Revealing the Political Affinity of Online Entities Through Their Twitter Followers[☆]

Giorgos Stamatelatos*, Sotirios Gyftopoulos, George Drosatos, Pavlos S. Efraimidis

Dept. of Electrical and Computer Engineering, Democritus University of Thrace, Kimmeria, Xanthi 67100, Greece

Abstract

In this work, we show that the structural features of the Twitter online social network can divulge valuable information about the political affinity of the participating nodes. More precisely, we show that Twitter followers can be used to predict the political affinity of prominent Nodes of Interest (NOIs) they opt to follow. We utilize a series of purely structure-based algorithmic approaches, such as modularity clustering, the minimum linear arrangement (MinLA) problem and the DeGroot opinion update model in order to reveal diverse aspects of the NOIs' political profile. Our methods are applied to a dataset containing the Twitter accounts of the members of the Greek Parliament as well as an enriched dataset that additionally contains popular news sources. The results confirm the viability of our approach and provide evidence that the political affinity of NOIs can be determined with high accuracy via the Twitter follower network. Moreover, the outcome of an independently performed expert study about the offline political scene confirms the effectiveness of our methods.

Keywords: Social Network Analysis, Twitter Followers, News Media, Political Affinity

[☆]A preliminary version of this work appeared as "Deriving the Political Affinity of Twitter Users from Their Followers" in the 11th IEEE International Conference on Social computing and networking (SocialCom 2018).

*Corresponding author

Email addresses: gstamat@ee.duth.gr (Giorgos Stamatelatos), sgyftopo@ee.duth.gr (Sotirios Gyftopoulos), gdrosato@ee.duth.gr (George Drosatos), pefraimi@ee.duth.gr (Pavlos S. Efraimidis)

1. Introduction

Twitter, an online news and social networking service, has been subject of scientific research for at least a decade. Users in Twitter can follow other users in order to receive short messages posted by them, which are called tweets. The follower relationships of Twitter naturally convey an inherent directed graph structure, where vertices are the user accounts and edges represent the follower-to-followee relationship. The interpretation of these links varies across contexts: they may represent intimate relationships, common interests, an intent in news briefing and many others.

The significance of Twitter in research is partially because it supplies means to comprehend social relationships and influence dynamics in human societies. Various studies examine the structural and topological characteristics of the Twitter network, for example via concepts related to user influence and centrality [1] while others focus on extracting information from the content of tweets [2]. Furthermore, the Twitter network has been previously used for a multitude of practical applications, for example stock market predictions [3], event detection [4] and geo-locating users [5].

A distinct characteristic of Twitter is the presence of politically related actors, for example politicians or other party representatives, public officials, candidates as well as news media. These actors engage on the social platform as part of their political campaigns or utilize it as a means of political deliberation and advocacy [6]. The topic of political deliberation in social networks is relevant and has received substantial attention, especially through its applications to the identification of political bias in news sources.

In this work, we study the topic of deriving the political affinity of particular nodes of interest (NOIs) by using the structural features of the Twitter network. More specifically, we consider the NOIs to be the members of the Greek Parliament (MPs) and the most popular news sources, although the set of NOIs can be enriched with other politically engaged actors as well. Our approach fo-

cuses on two primary objectives. The first objective is to confirm that Twitter links between the NOIs and their followers can be used to identify the political affinity of the MPs and establish suitable methods to accomplish this. Our findings indicate that the Twitter follower network can portray with very high precision the affinity of political actors. Our second objective is to extend the application of this methodology to determine the political affinity of the most popular news sources, a natural extension given the strong relationships among political actors and news media. We argue that the results are promising and in agreement with the actual political scene, although not as consistent as the findings of the first objective.

Since, however, there is no single interpretation of political affinity, we determine three different perspectives and establish analytical methods that comply with each one. The *group affiliation* refers to the identification of groups or clusters of NOIs with the same or similar affinity. The *bipolar arrangement* projects the NOIs in a one dimensional arrangement in respect to a relevant measure, for example the left-to-right political axis. Finally, the *influence factors* constitutes a way to quantify the affinity of each NOI relative to another entity, for example the political parties.

The methods we propose in this work rely only on the social ties formed among relevant parties in the network and don't require any prior knowledge regarding the political standing of the involved entities. Our approach utilizes the nodes of the implicit graph structure simply as their Twitter IDs, and no additional knowledge about these user accounts is required. Furthermore, our methods are easily reproducible and can be implemented without complex filtering or preprocessing. They attain very high accuracy, even on a complex political scene with a large number of political parties. An important feature of our methodology is that it leverages the knowledge of the network to determine the political standing of a NOI. This is in direct contrast with utilizing the NOI's own explicit profile, for example tweets or friends, which portrays what the NOI is trying to convey to the network, rather than the opinion of the network about them. An additional effect is that users of importance cannot easily handpick

their followers, who adapt to the online and offline political scene.

Studying the topic of deriving the political affinity of news sources is appealing because it constitutes a primitive technique of inducing higher level knowledge from public information. In particular, results about the political affinity of news sources can potentially characterize the news media scene of a country as a whole, for example if it is biased or it favors only part of the political spectrum. Moreover, another application might be the identification of political bias in particular news articles or even the classification of fake news.

Overall, our approach relies on the assumption that people's political preferences will, on average, reflect those of the politicians or the news sources they follow, a phenomenon described as *selective exposure*. Prior literature on this topic suggests this assumption is reasonable since people seek after information from those with similar political views [7]. In the context of our study, the interpretation of selective exposure dictates that the following decisions of Twitter users provide information about their own perceptions of both their ideological position and that of the political accounts they follow [8]. Previous research demonstrates that the assumptions that our approach is relied upon are well founded in a news media framework as well. For example, in [9] it is stated that readers have an economically significant preference for like-minded news, which is consistent with our assertions.

Finally, in this paper, we make use of the *overlap coefficient*, a measure of similarity between two sets, in particular the follower sets of the NOIs. This measure appears for example in [10] for the purpose of studying affiliation networks and in [11] for text mining applications. To the best of our knowledge, the overlap coefficient has not seen extensive use until now in the context of social network analysis.

The highlights of our contribution are summarized as follows:

- Further proof of the selective exposure phenomenon, targeted for the Twitter network, as well as additional evidence that followers can portray the political leaning of their followees.

– The analytical formulation of three distinct perspectives of political affinity (the group affiliation, the bipolar arrangement and the influence factors) and the suggestion of techniques suitable for each perspective.

– A structural dataset acquired via the Twitter API comprising the nodes of important political influence in Greece along with their follower sets.

– The application of novel techniques, specifically the Minimum Linear Arrangement problem, which is not mentioned in the Social Network Analysis literature.

– The promotion of the overlap coefficient as a measure of pairwise similarity.

The article is organized as follows. In Section 2 we explore the recent literature in respect to political concepts in social networks. The dataset used in this work as well as the methodology are described in Section 3. In Section 4 we demonstrate the existence of rich political information within the Twitter follower dataset by evaluating the effectiveness of our methods and, moreover, lay out the experimentation settings. In Section 5, our methodology is applied on the combined MP and news sources dataset to assess the political affiliation and orientation of the news sources. The results are evaluated against the replies of an expert survey that was conducted for this purpose. Finally, Section 6 concludes this paper and presents suggestions for future work.

2. Related Work

Previous research demonstrates that the concept of social network analysis in Twitter and other online social networks is a very active field. In [12] the authors build interest profiles of social network users based on the homophily principle; users tend to interact with users with common interests or preferences. The review in [1] summarizes methods of quantifying the influence and popularity of users, targeted at the Twitter network, while in [13] the similarity among users in social communities is detected based on the similarities of their spatial

history profiles. In this study we focus exclusively on the political aspect of social interactions while our objective is to identify the political interests of influential users based on their followers.

A number of previous studies have promoted concepts related to the detection and analysis of political affinity. In [14], the authors show that Twitter is used extensively for political deliberation and evaluate whether tweets reflect the current offline political sentiment. In [15], the values of user attributes such as political orientation or ethnicity are inferred, while in [16] an example application to determine political leanings from tweets is demonstrated. These methods operate by examining the content of tweets while the approach presented in this paper utilizes algorithms that only rely on the topological and structural characteristics of the Twitter network.

Furthermore, in [17], the authors construct the politician-journalist graph and attain multiple conclusions regarding the network structure. Moreover, the study of [18] is the identification of the characteristics of political parties and the political leaning of users in social media. The data scheme used in these reports is similar to the one used here but our focus and methodology are distinct.

Three studies that share common characteristics with our political affinity perspectives are [19], [20] and [21]. These works investigate political information within social networks but each from a different perspective. In particular, the authors of [19] employ clustering methods in respect to the left and right leaning tweets, a concept that is related to our clustering approach. The authors of [20] propose a methodology for positioning news media on a one-dimensional Euclidean political space via the Jaccard similarity of their follower sets. This format is in accordance with the scheme produced by the application of the minimum linear arrangement problem in this work. Similarly, in [21], the political slant of articles are evaluated through the projection of the journalists' political preferences. The methodology presented is able to quantify the slant of a news article in a scale of -1 to 1, which can be parallelized with the quantifiable measures from the DeGroot model application of this study.

The Greek political scene was previously studied in [22], in which the authors

employ a learning model to predict the voting intentions during the 2015 Greek bailout referendum. Relevant tweets dating before and after the referendum are leveraged in order to examine the intentions of this spontaneous in nature event. In the context of referendums, the Twitter network is utilized in [23] as well to study the effects of news media in the 2016 constitutional referendum in Italy. The potential of Twitter as a platform of information dissemination and dialogue in Greece is also examined in [24] by applying content and thematic analysis on the tweets of the two biggest Greek political parties.

While our case study is the Greek political scene, previous literature on the behavior of political actors in the USA is very common. In [25], the authors examine the Twitter linkages between five major American political leaders, among them US President Donald Trump, with eight America hate groups (e.g. Anti-Immigrant and White-Nationalist). This appears to be in parallel with our investigation of linkages among politicians and news media. The follower-followee connections of the Twitter network are also utilized in [26] to identify a latent ideological dimension concerning political actors in USAs political scene.

An interesting recent study in [27] attempts to infer the political leaning of news outlets in the US by characterizing the followers and then relaying the followers' preference to the news outlets that they opt to follow. The authors claim that, overall, users tend to follow politicians with similar views and that those who follow Congresspeople on Twitter may have more polarized political tendencies that the overall US population. The results are achieved using the American for Democratic Actions (ADA) scores. The objectives of our work are similar to those in [27] but in this paper we establish methods that work in a multi-party context, and, moreover, don't require a quantitative starting point, like the ADA scores or any other prior knowledge about the involved parties.

In [8], the author uses the structural characteristics of the Twitter network to extract the political positions of politicians, users and news sources in five countries. He proposes a *Bayesian spatial following model of ideology* based on the popularity of the politicians, the political interest of users and their estimated ideal points on the political spectrum in order to predict the probability

of a user following a politician. Although the author's hypothesis coincides with our hypothesis (i.e., the mere structure of the Twitter network suffices for the extraction of the political inclination of specific users), his proposed model applies extensive filtering on the users' dataset (e.g., geolocation, tweet activity, number of followers) while in our approach we use raw data for our algorithms without filtering and without any other knowledge of the users' characteristics.

Finally, in [28], the authors leverage the demographics of the audience of the news sources, obtained through the advertiser interfaces of social media sites like Facebook and Twitter, to infer biases of news sources. In a different work [29], it is shown that opposing views of Twitter users can be reflected on the personalization of the corresponding Google News aggregator. In [30], a method for extracting information about the slant of a news article using related retweets and followers of Landmark users from Twitter, is presented. Selected Landmarks and connections and tweets from Twitter are used in [31] with a global positioning algorithm to map news media on the political spectrum. The close association between MPs and news sources is also studied in [32], where the political belief of a Twitter user is being inferred based on their links with news sources. All these related works are evidence that supports the view that there is significant political information in the Twitter network and that this information can be used to infer bias about news sources and, consequently, news articles. In this work, we show that political information can be extracted from Twitter even by using only the follower relations and that this information can be used to infer the bias and the affinity of news media.

3. Dataset and Methodology

In this section, we provide a description of the dataset and an overview of the methodology utilized to study the political affinity of the NOIs. Initially, a dataset is assembled from Twitter (Section 3.1), an online social network with distinctive political nature. We then explain how this dataset can be interpreted as a bipartite graph and suggest the appropriate transformation stage

(projection) in order to reduce the dataset into manageable size for the direct application of our algorithms (Section 3.2). Finally, we provide an overview of the proposed methodology in order to study the political affinity in the dataset (Section 3.3).

3.1. Dataset Description

The dataset that we assemble and use is based on the Twitter accounts of actors that are relevant to the Greek political scene. More specifically, we focus on (a) the members of the Greek Parliament (MPs), and (b) a list of the most acquainted news sources with nation-wide audience. We refer to these actors as NOIs (i.e., nodes of interest) since they occupy a significant share of information about the political scene in Greece.

The set of MPs was acquired from the official website of the Greek Parliament[1] without any discrimination. Summarily, the set of NOIs consists of 300 MPs, of which 166 have a public Twitter account that was either advertised in their personal websites or was a result of a query in the Twitter search engine. As a result, 134 MPs with either a protected account (5) or no account at all (129) could not be included in the dataset. Among the disregarded MPs is the party *KKE*, one of the eight political parties, representing the left wing of the Greek Parliament, of which none of the MPs have a Twitter account. Furthermore, 4 of the MPs were independent (not members of any party listing). We only considered the 162 MPs with an explicit party militancy as part of the NOI set (and not the 4 independent MPs). This decision was due to our methodology, which is based on a strict profile of political parties for both the evaluation of our experiments and the analysis of the news media affinity.

We also include a set of 24 well known news media in our dataset. In particular, the media contained in the dataset are: 16 printed newspapers that are nationally distributed, 6 TV channels with national broadcast range, and 2 online blogs. The selection of the news media is based on their nationwide

[1] `https://www.hellenicparliament.gr/en/Vouleftes/Ana-Koinovouleftiki-Omada/`

Table 1: Breakdown of NOIs into groups. The MPs are shown in the left column and the news sources in the right. For the MPs the table shows the number of parliamentary seats for each party as well as the number of MPs present in the dataset. The underlined parties form the government coalition.

Parliamentary Group	**Dataset**	**Seats**	**News Group**	**Dataset**
SYRIZA	62	145	Newspapers	16
ND	61	76	TV Channels	6
DHSY	17	20	Blogs	2
XA	10	16		
KKE	0	15		
ANEL	4	9		
POTAMI	6	6		
EK	2	6		
Independent MPs	0	7		
Totals	162	300	Totals	24

coverage, their interest in political news, their presence in Twitter and on our commitment to cover, to the greatest possible extent, the political spectrum of Greece. We consulted a group of political scientists for advice on the coverage of the greek political scene by our dataset. We note that some well established news media of the Greek scene are not included in our dataset since they do not maintain, to the best of our knowledge, an official Twitter account. The political scientists confirmed that under these preconditions our dataset is representative of the political spectrum at that particular period. In total, we collected 186 Twitter accounts from the above categories. A breakdown of the NOIs is presented in Table 1.

We complete our collection by crawling the followers of each of the NOIs using the Twitter API to construct a dataset of 186 NOIs, 1,279,005 unique followers and 5,610,099 connections between NOIs and followers. For each of

the users (NOIs and followers) only the Twitter user IDs are stored, while the connection is simply a pair of a NOI ID and a follower ID. It is worth mentioning that during this process we ignore the connections where both endpoints are NOIs. The rationale behind this decision is associated with our proposition to determine the political affinity of the NOIs without using information provided by their own actions directly. However, the amount of the NOI-to-NOI relations were less than 1% of the total following relations. Moreover, the 4 independent MPs as well as the additional news sources that are not included in the NOIs set are presented in the dataset as followers.

Finally, as a result of our data acquisition process, ties among the followers and non-NOI users are not included in the dataset. The process of obtaining this information is very demanding given the massive amount of followers and limitations imposed by the Twitter API but it is not considered necessary either since our methods do not leverage these connections.

The dataset was constructed on April 2018 and, thus, reflects the connections between the selected NOIs and their followers in the Twitter network, and consequently the political background, at that time. The dataset along with other supplementary material about this work are available online[2].

3.2. The Projected Graph

The acquired dataset can be naturally represented as a bipartite graph $G(N, F, E)$ where the two disjoint sets of vertices are the NOIs (N) and the followers (F) respectively, and an edge E_{ij} between a NOI i and a follower j exists iff j is following i and j is not a NOI. Many real world networks are naturally modeled as bipartite graphs, especially in social systems, like the Twitter follower network we use in this work. The average degree of the NOIs is 30,162 while the majority of the NOIs (91.4%) have less than 10^5 followers.

The dataset, however, is massive and possibly incompatible with general graph processing algorithms due to its bipartite nature. Thus, we transform the

[2] `https://figshare.com/s/7214b0b8f52544c85df9`

graph to its one-mode projection onto the NOIs, an extensively used method for compressing information about bipartite networks [33]. The one-mode projection of a bipartite network $G = (X, Y, E)$ onto X (X projection for short) is a weighted, complete, unipartite network $G' = (X, E')$ containing only the X nodes, where the weight of the edge between i and j is determined by a weighting function $\beta_G(X_i, X_j)$. The weighting method may not necessarily be symmetrical but in this work we engage in a simpler approach with commutative weight functions so that $\beta_G(X_i, X_j) = \beta_G(X_j, X_i)$, resulting in an undirected projection. Typically, the weight function expresses a form of similarity among the vertices in order to preserve the semantics of the original graph.

While the projection allows simplification of the network and compatibility with unipartite algorithms, it constitutes a lossy graph compression operation and consequently incurs information deficit over the original bipartite graph. There exists, however, no global weighting method of minimizing information loss and the optimal weighting is heavily dependant on the nature of the network and the objectives of the study. Therefore, we proceed with a selection of set theoretic functions that expose the similarity among the NOIs, namely the *Overlap coefficient*, the *Jaccard index*, the *Ochiai coefficient*, the *Sorensen-Dice coefficient* and the *phi coefficient*. The definition and a brief description of each weighting method is provided in Appendix A.

All weighting methods express similarity and are, hence, generally positively correlated. However, each projection method interprets the concept of similarity from a different perspective and, in that sense, they are all unique. Other similarity techniques mentioned in the literature are computationally demanding, for example the original SimRank [34] algorithm has a space requirement of $O(n^2)$. In this work, we utilize methods that can be computed trivially even for large scale input data, such as the Twitter network. Our observations suggest little room for further improvement over the effectiveness of these simple similarity measures.

The projection methods were applied on the bipartite graph G so that the weight of the projected edge between two NOIs x and y is evaluated by a function

$weight(E'_{xy}) = \beta_G(x, y)$. We did not take self loops into consideration because the weights would be trivially set to the maximum value. Since the projected graph is complete, there are $|N|(|N| - 1)/2 = 17{,}205$ undirected edges in each of the projections, including possible edges with zero weight.

3.3. Methodology

Our proposed methodology utilizes the projections of the Twitter follower network in order to estimate the political affinity of the NOIs. Specifically, our methodology includes a combination of methods, namely the *modularity clustering*, the *minimum linear arrangement problem (MinLA)* and the *DeGroot model approach with stubborn agents*. The selection of these methods highlights the conceptual diversity in the interpretation of political affinity. Each of these approaches provides a different perspective of the political affinity encoded in the projection graphs.

3.3.1. Modularity Clustering

Clustering or *community detection* in a graph refers to the process of identifying the modules and, possibly, their hierarchical organization, by using only the information encoded in the graph topology. Community detection has been widely applied in real-world social systems and various methods with different characteristics have been suggested [35].

More specifically, we use the algorithm in [36], a heuristic method that is based on modularity optimization and is commonly known as *Louvain optimization*. The algorithm is well established and has seen extensive use in the field of social networks [37]. Its complexity is linear on typical and sparse data. The Louvain optimization algorithm unveils hierarchies of communities and allows to zoom in the network and to observe its structure with the desired resolution via the parameter r. Therefore, the *resolution* parameter determines the desired number of communities in the partition. The parameter can be tuned accordingly in order to accommodate the requirements of specific experiment settings.

The application of modularity clustering to the dataset enables us to study the political affinity of the NOIs from a perspective that conveys the group affiliation, i.e. the affiliation of a NOI with a specific political party. Thus, our expectation is the partition of the NOIs into sets of communities that represent the political parties.

3.3.2. Minimum Linear Arrangement

The *Minimum Linear Arrangement* (MinLA) problem consists in finding an ordering of the nodes of a weighted graph, such that sum of the weights of its edges is minimized . More formally, given a finite graph $G = (V, E)$ of order n with weighted adjacency matrix w, the MinLA problem is the problem of finding a vertex labeling $f \to \{1, 2, ..., n\}$ such that the sum $\sum_{(u,v)\in E} w_{uv}|f(u) - f(v)|$ is minimized over all possible labelings [38].

Regarding its computational complexity, on general graphs MinLA is an NP-complete problem, thus one has to resort to heuristics or approximation algorithms to obtain a solution. However, there are no "good" approximation guarantees for the MinLA problem, either. The best known result, is the $O(\sqrt{\log n}\log\log n)$-approximation algorithm presented in [39]. On the other hand, in [40] it is shown that no Polynomial Time Approximation Scheme (PTAS) exists for MinLA and in [41] that it is SSE-hard to approximate MinLA to any fixed constant factor.

The MinLA problem has been applied to various scientific fields, for example in VLSI design [42] in order to minimize the electrical resistance of a circuit, or in a theoretical level [43] but, to the best of our knowledge, its physical interpretation has not been studied on a specific social network theme in prior literature. We argue that the MinLA problem is suitable for application on this context and constitutes an innovative approach for the analysis and understanding of social networks. Our hypothesis is that the application of a solution of the MinLA problem to the graph projections will unveil the positioning of the NOIs in a bipolar spectrum and, eventually, in a political spectrum. The intuition behind this is that NOIs with similar political views and, hence, stronger bonds

in the projection, should occupy successive labels in the MinLA ordering, while NOIs that share weaker links should be distantly positioned.

For the purposes of this work , we designed and implemented a simplistic randomized local search algorithm, repeated over a set of uniformly random initial rankings, which approximately leads to the minimum LA. Given an initial guess of the arrangement we perform a sequential series of steps to determine a local minimum of the cost function, the *fast converge phase* and the *local converge phase*. During the fast phase, the algorithm performs random swaps on the elements of the arrangement for a number of repetitions, and maintains the best arrangement in terms of cost. The purpose of this phase is to allow the algorithm to quickly descend close to a local minimum while the number of repetitions involved determine the convergence rate. We selected to perform this step n^2 times as we empirically observed a sufficiently quick convergence for this setting. During the local phase we validate that the current LA is the local minimal cost LA by performing all possible swaps in it; if there is a swap that improves the cost we restart the process until we identify the local minimum. The above process of computing a local minimum is repeated several times with random initial arrangements and the best solution is kept. The scheme is presented in Appendix B.

3.3.3. The DeGroot Model Approach

Moris H. DeGroot presented in [44] a simple yet efficient model about opinion diffusion in a social graph. The core idea of his model is that individuals tend to adopt the opinions of their friends. According to the model's *opinion update rule*, given a social graph $G = (V, E, O)$, where V represents the vertices (i.e., individuals), E the edges amongst them (i.e., friendships) and O the opinions of nodes $i \in V$ about a specific topic as real valued o_i, each individual i updates o_i to o_i' by averaging the opinions of its friends. When *trust factors* are introduced to the friendships (i.e., weights), each member updates its opinion according to the weighted average of its friends' opinions. The process is repeated and, under certain condition, the opinions of the nodes converge signifying a consensus in

the graph. DeGroot underlined the mathematical coherence of the process to Markov chains. He proved that the final opinion, when convergence occurs, depends solely on the structure of the graph and the initial opinions of its members.

In [45], Ghaderi and Srikant enriched DeGroot's model with *stubborn agents* (i.e. nodes that are fully or partially biased towards an opinion) and studied its convergence. They remarked the common underpinnings of their extension with Markov chains and proved that "the model converges to a unique equilibrium where the opinion of each agent is a convex combination of the initial opinions of the stubborn agents". Moreover, the contribution of stubborn agent s in node's i final opinion is the probability of a random walk hitting s given it started from i, namely the *hitting probability*.

Based on the findings of Ghaderi and Srikant, we present a technique to estimate the political affinity of the NOIs. We consider each NOI projection as a social graph where the weights of the links produced by the projection methods correspond to the trust factors of the nodes to their neighbors. In the case of the phi projection, the graph contains edges with negative weights, a phenomenon that does not abide by the restrictions of the DeGroot model. Therefore, transformations of the original formula are utilized (namely the ϕ_G^{add} referred as Phi-A and ϕ_G^{exp} as Phi-E) as described in Appendix A. The undirected edges of the graph are duplicated into two opposite directed arcs and, in order to abide by the DeGroot model restrictions, the weights of each node's outgoing arcs are normalized. Seven additional nodes are introduced to the graph that represent the political parties and each MP's node is linked to the corresponding party.

Figure 1 presents an abstract example of a NOI projection with three MP nodes (MP_1, MP_2, MP_3) and one news media node (MM_1), enriched with two party nodes ($Party_1$, $Party_2$). According to the needs of individual experiments, we perform slight modifications to this structure to ensure compliance with the respective evaluation goals. Specifically, depending on the setting, a MP can be transformed into a stubborn node by removing its outgoing arcs to other

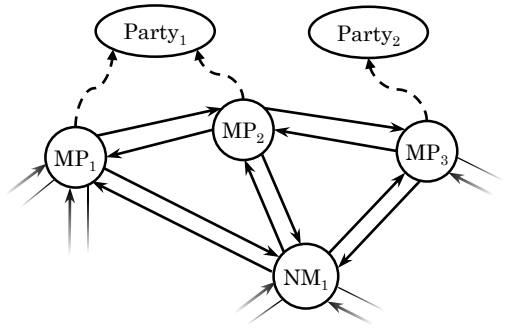

Figure 1: Abstract example of a NOI projection, enriched with party nodes.

NOIs and have its sole influence exerted by its respective party. Conversely, when a MP updates its opinion according to the DeGroot model, its friendships to other MPs (outgoing compact arcs) are used and its link with the party is ignored. Hence, the direct link of a MP to its party and the MP's friendships are mutually exclusive.

The use of random walks leverages the connections among the MPs as well as the links between the MPs and their parties to quantify the latent relationships of MPs with all political parties. Consequently, our heuristic can also uncover the associations among the news media and the parties through the intermediate edges with neighbouring MPs in the projections. This method reveals a perspective of political affinity that differs from the NOI clustering and the MinLA approach as it relies on the hitting probabilities of random walks to determine the influence factors between the various actors in the political graph.

4. Proof of Concept: The MPs Case

The fundamental concept of our study is that the mere structure of a social network consisting of nodes with political interests suffices for the extraction of rich political information through the use of innovative algorithms. In order to confirm this assertion, we firstly apply our proposed methodology on a subset

of the acquired dataset that ensures, to the greatest possible extent, the existence of political information and confines any source of politically irrelevant information that may falsify the results of our methods. We consider this as the proof of concept scenario of our proposed methodology.

4.1. The MPs Dataset and Projections

In the MPs case, we use a subset of the acquired dataset that contains the MPs and their Twitter followers as well as the connections between them. In particular, the dataset includes 162 NOIs (i.e., the Twitter accounts of the MPs), 740,580 followers and 2,403,200 connections between NOIs and followers. We note that the news media presented in Section 3.1 are considered as mere followers in the context of this scenario and any connection with the MPs is included. We can safely argue that this confined dataset incorporates to the greatest possible extent the available political information about the greek scene since the particular NOIs (i.e., the greek MPs) exhibit profound political behaviour and it is only natural to assume that their followers are politically motivated.

The raw data are perceived as a bipartite graph (as described in Section 3.2) and are transformed into projected graphs using the projection methods described in Appendix A. We refer to the bipartite graph and its projections as *MPs graph* and *MPs projections* respectively.

4.2. Experiments and Results

The proof of concept experiments utilize our methodology in order to confirm the existence of rich political information in the dataset. The application of *Modularity Clustering*, the *Minimum Linear Arrangement (MinLA)* and the *DeGroot Model Approach* are described separately and our results are being presented in the following sections.

4.2.1. Modularity Clustering

Initially, we apply the modularity maximization algorithm (Section 3.3.1) on the MP projection graphs in order to partition the vertex set into disjoint groups

of MPs and show that this method can reveal the underlying political structure of our dataset. Our hypothesis is that modularity clustering will partition the MPs into their respective political parties, or, equivalently, that the MPs of the same party will be classified into the same cluster. Consequently, the evaluation of the clustering method is performed towards the true partition of MPs in political parties (Table 1) which is an objective indication about their political affinity. As a result, we tune the resolution parameter so that the algorithm returns 7 clusters, the amount of political parties in the ground truth.

The quantifiable evaluation can be achieved by reducing the partition correlation problem into a set similarity problem using the concept mentioned in [46, Section 2.2.1]. More specifically, for some partition of the nodes $[N]$ into groups we consider the set S to comprise all unordered node pairs $\{i, j\}$, with $i \neq j$, where elements i and j belong to the same group in that partition and S' to consist of all other pairs. Naturally, it has to hold that $|S| + |S'| = |N| \cdot (|N| - 1)/2$.

The resulting sets S and S' can then be used as input to our evaluation methods, which are the Jaccard index, the Simple Matching Coefficient (SMC), the F1 score, the Normalized Mutual Information (NMI), the Pearson correlation and the Cosine similarity. These measures are used to assess the effectiveness of a partition and are different from the measures used to construct the projection, although some of them are used for both purposes. The results are shown in Table 2. The rows of the table refer to the similarity functions used for the projection. Each function is represented by the minimum and maximum values of the respective evaluation measure over all resolutions between 0.2 and 3.0 with a step of 10^{-3} that yielded 7 clusters. The columns of the table denote the evaluation measures. For comparison, the random partitioning is also included in the table. This random evaluation was produced separately for each measure/column by gradually generating random partitions of 7 communities (as many as the political parties) until the average of the correlations did not change beyond the 9th decimal digit.

The results deliver a strong evidence about the validity of our hypothesis, stating that MPs of the same political party will be classified into the same

Table 2: Clustering evaluation of the MP projections. Each projection displays the maximum (odd line) and minimum (even line) value of the respective evaluation measure.

β_G	Evaluation measure					
	Jaccard	SMC	F1	NMI	Pearson	Cosine
Overlap	**.8277**	**.9456**	**.9057**	**.6671**	**.8693**	**.9066**
	.7867	.9314	.8806	.6040	.8346	.8816
Ochiai	.5449	.8436	.7054	.3148	.6117	.7118
	.4724	.8197	.6417	.2515	.5459	.6544
Phi	.5110	.8384	.6764	.3055	.5968	.6911
	.4358	.8133	.6071	.2427	.5276	.6298
Jaccard	.4763	.8244	.6453	.2663	.5584	.6608
	.3759	.7831	.5464	.1661	.4452	.5678
Sorensen	.4576	.8144	.6279	.2386	.5312	.6418
	.3981	.7901	.5695	.1810	.4664	.5880
Random	.1062	.6475	.1920	.0000	.0000	.2046

group in the partition. The strength of the correlations among the weighting methods varies, although all methods had an above random association. In particular, the overlap coefficient firmly outperforms other functions commonly mentioned in the literature on all evaluation measures. Therefore, this indicates the existence of rich political affinity information within the Twitter follower network, and substantiates the suitability of modularity clustering for obtaining this information.

Figure 2 displays a force-directed visualization, produced by Gephi [47], of the MP projection using the overlap function with a resolution of 0.855. These settings correspond to the highest correlation achieved (the first line of Table 2). Vertices in this layout are colored by their modularity class. The respective party of each node (the ground truth) is given as a text label within the node while the party IDs are given in the legend in the top left corner. The partition of

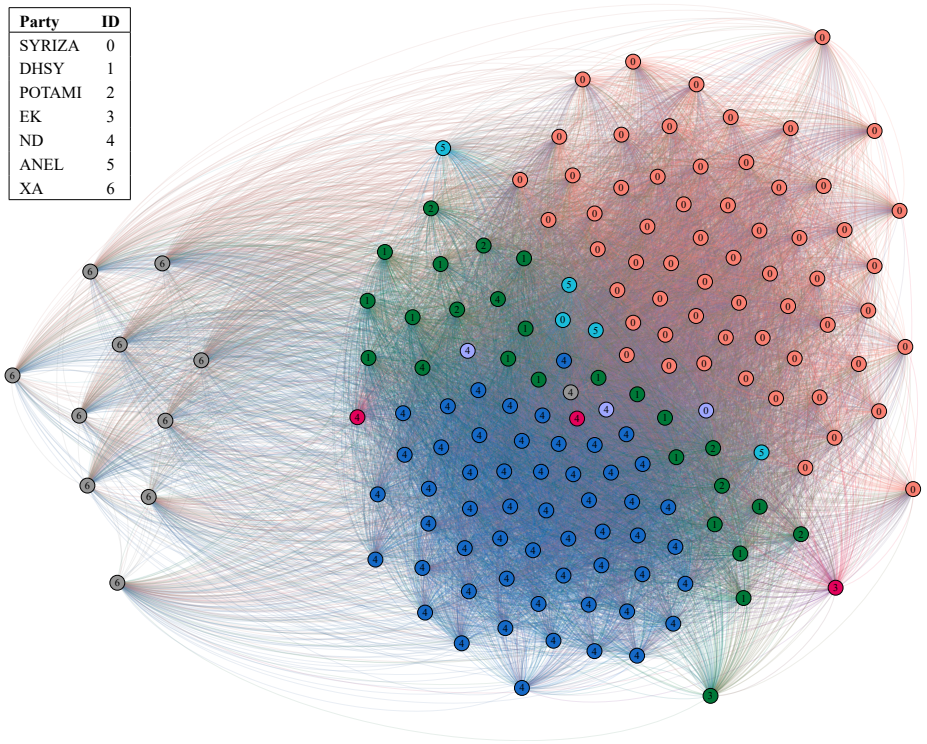

Figure 2: Force-directed visualization of the MP overlap projection.

the clustering illustrated in this figure is a result of only the Twitter follower-followee relations while the real distribution of MPs in political parties (the ground truth) is only used for the evaluation. An important visual observation is that the accuracy of the identified clusters is remarkably high, which coincides with the results in Table 2. In particular, the biggest political parties (SYRIZA, ND, XA) are clearly identified with the respective clusters almost flawlessly. Additionally, the layout provides a visual perception of the close association between modularity clustering and force-directed placement [48].

4.2.2. Minimum Linear Arrangement

In this experiment we apply our MinLA algorithm to the MP projections in order to arrange the MPs in a one-dimensional space and study the significance of this ordering. Our hypothesis is that MPs of the same political party will

appear consecutively inside the minimum cost arrangement of the MP projection vertices; the known affiliations of MPs in political parties enables us to evaluate this. Finally, we make an attempt to attribute the physical meaning of the minimum cost arrangement in relation to the left-to-right political axis.

Initially, we apply this algorithm to all of the MP projections and, for each, we obtained the minimum cost arrangement m and its cost $C(m)$. Since in this experiment we deal with ordinal data, we also define the concept of party ordering. Our dataset contains 7 parties, so there are $\rho = 7! = 5{,}040$ possible orderings. Each of these orderings can be flattened to a ranked list of MPs, where MPs of the same party are tied on the same rank. Thus, there are also ρ flattened MP ranked lists denoted as R_i, $1 \leq i \leq \rho$.

We assess the correlation of m with every ordering R_i using the Kendall tau-b (τ_B) correlation coefficient [49], which is a statistic used to measure the ordinal association between two measured quantities. The tau-b correlation coefficient is a generalization of the Kendall tau-a coefficient that accounts for ties in the input lists, specifically present in the R_i orderings. It is worth noting that Kendall tau-b is in range $[-1, 1]$ but, since in our context the linear arrangements (LAs) cannot contain ties, the maximum value is

$$K_{max} = \frac{T(n) - \sum_i T(t_i)}{\sqrt{T(n)}\sqrt{T(n) - \sum_i T(t_i)}}, \text{ where } T(x) = \frac{x(x-1)}{2},$$

which equals 0.8361 because $t = [62, 61, 17, 10, 6, 4, 2]$ (Table 1). Afterwards, we find the party ordering with the highest correlation to m, defined as R_q where

$$q = \arg\max_{i \in [1,\rho]} \tau_B(m, R_i).$$

Our results are presented in Table 3, which displays the tau-b correlation of each projection's minimum cost LA against its respective R_q. The results are in agreement with our findings in Section 4.2.1 in regards to the effectiveness of the overlap projection and the existence of rich political information within the Twitter follower dataset. Specifically, the $\tau_B(m, R_q)$ of the overlap coefficient is 86.8% of the maximum (0.7261/0.8361) proving that the MinLA problem definition highlights the clustering features of our dataset and confirms our

Table 3: MinLA evaluation of the various MP projections. The max τ_B is 0.8361 while the random τ_B is approximately 0.1149. The costs among the projections refer to different edge weights and, thus, are not comparable.

β_G	Minimum (Found) Cost	Random Cost	τ_B
Overlap	162,196	225,083	0.7261
Phi	51,230	83,352	0.7228
Ochiai	55,308	88,322	0.7008
Jaccard	19,628	38,216	0.3933
Sorensen	36,624	68,293	0.3867

hypothesis. Moreover, the phi and Ochiai based projections are also represented by very promising correlations that are only marginally below overlap. The random cost column of Table 3 is derived from the average distance of two nodes in a random LA which is $(n+1)/3$ (Lemma 1 in Appendix C).

Furthermore, we experimentally found that the $\tau_B(r, R_q)$ of the random MP ordering r is 0.1149. Consequently, it follows that, despite their differences, all of the projection methods reveal some amount of information from the follower network with above random significance.

A visual perception of the MinLA application of the overlap projection is shown in Figure 3. The top ruler in the figure displays the minimum cost LA m while the bottom represents the closest party arrangement R_q. Each ruler contains 162 MPs represented by points colored by the real party of the respective MP. The figure offers an alternative understanding of the magnitude of correlation between these vectors and, overall, the validity of our hypothesis.

Finally, we discuss some interesting observations about the closest political party ordering R_q of the overlap projection which is [SYRIZA, ANEL, POTAMI, DHSY, ND, EK, XA]. A comparison of the R_q to the arrangement of the parties based on their ideological identity reveals interesting properties of our result and of the inherent political information in our dataset. According to their self-identification and data from additional resources (e.g. Wikipedia), the most

(a) Resulting minimum cost LA

(b) Closest party arrangement

Figure 3: Visualization of the minimum LA for the overlap projection. The top figure is the resulting minimum cost LA and the bottom one of the closest party arrangement for it. The τ_B correlation between the two arrangements is 0.7261.

credible arrangement of the parties on the left-to-right political spectrum is [SYRIZA, DHSY, POTAMI, EK, ND, ANEL, XA]. In R_q, ANEL is adjacently positioned to SYRIZA, an oxymoron phenomenon that can be justified by the fact that ANEL and SYRIZA were in governmental coalition and, thus, their ties are strong in the Twitter follower dataset. Furthermore, the swap of DHSY and POTAMI in R_q is inconsequential, especially after their deliberations (in April 2018) about the formation of a new upcoming coalitional party (KINAL) for the upcoming elections. The misplacement of EK can be attributed to its small footprint (2 MPs) and, hence, by deficient information. In general, we can argue that R_q outlines the parties on one dimension according to the followers' criteria that are a combination of the left-to-right political perspective and the pro and anti-government feeling.

4.2.3. The DeGroot Model Approach

In this section, our DeGroot model approach is applied to the MP projections in order to determine the influence factors towards the political parties. These factors are then used to classify every MP to the party with the maximum influence factor. Our hypothesis coincides with the clustering hypothesis in Section 4.2.1; MPs will have their dominant influence factors on their respective affiliated party.

We perform a series of experiments for all the MP projections that are based on the concept of the leave-one-out cross-validation method, where each MP

Table 4: Hits of the leave-one-out cross-validation method for each projection.

	Overlap	Ochiai	Jaccard	Phi-A	Phi-E	Sorensen
SYRIZA	53	52	48	49	49	46
ND	59	57	53	55	57	53
DHSY	16	16	13	7	9	13
XA	10	10	10	10	10	10
ANEL	3	3	0	0	0	0
POTAMI	1	6	6	0	0	0
EK	0	0	0	0	0	0
Total Hits	142	144	130	121	125	109
Hit ratio	87.65%	88.89%	80.24%	74.69%	77.16%	67.28%

is selected individually. The directed arc of the selected MP to its party is ignored while the rest of MPs are transformed into stubborn agents by removing their outgoing arcs to other MPs as explained in Section 3.3.3. The selected MP's friendships with other NOIs are used to calculate a random walk's hitting probabilities to every party's node given it originates by the MP. Since the parliamentary groups are uneven, the evaluated probabilities are divided by the corresponding group's size in order to compute the *uniform per party influence* and avoid any dominance effect by the parties with large parliamentary groups. The uniform influences are used to classify each MP to a party based on the greatest uniform influence of their random walk. A hit is considered when the party with the greatest uniform influence on the MP coincides with its actual party. The experiments are implemented using PRISM [50], a tool that is widely used to analyze models that exhibit probabilistic behavior (e.g. Markov chains).

The results presented in Table 4 denote that our approach achieves surprisingly high hit ratio in almost all cases of projections, a clear indication that the MPs dataset and, consequently, the MPs projections contain significant political information. The highest hit ratio is achieved in the graph produced by the

Ochiai projection (88.89%) while the result of the corresponding overlap projection is slightly lower (87.65%). Moreover, the Ochiai and the overlap projections provide sufficiently robust graphs that enable the correct classification of MPs of smaller parties (e.g. ANEL, POTAMI). This property is also valid for the graph of the Jaccard projection although it exhibits noticeable weaknesses in the classification of MPs of the two largest parliamentary groups (SYRIZA and ND). In general, the vast majority of the MPs projections produce graphs that achieve high levels of accuracy in the classification of MPs to their parties.

4.3. Discussion

The results of all three methods presented in the previous sections provide clear indications that our assembled dataset contains significant political information and the applied algorithms are efficient in extracting it. The selected methods succeeded in revealing different aspects of the political information. The results of modularity clustering indicate that the followers of NOIs suffice for the efficient detection of the actual parties while Minimum Linear Arrangement produces a ranking of the NOIs that can be interpreted as a political bipolar. Finally, the DeGroot Model Approach exhibited surprisingly successful behavior in highlighting the affiliations of the NOIs to the political parties.

The results of our proposed methods also allow a comparative analysis of the weighting projection methods and their efficiency in conveying useful information in the projections. The Sorensen-Dice coefficient achieved the weakest scores in the applied algorithms while the Jaccard coefficient performed averagely in the DeGroot Model Approach and poorly in the other methods. The phi coefficient provided efficient scores in all methods applied and the Ochiai weighting method achieved the highest scores in the DeGroot Model Approach. In general, the overlap coefficient appears superior to other measures. Although, it does not always attain the best evaluation in all the scenarios, it qualifies for a very consistent and reliable weighting function among the proof of concept experiments in our dataset.

5. The Case of News Media

The promising results of our methodology on the purely parliamentary dataset of the previous section justify our attempt to deploy the presented algorithms on a politically obscure scene. We consider a set of popular news media in Greece as the case study and utilize our proposed methodology to examine its effectiveness. For evaluation purposes, we conduct an experts' survey and compare its findings to the results of our algorithms.

5.1. The News Media Dataset and Projections

In this case study, we utilize the complete dataset of our work. The dataset includes 186 NOIs (MPs and news sources), their followers and the connections between NOIs and followers (see Section 3 for further details). The bipartite graph formed by this data is projected onto the NOIs using the projection methods of Section 3.2 to create the *enriched projections*, which are then provided as input to the methods presented above.

5.2. Expert Survey

The purpose of the expert survey is to establish a factuality that we consider as ground truth about the political affinity of the news media that are present in our case study. We resort to the expertise of 8 scientists from the field of political science to provide us insight about the political affinity and orientation of news sources in Greece. We structured and provided a survey questionnaire about the 24 news media of our dataset.

The questionnaire involved two questions, which we refer to as *political affiliation* and *political orientation*. The first question aimed at providing means of quantifying the relationship among news sources and political parties. The participants were asked to label the relationships among all pairs of news sources and political parties with one of the options "-2 hostile", "-1 negative", "0 neutral", "1 positive" and "2 partisan". The second question aimed at classifying the news sources in a left-to-right political spectrum scale. The experts were

asked to label each news source with one of the options "far left", "left", "center", "right" and "far right". These options represent the position of the news media in the political spectrum and do not directly imply an association with a political party as in the first question.

The responses of the participants are aggregated using the average of individual answers. The data of the first question were processed into a 24×8 matrix (24 news media and 8 political parties) of political affiliation values in the range [-2,2] while the responses of the second question were assigned values in the range [-2,2] (i.e., -2 = "far left", -1 = "left", 0 = "center", 1 = "right", 2 = "far right") and the average values denote the political orientation of each news media derived from the opinions of the experts. We consider these findings of the survey questionnaire as the ground truth that we can deploy in the application of our methods in the news media case study. The raw answers of the survey are included in the supplementary material of this work given in Section 3.1.

The reliability and homogeneity of the expert survey was assessed using Cronbach's alpha [51]. The issue of missing data (245 of 1728 records) was solved using a variety of imputation techniques [52] (namely the k-nearest neighbours, multivariate imputation by chained equations (MICE), expectation maximization (EM), mean, mode, median and random imputation) and listwise deletion. The Cronbach's alpha values that were calculated for all these missing data handling methods ranged from 0.929 to 0.956 indicating acceptable (≥ 0.7) internal consistency of the conducted survey.

5.3. Experiments and Results

The final step of our case study involves the application of the Modularity Clustering, the Minimum Linear Arrangement (MinLA) and the DeGroot Model Approach to the enriched projections. The outputs of the algorithms outline the political profile of the news media under different prisms. The results are evaluated using the findings of the experts' survey.

5.3.1. Modularity Clustering

The methodology explained in Section 4.2.1 is reproduced for the 5 enriched projections and from the resulting partitions the news sources are filtered. It is then possible to use the expert responses of the political affiliation question for the evaluation of the clustering method. Specifically, we assign each news source to a cluster based on the political party that is most affiliated with that source and, hence, creating a comparable structure.

This process yields 5 groups of news sources but one of the news sources is tied in two parties, one party with 13 NOIs (including the tied news source) and a singleton community comprising only the tied news source. However, since the evaluation method relies on pairs of nodes inside the clusters, it is not able to distinguish a singleton cluster. Furthermore, the method cannot operate on overlapping partitions and, thus, we naturally place the tied news source into the bigger community, eliminating the singleton group. This action is inconsequential since that particular node has the same level of political affiliation for both parties. Therefore, the resolution parameter of the clustering method is, similar to Section 4.2.1, tuned for 4 communities.

The results are shown in Table 5, which has the same format as Table 2. For each projection method, the minimum and maximum values of the respective evaluation measure over all resolutions that yielded 4 clusters is displayed. For comparison, the random partition with 4 communities is also given. The table carries similarities with the experiments on the MP projections. More specifically, given the significance of the measures, the existence of political information within the Twitter follower dataset is further established. It also appears that the modularity optimization clustering is suitable for the examination of the political affinity of the news sources within our dataset. Moreover, the overlap similarity appears to retain more information about the objectives of this experiment, although all of the projections achieved better than average accuracy.

Table 5: Clustering evaluation of the enriched projections. Each projection displays the maximum (odd line) and minimum (even line) value of the respective evaluation measure.

β_G	Evaluation measure					
	Jaccard	SMC	F1	NMI	Pearson	Cosine
Overlap	**.5072**	**.7536**	**.6731**	**.1736**	**.4762**	.6734
	.4488	.7355	.6196	.1467	.4367	.6226
Ochiai	.4747	.6884	.6438	.1406	.3927	.6768
	.2865	.4638	.4455	.0009	.0343	.4470
Phi	.5029	.7065	.6692	.1566	.4386	**.6865**
	.2825	.5000	.4405	.0026	.0588	.4432
Jaccard	.4067	.6957	.5782	.0810	.3241	.5787
	.3033	.4674	.4655	.0000	.0051	.4698
Sorensen	.4067	.6812	.5782	.0764	.3187	.5832
	.2757	.4891	.4322	.0004	.0240	.4368
Random	.1722	.5685	.2927	.0000	.0000	.2987

5.3.2. Minimum Linear Arrangement

In Section 4.2.2, the application of the MinLA problem in the MP dataset (through the MP projections) demonstrated the suitability of our methodology as well as the existence of profound political information within the Twitter follower network. For the purposes of the case study, we apply the same methodology on the projections of the enriched graph. Our motive is to confirm the previous findings in Section 4.2.2 and to examine new hypotheses about the news sources by utilizing the results of the expert survey.

More specifically, the application of the same MinLA algorithm in each enriched projection yields an arrangement m of the NOIs in a line, which is similar to the arrangement in Section 4.2.2 but in this case study it contains the news sources in addition to the MPs. It is possible to use m as source to construct two sub-arrangements m_{mps} and m_{news} which consist of only the MPs and

Table 6: MinLA evaluation of MPs using two datasets.

β_G	τ_B of MP projections	τ_B of enriched projections
Overlap	0.7261	0.7587
Phi	0.7228	0.7617
Ochiai	0.7008	0.4315
Sorensen	0.3867	0.3809
Jaccard	0.3933	0.3875

news sources respectively, where the relative ordering of the NOIs is preserved inside the sub-arrangements. In the undermentioned text, we study these two sub-arrangements separately.

Initially, we evaluate m_{mps} against the MP distribution in the political parties in the same way as in Section 4.2.2 while also contrasting the results. The two arrangements differ only on the dataset used and, thus, this evaluation shows how the addition of the news sources affected the political information in the dataset. The results are shown side by side in Table 6. It is clear that the addition of the news sources in the dataset did not diminish the amount of political information within the dataset. In fact, the projection methods that did well with the MPs dataset (overlap and phi), also performed well in the enriched dataset. This is a strong indication that the clustering features within the Twitter follower network are maintained even after the enrichment with the news sources. Furthermore, the Sorensen and the Jaccard projections had negligible differences while the Ochiai projection received a significant drop in effectiveness.

The evaluation of m_{news} is performed against the replies about the political orientation in the expert survey. Specifically, the political orientation vector given in Section 5.2 indirectly creates a ranked list of the news media sorted by their orientation. The evaluation of m_{news} is performed against this ranked list. This process is semantically very different from the m_{mps} evaluation. More

Table 7: MinLA evaluation of the various enriched projections. The max τ_B is 0.9799. The costs among the projections refer to different edge weights and, thus, are not comparable.

β_G	Minimum (Found) Cost	Random Cost	τ_B
Overlap	260,787	351,758	0.6249
Phi	71,732	115,967	0.4030
Ochiai	77,727	124,053	0.4030
Sorensen	46,512	93,088	0.1960
Jaccard	24,816	51,930	0.1738

precisely, the political orientation of the news sources corresponds to a linear scale of the news sources in the left-to-right political spectrum. As such, the evaluation of m_{news} is an indication of the minimum cost MinLA news sources arrangement correlation with the left-to-right political spectrum orientation.

The results of our experiment are reported in Table 7, which has the same form as Table 3. The max τ_B was calculated using the formula given in Section 4.2.2; due to the low amount of ties, the max τ_B is much higher than the respective value in Section 4.2.2. The table confirms some of our previous findings regarding the effectiveness of the overlap projection and further demonstrates the potency of our methods.

The inspection of the τ_B correlation between the news sources minimum cost arrangement and the expert survey political arrangement provides interesting insights. More specifically, although the result is clearly very significant, the correlation is not as high as the experiments for the MPs dataset in Section 4.2.2, which can be attributed to a variety of factors. Initially, given the semantics of the two methods, it is not meaningful to directly compare them because the arrangement with the MP projections displays clustering correlation while the arrangement of the news sources shows a measure of precise arrangement in a linear axis, which is by its nature a more difficult problem. Moreover, it is possible that the minimum cost arrangement of the news sources might not correspond exactly to the left-to-right political spectrum because the dynamics

of the Twitter network are very complex and heterogeneous among its users.

5.3.3. The DeGroot Model Approach

The application of the DeGroot Model Approach on the enriched projections aims at the extraction of the political affinity and orientation of the news sources. The NOIs of the enriched projections are handled using the same guidelines of Section 4.2.3. Each undirected edge is duplicated into two opposite directed arcs and the weights of the outgoing arcs of every node are normalized. Furthermore, seven additional nodes are introduced that represent the political parties and the MPs are linked to their corresponding party. We note that the nodes of the news sources are not directly connected to any party but their connections with the MPs are the indirect link with the parties that we aim to examine and evaluate.

We transform the nodes of the MP's into stubborn agents (i.e., nodes that are solely influenced by their party and their connections to other MPs are ignored) and we estimate the political affinity of each news media node by evaluating the hitting probability of a random walk that originates from it to each party's node. The retrieved probabilities are then divided by the corresponding parliamentary group's size, in order to avoid any dominance effect by the largest groups.

A series of preliminary experiments prove that the addition of the 24 nodes of the news sources to the graph does not taint the validity of our approach. We apply the DeGroot model in order to classify each MP to a party (using the same methodology as in Section 4.2.3) and collate the results in Table 8. The hit ratios on the classification of the MPs to their parties are slightly decreased in most cases while in the case of the graph produced by the overlap projection the ratio is increased (from 87.65% to 88.27%). These encouraging results allow as to proceed to the evaluation of the political affinity and orientation of the news sources.

Political affinity extraction. The political affinity extraction of the news media to each party is achieved through the calculations of the hitting probabilities for

Table 8: Hit ratios of MPs based on the leave-one-out cross-validation method for purely parliamentary (only MPs) and enriched projections (MPs and news sources).

	Purely parliamentary graph	Enriched graph
Overlap	87.65%	88.27%
Ochiai	88.89%	87.04%
Jaccard	80.25%	80.25%
Phi-A	74.69%	72.84%
Phi-E	77.16%	74.69%
Sorensen	79.01%	77.78%

random walks that originate from the news media nodes to the party's node. Our hypothesis for this experiment states that the news media that share common ideological and political views with a specific party should also share common followers with it and, thus, their links to the party's MPs in the projection graphs should be strong and would result to an increased hitting probability of a random walk that originates from the news media node to the specific party's node.

We test our hypothesis on the social graph produced by the overlap projection since our preliminary experiments suggest that the addition of the 24 news media NOIs enhance the mechanism of the DeGroot model in the extraction of political information. The experiment produces a 24×7 matrix that contains the uniform per party influences of each party to the news media node. We round these influences to 2 decimal places to produce coarse grain results and avoid any jitter.

In order to evaluate the validity of our approach, we correlate the results of our experiment with the results of the political affiliation from the expert survey. More specifically, we produce separate rankings of the news media according to their influences to the nodes of all the parliamentary groups in our dataset and correlate them with the findings of the experts' survey. We assess the correlation using the Kendall tau-b (τ_B) and the Pearson correlation coefficient (ρ). The

results presented in Table 9 validate our approach. The values of the Kendall tau-b coefficients are in most cases greater than 0.5. The results for ANEL and EK (the two smallest parliamentary groups in the graph) are significantly lower, a phenomenon that could be attributed to the small number of corresponding NOIs that are included in our dataset (4 and 2 NOIs respectively). In general, the findings provide a clear indication of dependence between the produced rankings and the experts' survey results. Furthermore, the high values of the Pearson coefficient ρ in almost all cases confirm our hypothesis that our method reveals information about the political affinity of news media from our dataset.

In order to make these results more easily understandable to a broader audience, we utilize the metric of *Precision at K (P@K)* to calculate the precision of the top K elements in the produced rankings in respect with the top K elements of the findings from the experts' survey. The results presented in Table 9 further confirm the correlation of the produced rankings with the experts' survey results. The values of *P@5* are surprisingly high in the cases of SYRIZA, DHSY, POTAMI and XA while the low values of ND, ANEL and EK can be attributed to the small number of NOIs in the dataset (in the cases of ANEL and EK) and to divergence of opinions between the political scientists and the Twitter users about the news media that are the greatest supporters of ND. The values of *P@10* exhibit significant consistency ranging between 0.8 and 0.5 providing, thus, further support to our findings.

Political orientation extraction. We further extend our approach and alter our experimental scenarios to extract information about the political orientation of the news media from our dataset using the DeGroot model. We modify our produced social graph with stubborn agents by discarding the nodes of 5 parties (the nodes of ND, DHSY, ANEL, POTAMI and EK). The remaining 2 party nodes (SYRIZA and XA) are considered as the two poles of the left-right ideological spectrum, according to their self-identification, that are present in our dataset. We evaluate the hitting probabilities of random walks that originate from the news media nodes to the 2 parties' nodes. Our hypothesis is that the

Table 9: Correlation coefficients (Kendall tau-b and Pearson), P@5 and P@10 values of the news media rankings for all parliamentary groups of the DeGroot experiment compared with the experts' survey.

Ranking Probability	τ_B	ρ	*P@5*	*P@10*
SYRIZA	0.57063	0.76702	0.8	0.8
ND	0.51756	0.81495	0.2	0.6
DHSY	0.63157	0.75014	0.8	0.6
POTAMI	0.58757	0.75780	0.8	0.6
ANEL	0.11213	0.03983	0.2	0.5
XA	0.57366	0.64034	0.8	0.7
EK	0.31555	0.57076	0.4	0.5

MPs' location in the political spectrum is reflected in the arcs of our social graph and, thus, the news media links to the MPs result to high hitting probabilities to the pole of the political spectrum that are closer to.

The results of the experiment produce a ranking of the media based on their "distance" from the pole of SYRIZA (i.e., the party that is considered to represent the leftmost pole in our political spectrum). The ranking is then correlated to the findings of the second question from the experts' survey concerning the political orientation of the news media. The Kendall tau-b coefficient is evaluated to 0.58792, a result that indicates strong correlation between the two rankings and validates our hypothesis, while the Pearson coefficient is evaluated to 0.46833 that further supports our approach.

Figure 4 presents graphically the rearrangement of the news sources in the results of the DeGroot approach compared to the ground truth of the experts' survey. A prominent observation is that the rankings are in agreement about the locations of the sources that are closer to the poles (i.e., the leftmost and rightmost edges of the stripes). Furthermore, the differences in the arrangement of the news sources in the center of the political spectrum (i.e., the middle of the stripes) are noticeable but not extensive.

(a) Stripe produced by experts' survey results

(b) Stripe produced by MinLA algorithm (overlap)

(c) Stripe produced by DeGroot experiment's results (overlap)

Figure 4: Comparison of the three rankings produced by the experts' survey, the MinLA experiment and the DeGroot approach. Stripes (a), (b) and (c) consist of rectangles that visualize the ranking of the news media. Each rectangle represents a news source and is coloured according to the news source's affinity to the two poles of the political spectrum based on the results of the experts' survey (violet and yellow for the left and right pole respectively). The resulting colours are preserved in stripes (b) and (c) for comparison purposes.

5.4. Discussion

The purpose of the news media case study was to examine if our methodology can be applied to the enriched dataset to determine the political affinity of news sources. Overall, we have showed that modularity clustering, the MinLA problem and the DeGroot model are suitable methods. The results were evaluated against the replies of the expert survey and indicate that these methods can be used to study the news sources and determine their political affinity with significant precision.

While the accuracy of the methods in both Section 4 and Section 5 were very high, the evaluation over the MPs was finer and almost flawless. A possible explanation is a fundamental distinction among the MPs and the news sources as Twitter users in the context of our study. In particular, it is an established fact that the dominant act of politicians in Twitter is political deliberation and advocacy and, thus, it is reasonable to assume that other users follow them because of their political standing. In Section 4, we have proved this assertion

with significant accuracy by deriving the political standing of MPs via their followers. However, this does not always seem to be the case for the news sources. News sources convey perspectives of online presence other than politics and, as a result, users may follow them for reasons unrelated to politics, for example sports news. This fact is a form of interference on our methods which rely on the assertion that users follow NOIs for reasons related to the context (politics in this work). Filtering out these users is outside the scope of this study but we believe it would improve the accuracy of the method even more.

Finally, in Figure 4 we provide a visual juxtaposition between the news media rankings of the MinLA algorithm and the DeGroot model for the overlap projection. The two methods achieved similar τ_B evaluation (0.6249 and 0.5879 respectively) against the experts' survey ranking. An apparent similarity of the arrangements is the placement of the left wing (violet). While the two arrangements display similarities, it can be concluded that the MinLA ranking has more matches in the center of the spectrum while the DeGroot ranking has more matches in the right end.

6. Conclusions

The purpose of this work was to study and assess the possibility of deriving political affinity of particular entities (NOIs - Nodes of Interest) using the Twitter follower network. We initially applied our methodology on the Members of the Greek Parliament in order to a) classify them in political parties, b) arrange them in a bipolar spectrum and c) determine their correlation factors with political parties. Our results suggest additional evidence about the validity of the hypothesis that Twitter followers can portray the political leanings of their followees. Our work was later extended on the enriched dataset containing the MPs as well as popular news sources that operate under a political context.

Overall, our approaches are simple to implement and easy to reproduce while delivering very significant accuracy. Furthermore, the overlap coefficient, an underutilized measure, especially in the context of social networks, is high-

lighted and shown to achieve a considerably superior efficiency compared to other projection functions. Additionally, we applied concepts to this problem that have not been examined in prior literature, the MinLA problem and the DeGroot model with the presence of stubborn nodes, and showed that these techniques are perfectly suitable for the analysis of our dataset. We deem that the application of these novel ideas will have a theoretical impact on future research regarding Social Network Analysis. Our methods work without any prior knowledge of the ground truth related to the NOIs and do not employ heavy preprocessing or filtering on the raw data.

We argue that the proposed methodology could be utilized in other practical situations. While the examined scenarios of this work focus on the political attitude of the NOIs, a possible application of the methods could be targeted for other interest domains. For example, the tourism industry is a domain with extensive presence in online social networks. The analysis of online social networks' structure and nodes under the prism of their touristic interest could unveil beneficial aspects of their behaviour and provide valuable findings to tourism organizations.

There are several lines of research arising from this work which should be pursued. A natural extension of this work is the investigation of the scalability of these methods when applied to other countries and, consequently, different political systems. The political scenes among different countries are very diverse and it is interesting to see if the methods are suitable and under which settings. Moreover, the dataset used in this work can also be enriched with temporal information regarding the establishment and possible cancellation of followerships. Therefore, various observations can be resulted from a temporal study during critical points in time, such as elections. Finally, an interesting emerging topic in the field of graph applications is graph embedding [53], where vertices of the graph are represented in vector space. The application of graph embedding in our dataset, and possibly an extensive comparison with the results of this paper, could be addressed in future studies.

Acknowledgements

This work was supported by the project "Assessment of News Reliability in Social Networks of Influence" (Grant no. MIS 5006337) that has been co-financed by the Operational Program "Human Resources Development, Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek National funds. We also thank Chrysanthos Tassis and Costas Eleftheriou from the Department of Social Administration and Political Science, Democritus University of Thrace for their support in running the survey.

References

[1] F. Riquelme, P. Gonzlez-Cantergiani, Measuring user influence on twitter: A survey, Information Processing & Management 52 (5) (2016) 949–975. doi:10.1016/j.ipm.2016.04.003.

[2] A. Giachanou, F. Crestani, Like it or not: A survey of twitter sentiment analysis methods, ACM Computing Surveys 49 (2016) 1–41. doi:10.1145/2938640.

[3] T. M. Nisar, M. Yeung, Twitter as a tool for forecasting stock market movements: Ashort-window event study, The Journal of Finance and Data Science 4 (2) (2018) 101–119. doi:10.1016/j.jfds.2017.11.002.

[4] M. Hasan, M. A. Orgun, R. Schwitter, A survey on real-time event detection from the twitter data stream, Journal of Information Science 44 (4) (2018) 443–463. doi:10.1177/0165551517698564.

[5] M. Dredze, M. Osborne, P. Kambadur, Geolocation for twitter: Timing matters, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1064–1069. doi:10.18653/v1/N16-1122.

[6] J. H. Parmelee, The agenda-building function of political tweets, New Media & Society 16 (3) (2014) 434–450.

[7] R. K. Garrett, Politically motivated reinforcement seeking: Reframing the selective exposure debate, Journal of Communication 59 (4) (2009) 676–699. doi:10.1111/j.1460-2466.2009.01452.x.

[8] P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data, Political Analysis 23 (1) (2015) 76–91. doi:10.1093/pan/mpu011.

[9] M. Gentzkow, J. M. Shapiro, What drives media slant? Evidence from U.S. daily newspapers, Econometrica 78 (1) (2010) 35–71. doi:10.3982/ECTA7195.

[10] S. P. Borgatti, D. S. Halgin, Analyzing affiliation networks, in: J. Scott, P. Carrington (Eds.), The Sage handbook of social network analysis, Vol. 1, SAGE Publications Ltd, London, UK, 2014, Ch. 28, pp. 417–433. doi:10.4135/9781446294413.

[11] M. Vijaymeena, K. Kavitha, A survey on similarity measures in text mining, Machine Learning and Applications: An International Journal 3 (1) (2016) 19–28. doi:10.5121/mlaij.2016.3103.

[12] F. Zarrinkalam, M. Kahani, E. Bagheri, Mining user interests over active topics on social networks, Information Processing & Management 54 (2) (2018) 339–357. doi:10.1016/j.ipm.2017.12.003.

[13] M. Celik, A. S. Dokuz, Discovering socially similar users in social media datasets based on their socially important locations, Information Processing & Management 54 (6) (2018) 1154–1168. doi:10.1016/j.ipm.2018.08.004.

[14] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, in: Pro-

ceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10), 2010, pp. 178–185.

[15] M. Pennacchiotti, A.-M. Popescu, A machine learning approach to Twitter user classification, in: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11), 2011, pp. 281–288.

[16] D. Maynard, A. Funk, Automatic detection of political opinions in tweets, in: R. García-Castro, D. Fensel, G. Antoniou (Eds.), The Semantic Web: ESWC 2011 Workshops, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 88–99. doi:10.1007/978-3-642-25953-1_8.

[17] P. Verweij, Twitter links between politicians and journalists, Journalism Practice 6 (5-6) (2012) 680–691. doi:10.1080/17512786.2012.667272.

[18] A. Boutet, H. Kim, E. Yoneki, What's in Twitter, I know what parties are popular and who you are supporting now!, Social Network Analysis and Mining 3 (4) (2013) 1379–1391. doi:10.1007/s13278-013-0120-1.

[19] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, A. Flammini, Political polarization on twitter, in: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 2011, pp. 89–96.

[20] J. An, M. Cha, K. Gummadi, J. Crowcroft, D. Quercia, Visualizing media bias through Twitter, in: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM '12), Workshop on the Potential of Social Media Tools and Data for Journalists, 2012, pp. 2–5.

[21] H. Le, Z. Shafiq, P. Srinivasan, Scalable news slant measurement using Twitter, in: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17), 2017, pp. 584–587.

[22] A. Tsakalidis, N. Aletras, A. I. Cristea, M. Liakata, Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum,

in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, ACM, New York, NY, USA, 2018, pp. 367–376. doi:10.1145/3269206.3271783.

[23] F. Marozzo, A. Bessi, Analyzing polarization of social media users and news sites during political campaigns, Social Network Analysis and Mining 8 (1) (2017) 1. doi:10.1007/s13278-017-0479-5.

[24] S. Poulakidakos, A. Veneti, Political communication and twitter in greece: Jumps on the bandwagon or an enhancement of the political dialogue?, in: Censorship, Surveillance, and Privacy: Concepts, Methodologies, Tools, and Applications, IGI Global, 2019, pp. 1125–1152.

[25] R. Sainudiin, K. Yogeeswaran, K. Nash, R. Sahioun, Characterizing the twitter network of prominent politicians and splc-defined hate groups in the 2016 us presidential election, Social Network Analysis and Mining 9 (1) (2019) 34. doi:10.1007/s13278-019-0567-9.

[26] A. S. King, F. J. Orlando, D. B. Sparks, Ideological extremity and success in primary elections: Drawing inferences from the twitter network, Social Science Computer Review 34 (4) (2016) 395–415. doi:10.1177/0894439315595483.

[27] J. Golbeck, D. Hansen, A method for computing political preference among Twitter followers, Social Networks 36 (2014) 177–184. doi:10.1016/j.socnet.2013.07.004.

[28] F. N. Ribeiro, L. Henrique, F. Benevenuto, A. Chakraborty, J. Kulshrestha, M. Babaei, K. P. Gummadi, Media bias monitor: Quantifying biases of social media news outlets at large-scale, in: Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM '18), 2018, pp. 290–299.

[29] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, C. Wilson, Measuring personalization of web search, in: Pro-

ceedings of the 22nd International Conference on World Wide Web (WWW '13), ACM, New York, NY, USA, 2013, pp. 527–538. doi:10.1145/2488388.2488435.

[30] H. T. T. Le, Z. Shafiq, P. Srinivasan, Scalable news slant measurement using twitter, in: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17), 2017, pp. 584–587.

[31] J. An, M. Cha, K. Gummadi, J. Crowcroft, D. Quercia, Visualizing media bias through twitter, in: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM '12), 2012, pp. 2–5.

[32] H. Briola, G. Drosatos, G. Stamatelatos, S. Gyftopoulos, P. S. Efraimidis, Privacy leakages about political beliefs through analysis of twitter followers, in: Proceedings of the 22nd Pan-Hellenic Conference on Informatics (PCI '18), ACM, New York, NY, USA, 2018, pp. 16–21. doi:10.1145/3291533.3291557.

[33] T. Zhou, J. Ren, M. c. v. Medo, Y.-C. Zhang, Bipartite network projection and personal recommendation, Physical Review E 76 (2007) 046115. doi:10.1103/PhysRevE.76.046115.

[34] G. Jeh, J. Widom, Simrank: A measure of structural-context similarity, in: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02), ACM, New York, NY, USA, 2002, pp. 538–543. doi:10.1145/775047.775126.

[35] S. Fortunato, Community detection in graphs, Physics Reports 486 (3) (2010) 75–174. doi:10.1016/j.physrep.2009.11.002.

[36] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10) (2008) P10008. doi:10.1088/1742-5468/2008/10/P10008.

[37] A. Lancichinetti, S. Fortunato, Community detection algorithms: A comparative analysis, Physical Review E 80 (2009) 056117. doi:10.1103/PhysRevE.80.056117.

[38] I. Safro, D. Ron, A. Brandt, Graph minimum linear arrangement by multilevel weighted edge contractions, Journal of Algorithms 60 (1) (2006) 24–41. doi:10.1016/j.jalgor.2004.10.004.

[39] U. Feige, J. R. Lee, An improved approximation ratio for the minimum linear arrangement problem, Information Processing Letters 101 (1) (2007) 26–29. doi:10.1016/j.ipl.2006.07.009.

[40] C. Ambuhl, M. Mastrolilli, O. Svensson, Inapproximability results for sparsest cut, optimal linear arrangement, and precedence constrained scheduling, in: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07), 2007, pp. 329–337. doi:10.1109/FOCS.2007.40.

[41] P. Raghavendra, D. Steurer, M. Tulsiani, Reductions between expansion problems, in: Proceedings of the IEEE 27th Conference on Computational Complexity, 2012, pp. 64–73. doi:10.1109/CCC.2012.43.

[42] J. Petit, Experiments on the minimum linear arrangement problem, Journal of Experimental Algorithmics 8 (2003) 1–29. doi:10.1145/996546.996554.

[43] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, P. Raghavan, On compressing social networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, ACM, New York, NY, USA, 2009, pp. 219–228. doi:10.1145/1557019.1557049.

[44] M. H. Degroot, Reaching a consensus, Journal of the American Statistical Association 69 (345) (1974) 118–121. doi:10.1080/01621459.1974.10480137.

[45] J. Ghaderi, R. Srikant, Opinion dynamics in social networks: A local interaction game with stubborn agents, in: American Control Conference (ACC '13), 2013, pp. 1982–1987. doi:10.1109/ACC.2013.6580126.

[46] T. Alzahrani, K. J. Horadam, Community detection in bipartite networks: Algorithms and case studies, in: J. Lü, X. Yu, G. Chen, W. Yu (Eds.), Complex Systems and Networks: Dynamics, Controls and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016, pp. 25–50. doi:10.1007/978-3-662-47824-0_2.

[47] M. Bastian, S. Heymann, M. Jacomy, Gephi: An open source software for exploring and manipulating networks, in: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM '09), 2009, pp. 361–362.

[48] A. Noack, Modularity clustering is force-directed layout, Physical Review E 79 (2009) 026102. doi:10.1103/PhysRevE.79.026102.

[49] A. Agresti, Analysis of ordinal categorical data, 2nd Edition, John Wiley & Sons, 2010.

[50] M. Kwiatkowska, G. Norman, D. Parker, PRISM 4.0: Verification of probabilistic real-time systems, in: G. Gopalakrishnan, S. Qadeer (Eds.), Computer Aided Verification, Vol. 6806 of LNCS, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 585–591. doi:10.1007/978-3-642-22110-1_47.

[51] K. Krippendorff, Content analysis: An introduction to its methodology, Fourth Edition, Sage Publications, Inc., 2018.

[52] T. D. Pigott, A review of methods for missing data, Educational Research and Evaluation 7 (4) (2001) 353–383. doi:10.1076/edre.7.4.353.8937.

[53] P. Goyal, E. Ferrara, Graph embedding techniques, applications, and performance: A survey, Knowledge-Based Systems 151 (2018) 78 – 94. doi:10.1016/j.knosys.2018.03.022.

Appendix A. Projection's Weighting Methods

For the formulas given below, we use the following notation:

Set	Cardinality	Description
N	n	All the followers in the dataset.
N_z	n_z	The followers of NOI z.
N_{xy}	n_{xy}	The common followers of NOIs x and y, $N_{xy} = N_x \cap N_y$.

Since all our weighting methods rely only on the follower sets, it holds that for any weighting method β_G, if $N_x = N_y$, then $\beta_G(x, z) = \beta_G(y, z)$. This property is easy to prove via the following formulas.

Jaccard Index. The Jaccard index of nodes x and y is defined as the intersection of the nodes' follower sets over their union:

$$j_G(x, y) = \frac{|N_x \cap N_y|}{|N_x \cup N_y|}.$$

It has values in $[0, 1]$, with 0 signifying no common follower and 1 an equivalence in the follower sets.

Ochiai Coefficient. The Ochiai coefficient between two NOIs x and y is identical to the cosine similarity when applied to binary vectors (presence or absence of an edge):

$$c_G(x, y) = \frac{n_{xy}}{\sqrt{|N_x||N_y|}}.$$

The Ochiai coefficient can be described as the intersection over the geometric mean and is also a measure lying in $[0, 1]$.

Sorensen-Dice Coefficient. The Sorensen-Dice coefficient is also known as the F1 score and is another statistic used for comparing the similarity of two follower sets:

$$s_G(x, y) = \frac{2n_{xy}}{|N_x| + |N_y|}.$$

It can be shown that there is a relationship between Sorensen-Dice coefficient and the Jaccard index:

$$s_G(x,y) = \frac{2j_G(x,y)}{1+j_G(x,y)}.$$

As in the above methods, the Sorensen-Dice coefficient is in $[0,1]$ and is equal to the intersection over the arithmetic mean of the sets.

Phi Coefficient. The phi coefficient is equivalent to the Pearson correlation coefficient when applied to binary vectors and is formulated as:

$$\phi_G(x,y) = \frac{nn_{xy} - n_x n_y}{\sqrt{n_x n_y (n-n_x)(n-n_y)}}.$$

This measure differs from the other similarity functions as it can be in the range $[-1,1]$, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. In some scenarios, however, a negative weight is either not meaningful or not compatible with the setting at all. In these cases we use two phi coefficient transformations instead that eliminate any negative value:

$$\phi_G^{add}(x,y) = \phi_G(x,y) + 1$$
$$\phi_G^{exp}(x,y) = e^{\phi_G(x,y)}.$$

Overlap Coefficient. The overlap coefficient (also known as Simpson coefficient) is a measure in $[0,1]$ that measures the overlap between two sets. In our context it assumes the value of 1 if the nodes are identical and a value of 0 if they have no common follower. More specifically, it is defined as the size of the intersection divided by the smaller of the cardinalities of the two follower sets:

$$h_G(x,y) = \frac{|N_x \cap N_y|}{min(|N_x|,|N_y|)}.$$

Appendix B. Algorithms

Algorithm 1: Local Search MinLA algorithm

```
Function localMin (a: Array)
    Set n ← size(a)
    for n^2 times do                                   // fast converge
        Perform a random swap on a to create a'
        If it reduces the cost set a ← a'
    end
    while changed do                                   // local converge
        Set changed ← false
        for x in [1, n), y in (x, n] do
            Perform the swap (x, y) on a to create a'
            If it reduces the cost set a ← a' and changed ← true
        end
    end
end

Function main (a: Array, reps: Int)
    for reps times do
        Shuffle a to create a' and invoke localMin(a')
        If the cost of a' is lower than a set a ← a'
    end
end
```

Appendix C. Lemmas

Lemma 1. *The average distance of two nodes in a random LA is* $(n+1)/3$.

Proof. Let X and Y be two random variables for the positions of the two nodes, respectively, in the LA. First, assume $X < Y$. Then, the following sum S_1 is:

$$\sum_{x=1}^{n} \sum_{y=x+1}^{n} P[X = x] \cdot P[Y = y | X = x] \cdot (y - x)$$

$$= \frac{1}{n}\frac{1}{n-1}\sum_{x=1}^{n}\sum_{y=x+1}^{n}(y-x) = \frac{1}{2n(n-1)}\sum_{x=1}^{n}(n-x)(n-x+1)$$

Assuming $X > Y$, the corresponding sum S_2 has the same value $S_2 = S_1$. The average distance is equal to the sum $S_1 + S_2$. Adding S_1 and S_2, and then simplifying gives $(n+1)/3$. ■