Clustering, eigenvector centrality and optimization: An innovative model for network analysis

Emilio Carrizosa[*] Alfredo Marín [†] Mercedes Pelegrín [†]

Abstract

The interest towards key nodes in networks emerged in the past century as a subject of Mathematical Sociology and Graph Theory, and has grown to the point of becoming one of the most noteworthy challenges for understanding social systems. State-of-the-art of network analysis includes discerning the relevance of a group of nodes as network representatives. Many sophisticated approaches based on techniques borrowed from disciplines such as Physics and Statistics have been developed. However, only a few authors use mathematical optimization to address group relevance.

This work presents a mathematical programming formulation for identifying the group of most relevant nodes and their communities. The initial idea was to explore the use of eigenvector centrality, a well-known measure for individual nodes, to spot the group of key members of a network. We realized that real networks usually display a modular structure that emanates from the combination of functional subunits, known by social scientists as *communities*. Aiming at coverage, it appeared natural to assume that targeted key members will belong to different communities. Our approach emerged then as a combination of clustering, which uncovered the communities, and eigenvector centrality, which quantified group relevance. Namely, a representative is chosen for each cluster to be in the group of key members. The network clustering yielding maximum overall relevance of the representatives, which is calculated as an eigenvector centrality, is selected.

1 State of the art

When social networks analysis was still in its infancy, different strategies were explored in order to determine the relevance of a single node. This gave rise to the so-called centrality measures. Some of the classics are based on local criteria such as the number of connections with other nodes of the network (*degree centrality*), the number of shortest paths that contain the node (*betweenness centrality* [1, 2]) or the distance between the node in question and the rest of the nodes in the network (*closeness centrality* [3, 4]). A different approach assumes that one node's importance not only depends on its connections to the rest of the network, but also on the importance of its neighbors. Translating such recursive definition into a mathematical formula yields the search of the eigenvectors of a matrix, named *matrix of relationships* in Sociology and *adjacency matrix* in Graph Theory. The result is a decentralized measure that has been known as *eigenvector centrality* [5, 6] and inspired popular Google's method for rating web pages, PageRank [7]. Among later approaches to measure the relevance of individual nodes, we can find *coreness* [8] or *h-index* [9].

[*] Department of Statistics and Operational Research and Instituto de Matemáticas (IMUS), Universidad de Sevilla, Spain

[†] Department of Statistics and Operational Research, Universidad de Murcia, Spain

State-of-the-art challenges in network analysis include discerning the relevance of a group of nodes [10, 11, 12]. In order to accomplish the new task, a first thought could be to leverage previous knowledge of the problem on a single node. First attempts to approach joint relevance rely upon adaptations of classical degree, closeness and betweenness centralities [13, 14]. Alternatively, they try to identify the minimal set of nodes whose simultaneous removal will fragment the network [13, 15]. In this context, [15] presented an integer linear programming formulation that determines which nodes to remove so that the remaining ones has minimum pair-wise connectivity.

More intricate approaches, coming from very different domains, study influence propagation. In [10] the problem of finding the minimal set of nodes to fragment the network is mapped onto optimal percolation. They developed a heuristic method scalable for big data. Other works use infectious models of Epidemiology to study information spreading through the network [11]. A proportional hazard model is proposed in [16] to shape the spread of behaviors. This survival model served to measure influence and susceptibility on 1.3 million Facebook users. A recent work proposed several integer linear programming formulations for least cost influence propagation [17]. They incorporated activation functions to represent which nodes are reached by influence propagation. Based on the formulations, the authors developed an exact method and also a heuristic algorithm that hinges on column generation.

In recent years, community discovery led by key nodes identification has become a hot research topic in network analysis. Networks topology is rarely homogeneous: real networks usually display a modular structure where one can distinguish different communities. In this paradigm it is assumed that targeted key members will tend to be in different communities [18]. A recent work identifies key nodes in accordance to their relevance and relative dispersion, which are then used as seeds for a k-means clustering [19]. A different strategy also uses the key nodes as seeds but interprets clustering as the result of a non-cooperative game where every node tries to maximize its own social identity [20]. Other approaches use statistical models, such as kernel functions [18, 21] or unbiased random walks [22], to represent relevance and select a number of centers. In a second step, community memberships are determined by evolving a dynamical system where centers memberships are immutable [18, 22]. However, none of these approaches consider joint optimization of the group of key nodes and community discovery.

2 The model

We propose to embed eigenvector computation into a k-centroids clustering framework. In this paradigm for cluster analysis, each group is represented by a centroid and individuals belong to the group with closest centroid. Different sample statistics can be used to determine the clusters representatives, being the mean or median the most widespread ones. As a novelty, our model uses eigenvector centrality instead. This way, centroids are interpreted as the most relevant members within each group, while clusters correspond to their spheres of influence.

As every k-centroids clustering, our approach is made of two main subtasks. One is to determine the centroids and the other is to decide cluster memberships. If we assume that node-cluster memberships are known, eigenvector centrality can be applied within each cluster. The nodes with highest centrality within their clusters, the centroids, will then make the top group for such network partition. Consider now the reverse problem where centroids are given and memberships need to be decided. Unlike in k-means, there is no indication here regarding how nodes should be assigned to clusters. In other words, given the centroids, there is no natural node-cluster assignment that allows to iterate as in Lloyd's k-means. In k-means clustering, each node belongs to the cluster with closest centroid, but here nodes belong to the cluster whose centroid's influence they contribute the most, and this is something that we cannot determine individually for each node. The problem we address here is purely combinatorial:

exploring all the exponentially many partitions of the network and computing the associated centroids would yield the optimal partition and, with it, our group of most relevant members. Mathematical Programming is what allows us to find an optimal solution to this approach without using brute force.

Several interesting questions concerning the field of Operations Research arose when modeling eigenvector centrality with clustering. The first one was to write a mathematical program that we could manage. Naive attempts to model eigenvector centrality over the clusters lead to highly nonlinear programs. Solving them by a sequential linearization approach turned out inoperative. In the end, a more elaborated modeling of eigenvector calculation over the clusters, which includes additional decision variables and constraints, resulted in a mixed-integer linear program for the problem. Variables reduction was investigated for undirected networks. A second difficulty concerned symmetry breaking in integer programming. Symmetry arose from decision variables and constraints to model network clustering. This produced a particular case of set partitioning, a seminal problem in integer programming. Rediscovering the facets of the partitioning orbitope introduced by [23] allowed us to break the symmetry and enhance our formulation.

3 Contributions

The resulting model gives an exact approach to identify the group of most relevant nodes, where relevance is based on eigenvector centrality. Originally thought to give support to eigenvector computation, clustering became a key aspect of our proposal, whose result is twofold. Clustering partition and group relevance optimization interact in a two-directional feedback mechanism to reveal both network modular structure and key members of the network. Network partition directly affects group relevance estimation, while optimizing such results in a more suitable partition in clusters. Computational experiments on real-life and synthetic networks support these statements. Results on *Les Misérables*, Zachary's Karate Club and American Political Book networks reveal previously unnoticed key members. Additionally, clusters are consistent with previous knowledge on the community structure of these networks. Our computational experience on synthetic networks demonstrates an adequate scalability of the method and validates community discovery.

The main contributions of the work can be summarized as follows. First, an innovative model that combines optimization of group relevance and community discovery in the same process is proposed. Second, our exact approach shows the potential of mathematical programming to uncover complex network structures, a context where heuristics abound. Finally, the proposed model serves as a suitable adaptation of widespread PageRank to the problem of group centrality.

References

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

[2] M.E. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27:39–54, 2005.

[3] L.C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1:215–239, 1978.

[4] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581–603, 1966.

[5] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:113–120, 1972.

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

[7] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[8] S.N. Dorogovtsev, A.V. Goltsev, and F.F. Mendes. k-core organization of complex networks. *Physical Review Letters*, 96:040601, 2006.

[9] J.E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the national academy of sciences*, 102:16569–16572, 2005.

[10] F. Morone and H.A. Makse. Influence maximization in complex networks through optimal percolation. *Nature*, 524:65–68, 2015.

[11] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, and H.A. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6:888–893, 2010.

[12] J.X. Zhang, D.B. Chen, Q. Dong, and Z.D. Zhao. Identifying a set of influential spreaders in complex networks. *Scientific Reports-UK*, 6:27823, 2016.

[13] S.P. Borgatti. Identifying sets of key players in a social network. *Computational & Mathematical Organization Theory*, 12:21–34, 2006.

[14] M.G. Everett and S.P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23:181–201, 1999.

[15] A. Arulselvan, C.W. Commander, L. Elefteriadou, and P.M. Pardalos. Detecting critical nodes in sparse graphs. *Computers & Operations Research*, 36:2193–2200, 2009.

[16] S. Aral and D. Walker. Identifying influential and susceptible members of social networks. *Science*, 337:337–341, 2012.

[17] M. Fischetti, M. Kahr, M. Leitner, M. Monaci, and M. Ruthmair. Least cost influence propagation in (social) networks. *Mathematical Programming*, 170:293–325, 2018.

[18] H.J. Li, Z. Bu, A. Li, Z. Liu, and Y. Shi. Fast and accurate mining the community structure: integrating center locating and membership optimization. *IEEE Transactions on Knowledge and Data Engineering*, 28:2349–2362, 2016.

[19] Y. Li, C. Jia, and J. Yu. A parameter-free community detection method based on centrality and dispersion of nodes in complex networks. *Physica A*, 438:321–334, 2015.

[20] Z. Bu, J. Cao, H.J. Li, G. Gao, and H. Tao. Gleam: a graph clustering framework based on potential game optimization for large-scale social networks. *Knowledge and Information Systems*, 55:741–770, 2018.

[21] J. Zhang, K. Zhang, X.K. Xu, K.T. Chi, and M. Small. Seeding the kernels in graphs: Toward multi-resolution community analysis. *New Journal of Physics*, 11:113003, 2009.

[22] A. Stanoev, D. Smilkov, and L. Kocarev. Identifying communities by influence dynamics in social networks. *Physical Review E*, 84:046102, 2011.

[23] V. Kaibel and M. Pfetsch. Packing and partitioning orbitopes. *Mathematical Programming*, 114:1–36, 2008.