frequency with which it is shared indicate its importance within a network and at least partially reflect the influence of the respective author. In turn, such data should be included when determining opinion leadership. Theoretically, LeaderRank has the capacity to incorporate aforementioned additional data. However, if this data were to be included in a network graph, then each *like*, *share* or anything similar would be seen as a unique edge from one node to another, just like regular forms of communication. This introduces two major problems, a theoretical one and a practical one. Firstly, is a *like* on a post equally as valuable as an actual reply and then how influential is a *share*? Evidently, there is a difference between the interaction activities, such as liking, sharing, writing or replying to a post, but this discrepancy is difficult to capture with the LeaderRank. Either one accepts that *likes* and *shares* have similar value to a written reply or one needs to additionally implement weights for different types of edges within a network.

Secondly, including *likes* as edges between nodes poses a practical problem: partial networks. When considering an individual post, then ideally the name of every individual who has liked this post is available in our data set, but in a real world example this is usually not the case. For example, when analyzing a twitter network one can discover how many people liked an individual post quite easily, but recovering the names of those individuals is highly restricted as twitter only provides a shortened list of names. It might be possible to recover all the names for a tweet with only 15 *likes*, but the list of names for a tweet with 100 *likes* can have the same length as the list for a tweet with 1.000 *likes*. Clearly, we lose a significant amount of information with exactly those tweets that are of great interest for opinion leadership, that is, tweets with seemingly the most influence over other users. When faced with similar restrictions on different platforms the total count of *likes* or *shares* might be more useful than a drastically reduced and limited list of names. In a similar manner it makes more sense to determine the popularity of politicians by counting the attendees of a political event compared to getting the names of only the first hundred attendees. Hence, it makes more sense to define people posting on social media as "politicians" speaking on a stage whereas users liking or sharing their content can be seen as attendees nodding in agreement or sending pictures of the stage to their friends.

On social media we have many attendees, virtually nodding their heads by clicking *like* or retweeting or sharing interesting content but they do not contribute by producing new posts. Incomplete data sets may not include the name for every person that likes a contribution, but these users can still be influenced and may even shape the network, since *likes* and *shares* present a measure for authority, credibility and approval in a given network. As a result, accounts partaking in the network through *likes* and *shares* should receive recognition as they silently enable cognitive biases, like the bandwagon effect [23] or herding mentality [24], that in turn alter how well-liked content appears to be, consequently, making it more or less influential. Ideally, LeaderRank does not only find opinion leaders in complete networks, but also discovers them in incomplete data sets. As a result, accounts that cannot be represented in the graph due to the absence of a name should still be considered when determining opinion leadership. A magnitude of nameless accounts cannot be included in a graph and thus they will not receive LeaderRank-Scores themselves, but seen as a collective they may help in shaping a network and identifying truly influential opinion leaders.

In this case study, two different networks are being examined. Namely, the network around the Facebook page as well as the Twitter network of the German left-winged political party "DIE LINKE". Firstly, the star topology of the Facebook network is being evaluated and secondly a novel approach to include *likes* and *retweets* is tested on the Twitter network.

In the first case study, the Facebook network under investigation shows an extreme case of a star topology in which the owner of the political Facebook page "DIE LINKE" acts solely as the central actor (for more information see Section IV). Since the LeaderRank emphasizes the strategic social location of a user, their competence seems to be improperly valued. In star-shaped network topologies, high centralities of only a fraction of nodes leads to a heavily skewed LeaderRank score distribution.

In contrast, one could argue that someone is more important if any activity generates a high number of responses. Such a case is regularly given by political networks which are dominated by the central node of the page owner. Consequently, a straightforward modification of the LeaderRank score is proposed in Section V-A addressing the imbalance the LeaderRank algorithm yields in such networks.

In the following paragraph a quantitative measure of LeaderRank distribution skewness is proposed that could aid to ensure proper applicability of the LeaderRank algorithm for any given network. This measure is further compared to the classic measure of network entropy. Tests on simulated data show the LeaderRank skewness to be superior to network entropy with respect to topological changes.

A. Definition of LeaderRank Distribution Skewness

Let $LR = \{S_1, ..., S_i, ..., S_N\}$ be the LeaderRank scores of all nodes. Further, $\overline{S}$ and sd_{LR} denote the arithmetic mean and standard deviation of LR. Based on the z-scaled LeaderRank scores (6), the skewness ν of the LeaderRank distribution is calculated as shown in (7).

$$z(S_i) = \frac{S_i - \overline{S}}{sd_{LR}} \tag{6}$$

$$\nu_{LR} = \left| \frac{1}{N} \sum_i z(S_i)^3 \right| \quad (7)$$

As discussed above, score distribution skewness is correlated with network topology. Yet, normalization of computed skewness is required in order to make a statement about the topology and whether a star-like topology is present. Hence, upper and lower bounds, ν_{min} and ν_{max}, are needed. In this paragraph, derivation of both bounds are given.

Trivially, ν converges to the lower bound – the theoretical minimum ($\nu = 0$) – in almost-regular graphs. Such graphs are regular graphs with one edge being removed. With N being sufficiently large, the supposition that $S_i \approx S_j$ for any pair of randomly selected vertices of a social network graph $v_i, v_j \in V$ holds true and a limit of $\lim_{sd_{LR} \to 0} \nu = 0$ can be assumed. In regular graphs however all LeaderRank scores are equal by definition, resulting to $sd_{LR} = 0$ and ν being undefined in this case.

In contrast, ν is equal to the theoretical maximum if the network graph exhibits a strictly star-shaped topology. Directed star graphs are graphs with a central vertex v_c and $N-1$ leaf vertices connected to v_c. One can re-write the set of star graph vertices as $V = \{v_c, v_2, ..., v_N\}$ and denote the LeaderRank score set as $LR = \{S_c, S_2, ..., S_N\}$.The LeaderRank scores of any randomly selected pair of vertices v_i and v_j with $v_i, v_j \neq v_c$, with v_c being the central vertex, are then not distinguishable, i. e., $S_i = S_j$, according to the LeaderRank's definition. Furthermore, the sum of LeaderRank scores equals N leading to $\overline{S} = 1$ for any given graph. Given the central node's score S_c, each S_i can thus be calculated as shown in (8).

$$S_i = \frac{N - S_c}{N - 1} \quad (8)$$

Thus if S_c is known, the set of LeaderRank values $\{S_c, S_2, ...S_i, ..., S_N\}$ and the resulting ν_{max} can be derived. In the following text we shall give an explicit relationship between the number of nodes N in a directed star graph and the corresponding score set LR. For this, let $\mathbf{s}$ be the scores vector at the steady-state to which $\mathbf{s}(t)$ converges according to the update rule (see Equation (2)). Then the identity given in Equation (9) holds, since $\mathbf{s} = \mathbf{s}(t+1) = \mathbf{s}(t)$.

$$\mathbf{s} = \tilde{\mathbf{A}}\mathbf{s} \quad (9)$$

Thus equation (9), in conjunction with the relation given in equation (5), yields a set of $N+2$ equations from which $\mathbf{s}$ can be (theoretically) obtained for any given graph, if a sufficiently efficient solver algorithm exists. However, for directed star graphs solving these equations is straight-forward, and leads to an explicit formalism for $\mathbf{s}$ and the LeaderRank scores LR accordingly. Solving this set of equations involves that $\tilde{\mathbf{A}}$ can be explicitly written as

$$\tilde{\mathbf{A}} = \begin{pmatrix} 0 & 1/2 & 1/2 & ... & 1/2 & 1/N \\ 0 & 0 & 0 & ... & 0 & 1/N \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & ... & 0 & 1/N \\ 1 & 1/2 & 1/2 & ... & 1/2 & 1/N \end{pmatrix}. \quad (10)$$

for any given directed, extended star graph with vertices $V = \{v_c, v_2, ..., v_N, v_g\}$. One henceforth obtains the steady-state score vector $\mathbf{s} = (s_c, s_2, ..., s_N, s_g)^\intercal$ from the resulting set of equations which can be derived by simply re-arranging Equations (9) and (5):

$$s_c = \frac{N^2}{5N-1} + \frac{N}{5N-1} \quad (11)$$

$$s_i = \frac{2N}{5N-1}, \forall i = 2, ..., N \quad (12)$$

$$s_g = \frac{2N^2}{5N-1}. \quad (13)$$

This explicit formalism of $\tilde{\mathbf{A}}$ also highlights that the leaf vertices (denoted as v_i for textual cleanness in the following text) are indistinguishable with respect to the weighted adjacency matrix values $\tilde{A}_{i\cdot}$. Thus, the obtained LeaderRank scores S_i are identical as well. Plugging the computed values of $\mathbf{s}$ into the final update rule (see Equation (4)) yields the LeaderRank score for the central vertex v_c:

$$S_c = \frac{N^2}{5N-1} + \frac{3N}{5N-1} \quad (14)$$

$$(15)$$

Then the equal LeaderRank score S_i of the leaf nodes can be calculated according to Equation (8), from which the upper skewness bound ν_{max} is readily computed. Subsequently, for any irregular network graph the LeaderRank skewness can be calculated and normalized subsequently using a min-max normalization as denoted in (16), whereas ν_{min} can be assumed as 0 as discussed above.

$$\hat{\nu} = \frac{\nu - \nu_{min}}{\nu_{max} - \nu_{min}} = \frac{\nu}{\nu_{max}} \quad (16)$$

B. Detection of star topology

LeaderRank skewness $\hat{\nu}$ can be utilized to indicate adverse leader ranking by means of LeaderRank scores. In this section, we compare ν to the classic measure of network entropy (denoted as H in the following text). In order to allow direct comparison to $\hat{\nu}$ as well as to entropies computed from other graphs, H is required to be normalized analogously to $\hat{\nu}$. In this subsection, we give a brief overview on how normalization can be conducted.

Let A be the adjacency matrix of a network with N vertices, where each element $a_{ij} := 1$ if there exists a directed edge e_{ij} between adjacent vertices v_i and v_j. Each element of the principal diagonal a_{ii} is defined as $a_{ii} := \deg(v_i)$ and thus corresponds to the degree – the sum of the incoming and outgoing edges – of vertex v_i. The trace of A is defined as the sum of all elements of the principal diagonal: $tr(A) = \sum_{i=1}^{N} a_{ii}$. The formalism for graph entropy used by Passerini and Severini $H(\rho) = -tr(\rho \log_2 \rho)$ [25] is based on the von Neumann entropy and can be adapted as shown in

(17).

$$\begin{aligned} H(\rho) &= -tr(\rho \log_2 \rho) \\ &= -\sum_{i=1}^{N} \rho_i \log_2 \rho_i \\ &= -\sum_{i=1}^{N} \frac{a_{ii}}{tr(A)} \log_2 \frac{a_{ii}}{tr(A)} \\ &= -\sum_{i=1}^{N} \frac{\deg(v_i)}{\sum_{j=1}^{N} \deg(v_j)} \log_2 \frac{\deg(v_i)}{\sum_{j=1}^{N} \deg(v_j)}. \end{aligned} \tag{17}$$

This formalism, which is the entropy of the density matrix of a graph, describes the distribution of incoming and outgoing edges. In a randomly generated graph one expects $\deg(v_i) \approx \deg(v_j)$. In this case, the graph entropy H is close to the theoretical maximum entropy H_{max}. Therefore, the graph entropy only reaches its maximum if G is a regular graph where $\deg(v_i) = \deg(v_j) = D$. Because $\rho_i = D/DN = 1/N$ in a regular graph, one has H as shown in (18).

$$H = H_{max} = -\sum \rho_i \log_2 \rho_i = \log_2 N \tag{18}$$

In contrast, the minimum graph entropy H_{min} is observable in networks showing star topology. The trace $tr(A)$ of such a graph corresponds to $2N-2$ and the degree of its central vertex is $\deg(v_c) = N-1$. Consequently, the entropy of the central vertex H_c is calculated as shown in (19).

$$H_c = -\frac{N-1}{2N-2} \log_2 \frac{N-1}{2N-2} = -\frac{1}{2} \log_2 \frac{1}{2} = 0.5. \tag{19}$$

The degree of any other vertex is $\deg(v_i) = 1$. Hence, the entropy of a graph constituted as a star is calculated as follows:

$$\begin{aligned} H &= H_{min} \\ &= 0.5 + \sum_{V \setminus v_c} -\frac{1}{2N-2} \log_2 \frac{1}{2N-2} \\ &= 0.5 + \frac{1}{2} \log_2 (2N-2) \\ &= 1 + \frac{1}{2} \log_2 (N-1). \end{aligned} \tag{20}$$

The normalized network entropy can be finally computed according to (21):

$$\hat{H} = \frac{H - H_{min}}{H_{max} - H_{min}}, \hat{H} \in [0,1] \tag{21}$$

In order to illustrate expressiveness of $\hat{H}$ and $\hat{\nu}$ with respect to the underlying network topology, a straightforward experiment was carried out in which synthetic networks exhibiting star topologies were continuously mutated over time, resulting in almost regular graphs after numerous generations.

This simulated process consequently yields a continuous change of network topology for each graph. $\hat{H}$ and $\hat{\nu}$ were accordingly computed for every generation and tracked. The time series of both measures are shown in Figure 1. More precisely, simulations of topological change were conducted by starting with star graphs of fixed sizes ($N = 16, 32, 64, 128, 256$ and 512 vertices). In every generation, edges between every pair of vertices were randomly added and respectively removed.

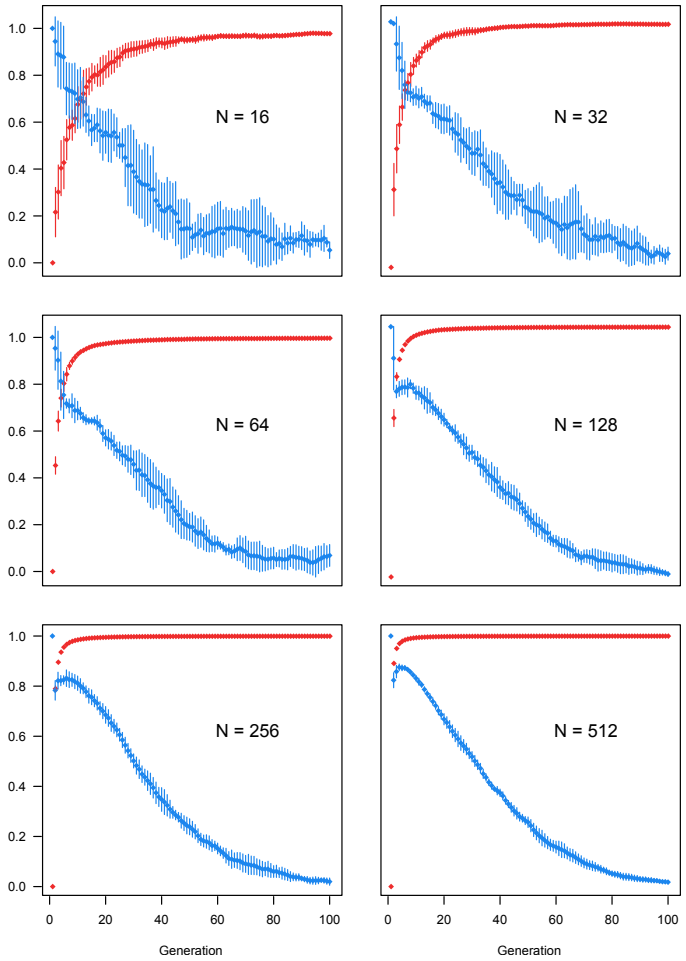

Figure 1. Simulation results of networks with various sizes N, whereas the red line represents $\hat{H}$, the blue line $\hat{\nu}$ and vertical bars indicate standard deviations.

For each graph size, six runs were conducted in an effort to estimate variance.

As shown in Figure 1, both measures converged after 100 generations. All entropy trajectories show fast convergence compared to $\hat{\nu}$ trajectories, with the convergence time decreasing with increasing N. Although $\hat{\nu}$ yield larger variances (especially for $N \leq 32$), its slower convergence and qualitatively similar trajectories for all graph sizes N illustrates greater sensitivity to topological changes. In that respect, matrix entropy loses significance with increasing graph size.

IV. Datasets

In this study, two different networks, namely Facebook and Twitter, of the German party "DIE LINKE" were analyzed, because both exhibit a star-like topology, yet to a different degree. As a comparison, a part of the Epinions social network, as an example for a nearly regular graph, was also included.

A. Facebook Dataset

Figure 2 depicts the network of the Facebook page "DIE LINKE" from January 2017 as a graph in which the size of each node corresponds to the out-degree (number of out-links). As can be seen, the network is dominated by the central node

of the page owner and, therefore, closely resembles a star-shaped topology.

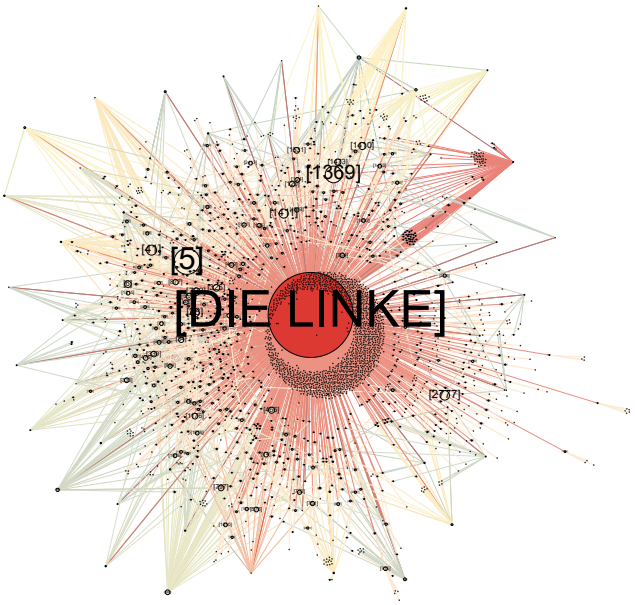

Figure 2. The network of the facebook page "DIE LINKE" of January 2017.

The central node often has the highest activity, meaning the most in- and out-links. The communication on the page was explored over a period of five months, from January 2017 up until May 2017, whereas all posts, comments and replies were taken into account as can be seen in Table I.

TABLE I. SUMMARY OF THE DATA INCLUDING NORMALIZED ENTROPY AND SKEWNESS OF THE CONSIDERED NETWORKS.

month	actors	posts	comments	replies	$\hat{H}$	$\hat{\nu}_{LR}$
January	2,878	26	2,955	3,471	0.19	0.98
February	2,146	33	2,196	2,062	0.24	0.98
March	3,196	40	3,501	3,245	0.17	0.97
April	2,432	26	2,558	3,295	0.22	0.98
May	4,765	31	4,130	5,674	0.10	0.98

Furthermore, it shows the normalized entropy and LeaderRank skewness of the "DIE LINKE" network, separately calculated for each month. It can be clearly seen that obtained $\hat{H}$ values fluctuate over time, whereas the LeaderRank skewness $\hat{\nu}_{LR}$ remains stable.

During the initial analysis of the dataset, it was observed that $12,031$ individuals were active throughout the five months. However, as shown in Figure 3, only 104 of these individuals were active in every single month. In general, it can be stated that users showed rather sparse and sporadic activity, with only a minority being recurrent users. Thus, yet again, this supports the assumption this network has a star-like topology. Additionally, this may indicate that the activity of users and, subsequently, the degree of opinion leadership, depends on the topics being discussed in a certain time period. However, in order to support this claim, further analyses need to be undertaken, which will be covered in a future study.

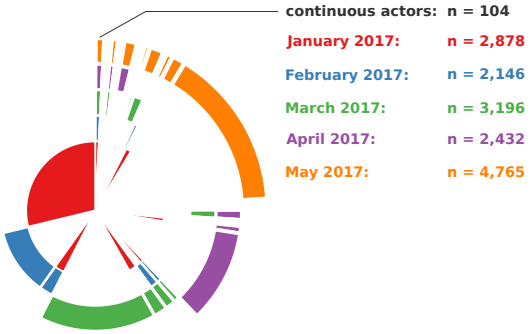

Figure 3. Sunburst chart of actor activity in the Facebook network consisting of one radial segment for each user, whereas a user's segment in a time layer is left out if said user was observed to be inactive in that time period.

B. Twitter Dataset

In a subsequent analysis the Twitter network of "DIE LINKE" was evaluated.

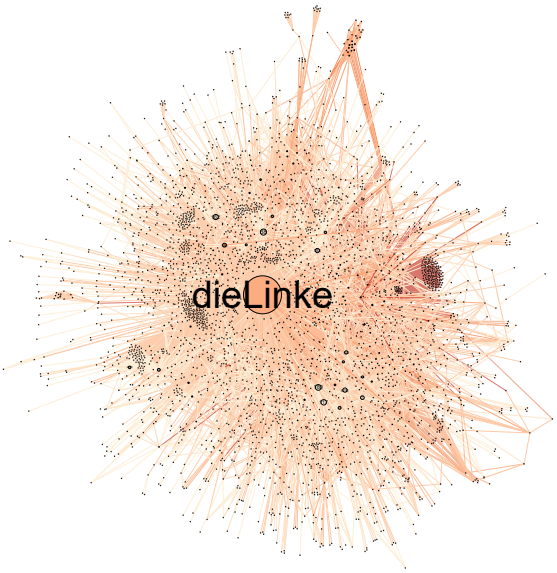

Figure 4. Twitter network of "DIE LINKE" of January 2018.

As can be seen in Figure 4 the star topology is less predominant for this network in comparison to the Facebook dataset. Consequently, a star topology is recognizable but at the same time some accounts besides "DIE LINKE" emerge.

The Twitter data set consists of tweets authored by "DIE LINKE", tweets addressing "DIE LINKE" and replies to the respective tweets. Aforementioned data was collected for the entire year of 2018 and on average twice as many actors where involved in the network compared to the Facebook data.

With an $\bar{\hat{\nu}} = 0.73$ the statistical analysis of the data shows that even though the star-like topology is not as distinctive as for the Facebook network it is still relatively strong as could already be seen in Figure 4. Furthermore, in comparison to the Facebook network the values for the skewness in the Twitter network show a greater fluctuation or to be precise cover a

TABLE II. SUMMARY OF THE TWITTER DATA INCLUDING NORMALIZED ENTROPY AND SKEWNESS.

month	actors	tweets	$\hat{H}$	$\hat{\nu}_{LR}$
January	5,966	10,695	0.39	0.74
February	6,194	11,466	0.40	0.79
March	7,677	14,820	0.44	0.86
April	7,179	12,711	0.38	0.84
May	7,529	14,349	0.36	0.77
June	8,864	21,407	0.14	0.86
July	6,612	13,951	0.22	0.67
August	6,834	13,033	0.24	0.79
September	8,072	16,631	0.33	0.79
October	6,943	13,974	0.26	0.87
November	5,757	10,249	0.32	0.76
December	5,642	9,119	0.38	0.75

greater range ($R^{FB}_{\hat{\nu}} = 0.1, R^{T}_{\hat{\nu}} = 0.2$). However, they are still more stable than the corresponding values for the entropy ($R_{\hat{H}} = 0.3$).

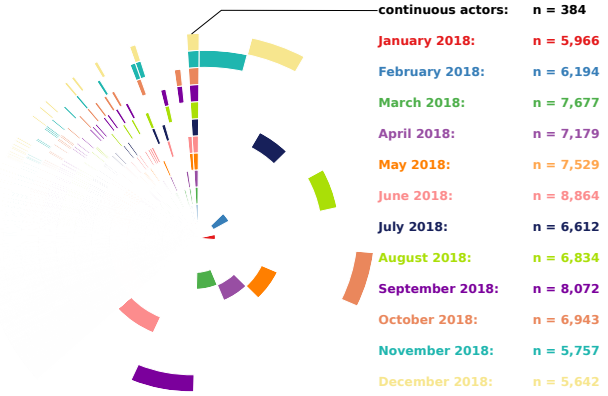

Figure 5. Sunburst chart of actor activity in the twitter network consisting of one radial segment for each user, whereas a user's segment in a time layer is left out if said user was observed to be inactive in that time period.

As can be seen in Figure 5, similar to the Facebook network, only a small amount of users is active throughout the entire year, yet rather their activity is concentrated on certain months.

C. Epinions Dataset

Figure 6 shows part of the *Epinions* social network [26] which, in contrast to the previous datasets, tends to be regular. Subsequently, there is no node, which dominates all others in terms of its degree. In this figure, due to the size of the network, it was necessary to arbitrarily limit the depiction by applying k-core ≥ 80 [27] showing only the most active nodes.

In comparison to the other networks, the *Epinions* social network [26] consisting of 75,879 actors shows a normalized network entropy $\hat{H} = 0.65$ and a normalized leader rank skewness $\hat{\nu}_{LR} = 0.07$, indicating a considerably less skewed LeaderRank score distribution.

The three discussed real world examples support the results of the simulation experiment discussed in Section III, whereas the normalized network entropy is less expressive in regards to an evaluation of the network topology than the LeaderRank skewness.

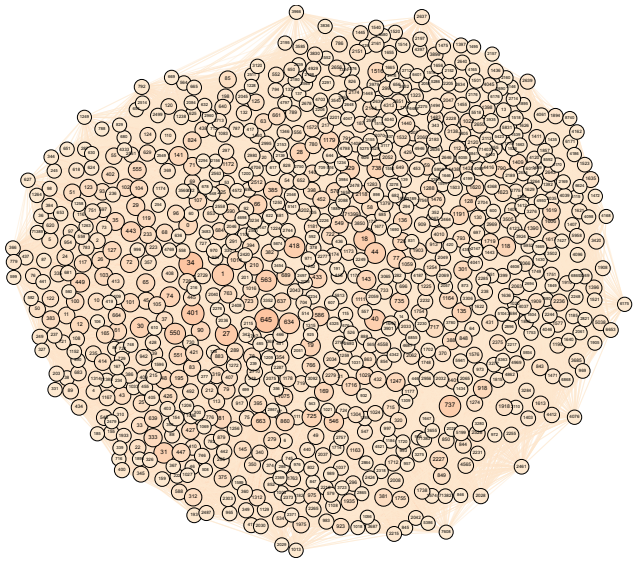

Figure 6. Part of the *Epinions* social network [26] (filtered by k-core ≥ 80).

V. COMPETENCE BASED RANKING APPROACHES

To address the issues discussed in Section III, we present a modification of the original LeaderRank referred to as CompetenceRank as well as some additional heuristics as improvements to incorporate specific features found in social networks.

A. CompetenceRank

In order to counteract the skewness of the LeaderRank in graphs with a star-like topology, the LeaderRank score of actors with a high degree of interaction, who at the same time only receive minimal attention by others, needs to be penalized. Similar to a citation network the relevance of a vertex does not only depend on the number of its interactions, yet it rather depends on a balanced ratio of own interactions and references by others. If this ratio is used as a weighting of the LeaderRank only those actors remain in the top ranks whose influence is based mainly on their competence.

Therefore, let V be the set of all vertices representing the actors of a social network and E be the set of all directed edges representing the relationship between vertices for example the communication or followers. The CompetenceRank CR_i of a particular actor $v_i \in V$ lowers the LeaderRank score S_i depending on the ratio of out-going and in-coming edges.

$$CR_i = \frac{S_i}{1 + \frac{e^{out}_{v_i}}{|E|} \sum_{v \in V} S_v} \tag{22}$$

The CompetenceRank as shown in (22) is subsequently calculated by dividing the original LeaderRank score S_i by a fraction of the cumulative sum of LeaderRank scores defined by the vertex's share of network activity, with $e^{out}_{v_i}$ being the number of its outgoing edges. By definition, the sum of LeaderRank scores of all vertices in the social network graph is equal to the number of actors N. When considering regular graphs, one observes LeaderRank distribution skewness $\hat{\nu} = 0$ as well as $e^{out}_{v_i} = e^{out}_{v_j} = D$ for any pair of randomly chosen vertices v_i and v_j. Thus, $|E| = ND$. From this, (22) can be

rewritten as

$$CR_i = \frac{S_i}{1 + \frac{D}{ND}N} = \frac{1}{2}S_i. \quad (23)$$

We finally define the CompetenceRank based on the assumption that $S_i = CR_i$ in regular graphs which is thus simply achieved by multiplying the expression in (23) by 2 as given in (24).

$$CR_i = \frac{2S_i}{1 + \frac{e_{v_i}^{out}}{|E|}N} \quad (24)$$

As shown in (25) one can calculate the discrepancy between the CompetenceRank and the LeaderRank in terms of the root-mean-square deviation $RMSD$ (note: vertical line denotes average sum). In turn that value can be seen as a further function of network regularity besides the measures discussed in Section IV.

$$RMSD = \sqrt{\sum_{i=1}^{N} [CR_i - S_i]^2} \quad (25)$$

On average one receives an $RMSD$ of 11.3 for the Facebook network, of 5.7 for the Twitter network and of 4.9 for the Epinions network.

B. Improved CompetenceRank

Especially social networks include many additional features that support the idea of a competence based ranking. In particular *likes* and *shares* play a special role in social media and reflect the acceptance of an expressed opinion and, as a consequence, should be considered when assessing the competence. Neither the LeaderRank nor the CompetenceRank as reported in [1] take these features into consideration or are even designed to include additional features. In the following paragraphs heuristics of the most important features of social network are designed and step by step combined in a weighted manner in order to reflect the relevance of different features regarding the competence in various types of social media platforms.

1) Pivoted post frequency normalization: As already discussed, if an actor posts messages with a high frequency without receiving much response from the network, their activity becomes less valuable and their LeaderRank score needs to be lowered. This means, when looking at it from another point of view, the fewer messages an actor posts, while at the same time receiving great response from the rest of the network, the more valuable this actor becomes. Consequently, their score needs to receive a higher rank. How much the rank needs to be lowered or raised has to depend on how much the individual's posting frequency deviates from the average posting frequency of all actors in the network. A similar behavior was described by Singhal et al. in 1995/1996 [28] with the pivoted length normalization for the text retrieval problem.

$$normalizer = 1 - b + b\frac{PF_i}{\sum_{i=1}^{N} PF_i} \quad (26)$$

Its original core idea is to reward or penalize a document based on the document length in relation to the average document length within a given collection of documents. For social networks this easily adapts to rewarding or penalizing actors when their total activity is either above or below the average activity in a given network. This leads to the equation as shown in (26), whereas the total activity is measured by the post frequency PF_i of the individual actor v_i and b controls how much an actor's activity is rewarded or punished. Depending on the network, the extent to which the activity is rewarded or penalized differs. In general, achieving a high degree in opinion leadership within a network requires individuals to understand and conform to its code of conduct. For example, when comparing a network of scientific publications and citations to a twitter network, then the former is defined by a rather low publication or post frequency but with a high quality whereas the latter favors a high activity but limits the depth and quality with a length limitation on each tweet. Moving from twitter to the scientific domain and vice versa inevitably requires an adaption to the new circumstances and only if this transition in behavior is achieved will one be able to maximize their influence in the respective area. In summary, the pivoted post frequency normalization rewards individuals that maintain a post frequency in line with or higher than average.

2) Sublinear post frequency transformation: Especially in networks that tend to have a star-like topology, few very active actors dominate the entire network. In the field of information retrieval a similar problem is addressed with a sublinear term frequency transformation, whereas one of the most popular approaches is Robertson's BM25 [29]. Here, the gain is lowered with an increasing term-frequency, while, at the same time, an upper bound of the term frequencies is defined. When adapted to the problem of highly active actors in social networks the impact of increasing posting frequencies can be lowered and with $k + 1$ an upper bound can be defined as shown in (27).

$$gain = \frac{(k+1)PF_i}{k + PF_i} \quad (27)$$

As previously discussed, the degree of opinion leadership partially relies on respecting the circumstances. While the pivoted post frequency normalization ensures that low activities are being penalized it also offers the chance of a disproportionate reward for users that are drastically more active than average. Therefore, the sublinear post frequency transformation diminishes returns that result from high activity and introduces an upper limit that prevents actors from extensively receiving a disproportionate gain. This concept allows users to benefit from being slightly more active than average while at the same time approaching the upper boundary requires a significant increase in activity.

3) Post frequency normalized LeaderRank: Using a combination of the pivoted post frequency normalization and the sublinear post frequency transformation as a weight for the LeaderRank score leads to a post frequency normalized LeaderRank nS_i as shown in (28).

$$nS_i = S_i \left[1 - b_1 + b_1 \frac{\frac{(k_1+1)PF_i}{k_1+PF_i}}{\sum_{i=1}^{N} PF_i}\right] \quad (28)$$

Using this equation the original LeaderRank is weighted by a fraction of an actor's activity in the entire activity of all network actors, whereas with k_1 the dominance of extreme activity over all other activities is minimized. Furthermore, with b_1 it is possible to control how much the degree of activity above or below the average is punished or rewarded,

respectively. The normalized LeaderRank as shown in (28) has a similar effect as the CompetenceRank in (24). However, with the parameters it is possible to adjust the normalized LeaderRank to the conditions of a specific network. For example, for a platform that focuses on posts with a high quality one may chose a low value for b_1, because of the low importance of the post frequency. Contrarily, for a platform like Twitter, which focuses more on activity, a higher a value can be chosen.

4) Incorporating likes and shares: likes are a key aspect of social media platforms as they show the acceptance of an actor by other actors and are, thus, an expression of competence. Therefore, they need to be taken into account when evaluating the impact of any given individual on such a platform and the post frequency normalized LeaderRank is combined with the average number of *likes* $\frac{LF_i}{PF_i}$ a user receives per post, whereas LF_i denotes the like frequency of a certain actor. Averages are used since a user could, for example, have an especially high number of 300 posts and only acquire one *like* per posted message. Contrary, a user posting three times could be receiving 100 *likes* per message. The total *like* count might be similar, yet the impact of the former appears to be marginal while the content of the latter seems to be well received and quite influential. In contrast to other activities in a network, such as creating posts, normally *likes* are connected to a post or message and not a certain actor. This means that everyone who can read the post can *like* it, even though they might not be part of the observed network and it is impossible to ensure that only those *likes* are considered that are from actors also active in this network. For example, in the "DIE LINKE" Twitter network, a tweet from a member of a right-winged party could appear in the network if it is directed to "DIE LINKE". This tweet might receive a lot of attention from other actors, active in the right-winged party network, yet not much attention from actors of the twitter network "DIE LINKE". Nonetheless, the tweeting actor would receive a high number of average *likes* for the "DIE LINKE" Twitter network. A similar effect could be achieved if *likes* are received through bots or are bought. Therefore, the normalized like score nLS_i of an actor v_i is calculated as the average number of *likes* this actor's posts receive weighted with the fraction of the actor's activity in the overall activity of the network as shown in (28).

$$nLS_i = \frac{LF_i nS_i}{PFv_i \sum_{i'=1}^{N} nS_{i'}} \tag{29}$$

Another important aspect of social networks is the number of posts by an actor that have been shared by other actors. In comparison to the number of *likes* a post receives, a highly shared post/tweet extends its reach significantly, consequently allowing the individual to influence more actors than they normally could. Similarly to (28) concepts like pivoted shares frequency normlization and sublinear shares frequency transformation are utilized together with their parameters k_2 and b_2 to maintain a controlled environment without too heavily benefiting extreme cases, resulting in a normalized share score nSS_i as shown in (30), whereas $\overline{SF}_i$ denotes the average share frequency an actor v_i receives.

$$nSS_i = 1 - b_2 + b_2 \frac{\frac{\left[1+k_2 \sum_{i'=1}^{N} \overline{SF}_{i'}\right]\overline{SF}_i}{k_2 \sum_{i'=1}^{N} \overline{SF}_{i'} + \overline{SF}_i}}{\sum_{i'=1}^{N} \overline{SF}_{i'}} \tag{30}$$

Finally, all components are combined resulting in the improved CompetenceRank CR_i as shown in (31) with α being the parameter that weights the normalized like score depending on the importance of *likes* in the observed network.

$$CR_i = [nS_i + \alpha nLS_i]\, nSS_i \tag{31}$$

VI. Results

Since the required additional data, i.e., *likes* and *shares*, were not available for the Facebook dataset, only the Twitter network of "DIE LINKE" for the year 2018 was analyzed with the new improved CompetenceRank and compared with the results of the LeaderRank. In the analysis, the parameter b was set to 0.7, k_1 was defined as the average tweet frequency in the entire network, k_2 as double the amount of the average tweet frequency and α was set to two assuming that liking is twice as important for competence as activity in the considered network.

An overview of the five highest opinion leader scores, indicating the discrepancy between the results for the LeaderRank and the improved CompetenceRank, is shown in Figure 7. As can be seen, the five accounts with the top scores are for the LeaderRank less diverse over the entire year as compared to the improved CompetenceRank. Lacking diversity in itself is not necessarily negative, however, when looking at the results for the LeaderRank it can be noticed that the accounts in Figure 7 include several political parties. Over the duration of 12 months, excluding the account of "DIE LINKE" (the owner of the network), with the LeaderRank it was possible to identify 19 accounts of possible opinion leaders, of which 9 belong to political parties (e.g. "afd", "cdu", "fdp", "linke_sh"). In comparison, a total of 23 accounts were identified using the improved CompetenceRank of which only 5 belonged to political parties.

It is not surprising that political parties appear in the top ranks, as they are a quintessential part of political discourse and thus it is their aim to shape the political opinion of the citizens. However, political parties reflect the consensual opinion of their members. Nevertheless, the ideas shaping the opinion of others and thus the political discourse as such often come from individuals. These opinions and ideas are not necessarily conform with the congruent opinion of the party. Still, they inspire the discussion and have the potential to influence the consensus. When only considering the activity of an account, as does the LeaderRank, such accounts, cannot compete with the accounts of political parties that are used to inform the public about the activity of the party and are thus highly active within a network. The improved CompetenceRank is able to raise the ranking of these accounts and to lower the ranking of those accounts that only receive a high rank because of their activity.

Deeper insight was provided by a thorough analysis of the monthly datasets. The LeaderRank and the improved CompetenceRank were calculated, providing us a total of two different ranked lists. Subsequently, to minimize the potential of performing well by chance on the first five accounts, the analysis of a list of five accounts per month was extended to the 20 highest ranking accounts per month. Ranked lists need to be evaluated in a way that reflects increased or decreased performance. Hence, the identified accounts were divided into six different categories: Individuals, Journalists, News, Political Parties, Politicians, Other and Unknown.

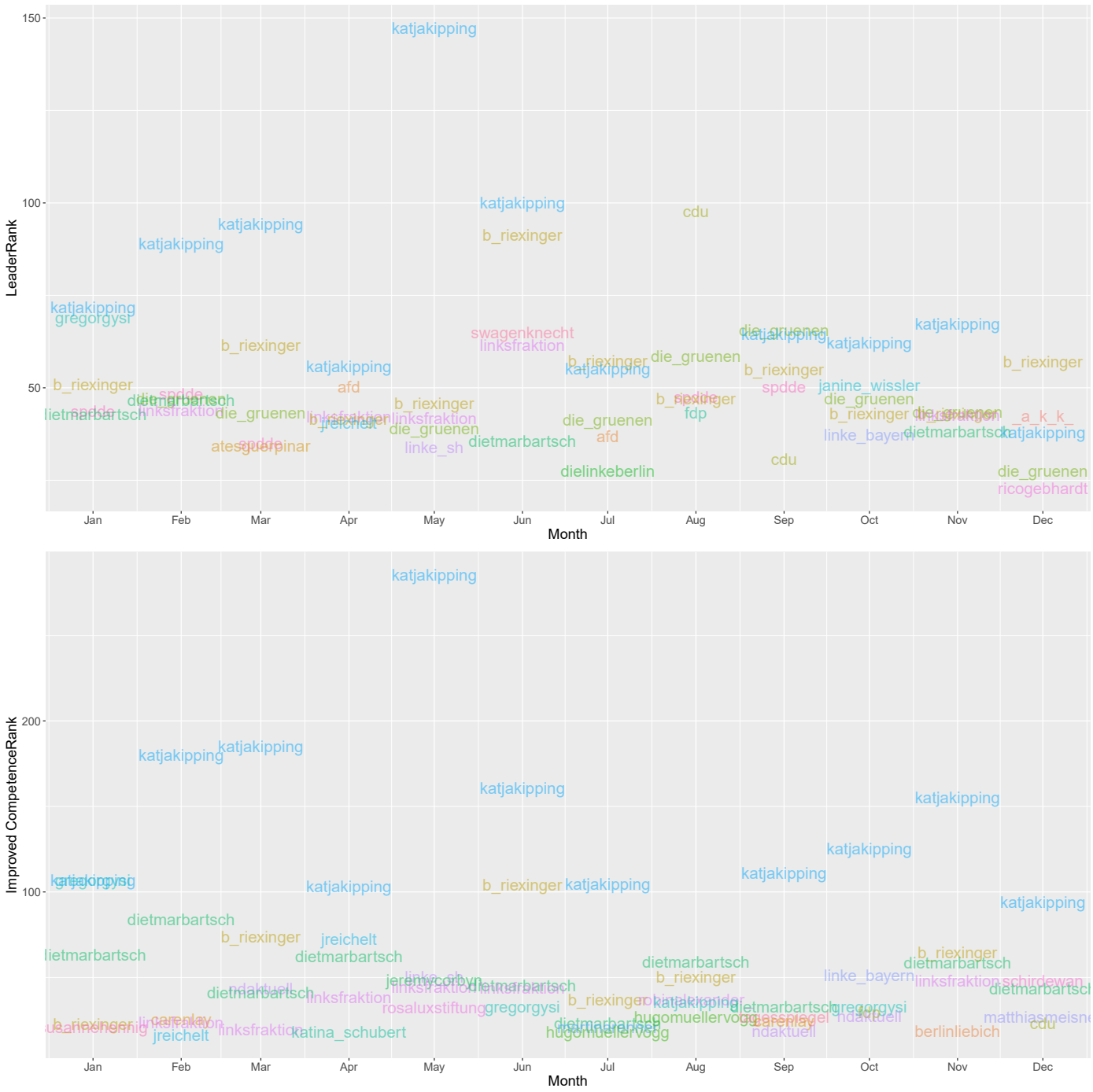

Figure 7. Comparison of the fivea accounts with the highest LeaderRank (upper graph) and improved CompetenceRank (lower graph) scores for the year 2018 for the Twitter Network of “DIE LINKE” .

Each of the 20 accounts per month received a label that was derived through manual evaluation of their Twitter profiles. Ordinary Twitter accounts, seemingly run by individuals without an obvious political office or a position in journalism were labeled as “Individual”. Following this procedure individuals with an obvious background in the field of the news industry were labeled as “Journalists”, whereas accounts tweeting on behalf of a news organization, accordingly do not represent the opinion of a single individual hence giving them the label “News”. In the same manner “Politician” refers to a single individual being either active in a political party or involved in a political office. Analogously to “News”, “Political Party” refers to an account tweeting on behalf of a political party. “Other” includes everyone not fitting into previously mentioned categories (e.g. companies, bands, NGOs etc.) and finally “Unknown” includes suspended and deleted accounts. The total results are displayed in Table III.

In the given network, it becomes evident that political

TABLE III. SUMMARY OF TYPE OCCURRENCES FOR MONTHLY DATA SUMMED UP ACCORDING TO THE USED MEASURE

Type	LR	Improved CR
Individual	10	25
Journalist	2	26
News	11	30
Politician	99	115
Political Party	104	36
Other	11	4
Unknown	3	4

parties are down-ranked according to their influence by the improved CompetenceRank, whereas individuals, journalists and news outlets receive higher ranks. This result confirms the assumption that the improved CompetenceRank counteracts the skewness of star-shaped topologies, as can be found for example in political networks, and further allows to distinguish the real initiators that trigger the intraparty pattern of opinions from the mass of other unimportant accounts in the network.

Furthermore, an account identified as an opinion leader should be associated with a small group or even a single person. This can be brought back to Katz' original thesis that large parts of society are not influenced by mass media or in our specific case by organizations and political parties but rather by trustworthy, influential opinion leaders. In turn, identifying 36 instead of 104 political parties is a considerable improvement, because it allows to identify more individuals, more journalists and more politicians. Moving away from pointing out the general importance of political parties and instead selecting specific individual accounts exerting their influence over a given social network is of tremendous value.

In this experiment the improved CompetenceRank outperforms the LeaderRank as it returned fewer political parties, fewer accounts of category "Unknown" and fewer suspended or deleted accounts. The analyzed Twitter network is less skewed than the Facebook network, as shown in Section IV. Therefore, it can be assumed that the improvements become even more distinctive when analyzing a highly skewed network.

VII. CONCLUSION AND FUTURE WORK

The analysis of social networks, and in particular identifying influential and opinion-influencing profiles, is of great interest in forensic research for a variety of reasons. In the present study, it was shown that the usual centrality-based approaches, and in particular the LeaderRank, produce erroneous results in star-like networks, such as Facebook pages of political parties. Furthermore, LeaderRank skewness was presented as an appropriate measure to quantify the degree of distortion of a network or in other words its proximity to a star-shaped topology.

Subsequently, CompetenceRank was introduced as a measure to overcome the shortcomings of the popular LeaderRank in star-like network topologies.

Additionally, an improvement of the CompetenceRank was provided incorporating fundamental interaction data such as "likes" and "shares". This methodology was tested on the Twitter network of "DIE LINKE". Identifying political parties as dominant and influential accounts on social media does not yield significant new insight into a political network since the importance of such accounts can be derived prior to any analysis as political discussions are frequently centered around political parties. However, pointing out influential individual politicians or individuals in general aligns more with the goal and image one has in mind when talking about an opinion leader. It was shown that the new measure outperforms the LeaderRank by identifying considerably more individual Twitter accounts and attributing less importance to accounts run by political organizations.

In following studies, it would be interesting to analyze the observed phenomena in more fine-grained time ranges. Additionally, it is necessary to take more and different network topologies into account. Furthermore, it was noticed that the texts in the Facebook data used were surprisingly well written. This provides an opportunity to conduct further textual analyses especially to answer the question whether there is a correlation between topics and opinion leaders and if so, how both develop over time.

REFERENCES

[1] M. Spranger, F. Heinke, H. Siewerts, J. Hampl, and D. Labudde, "Opinion Leaders in Star-Like Social Networks: A Simple Case?" in The Eighth International Conference on Advances in Information Mining and Management (IMMM). Barcelona, Spain: IARIA, July 2018, pp. 33–38.

[2] E. Katz, "The two-step flow of communication: An up-to-date report on an hypothesis," Public Opinion Quarterly, vol. 21, no. 1, Anniversary Issue Devoted to Twenty Years of Public Opinion Research, 1957, p. 61.

[3] M. Spranger, F. Heinke, S. Grunert, and D. Labudde, "Towards predictive policing: Knowledge-based monitoring of social networks," in The Fifth International Conference on Advances in Information Mining and Management (IMMM 2015), 2015, pp. 39 – 40.

[4] M. Spranger, H. Siewerts, J. Hampl, F. Heinke, and D. Labudde, "SoNA: A Knowledge-based Social Network Analysis Framework for Predictive Policing," International Journal On Advances in Intelligent Systems, vol. 10, no. 3 & 4, 2017, pp. 147 – 156.

[5] M. Spranger, S. Becker, F. Heinke, H. Siewerts, and D. Labudde, "The infiltration game: Artificial immune system for the exploitation of crime relevant information in social networks," in Proc. Seventh International Conference on Advances in Information Management and Mining (IMMM), IARIA. ThinkMind Library, 2017, pp. 24–27.

[6] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Comput. Netw. ISDN Syst., vol. 30, no. 1-7, Apr. 1998, pp. 107–117.

[7] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," PloS one, vol. 6, no. 6, 2011, pp. 1–9.

[8] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted leaderrank," Physica A: Statistical Mechanics and its Applications, vol. 404, no. Supplement C, 2014, pp. 47 – 55.

[9] E. M. Rogers, Diffusion of innovations. New York: The Free Press, 1962.

[10] B. Huang, G. Yu, and H. R. Karimi, "The finding and dynamic detection of opinion leaders in social network," Mathematical Problems in Engineering, vol. 2014, 2014, pp. 1–7.

[11] P. Parau, C. Lemnaru, M. Dinsoreanu, and R. Potolea, "Opinion leader detection." in Sentiment analysis in social networks, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds., 2016, pp. 157–170.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, Mar. 2003, pp. 993–1022.

[13] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07, M. J. Silva, A. O. Falcão, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, and Ø. H. Olsen, Eds. New York, New York, USA: ACM Press, 2007, pp. 971 – 974.

[14] A. Aleahmad, P. Karisani, M. Rahgozar, and F. Oroumchian, "Olfinder: Finding opinion leaders in online social networks," Journal of Information Science, vol. 42, 09 2015.

[15] Y. Chen, L. Hui, C. I. Wu, H. Liu, and S. Chen, "Opinion leaders discovery in dynamic social network," in 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), 2017, pp. 1–6.

[16] L. Jain and R. Katarya, "Discover opinion leader in online social network using firefly algorithm," Expert Systems with Applications, vol. 122, 2019, pp. 1 – 15. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741830811X

[17] B. Amor et al., "Community detection and role identification in directed networks: Understanding the twitter network of the care.data debate," CoRR, vol. abs/1508.03165, 2015.

[18] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Za&iuml and O. R. ane, Eds. New York, NY: ACM, 2002, p. 61.

[19] W. Oueslati, S. Arrami, Z. Dhouioui, and M. Massaabi, "Opinion leaders' detection in dynamic social networks," Concurrency and Computation: Practice and Experience. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5692

[20] C. Egger, "Identifying key opinion leaders in social networks: An approach to use instagram data to rate and identify key opinion leader for a specific business field," Master Thesis, TH Köln - University of Applied Sciences, Köln, 2016.

[21] M. Z. Shafiq, M. U. Ilyas, A. X. Liu, and H. Radha, "Identifying leaders and followers in online social networks," IEEE Journal on Selected Areas in Communications, vol. 31, no. 9, 2013, pp. 618–628.

[22] Z. H. Zhang, G. P. Jiang, Y. R. Song, L. L. Xia, and Q. Chen, "An improved weighted leaderrank algorithm for identifying influential spreaders in complex networks," in 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 1, July 2017, pp. 748–751.

[23] R. Nadeau, E. Cloutier, and J.-H. Guay, "New evidence about the existence of a bandwagon effect in the opinion formation process," International Political Science Review, vol. 14, no. 2, 1993, pp. 203–213.

[24] R. M. Raafat, N. Chater, and C. Frith, "Herding in humans," Trends in Cognitive Sciences, vol. 13, no. 10, 2009, pp. 420 – 428.

[25] F. Passerini and S. Severini, "Quantifying complexity in networks: The von neumann entropy," Int. J. Agent Technol. Syst., vol. 1, no. 4, Oct. 2009, pp. 58–67.

[26] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic web," in The Semantic Web - ISWC 2003, D. Fensel, K. Sycara, and J. Mylopoulos, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 351–368.

[27] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC bioinformatics, vol. 4, January 2003, p. 2.

[28] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '96. New York, NY, USA: Association for Computing Machinery, 1996, pp. 21–29.

[29] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '94. Berlin, Heidelberg: Springer-Verlag, 1994, pp. 232–241.