PK-Clustering: Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering

Talk submission

Alexis Pister, Paolo Buono, Jean-Daniel Fekete, Catherine Plaisant, Paola Valdivia

Université Paris Saclay, CNRS, Inria and Telecom Paris, France

Abstract. Social scientists often want to cluster the social networks they study to gain information or validate some of their hypotheses. Many clustering algorithms exist but analysts often have trouble choosing which one to use. Indeed, most of the algorithms do not provide any explanations on their output and do not take into account the prior knowledge that the user may have. In this work, we propose PK-clustering, a new approach for creating meaningful clusters in a mixed-initiative setting and which uses the prior knowledge of the user. A partition is constructed iteratively with the system giving suggestions that the user can validate or not. We demonstrate this technique with a prototype.

Keywords: Social network analysis, network visualization, clustering, mixed-initiative, prior knowledge, user interface

1 Motivation

Many social scientists study networks composed of persons of interest. As their data grow in size, the analysis can not remain at a node level and they need to aggregate the persons into groups. This can be accomplished by applying clustering algorithms to the data. However, most of the systems do not provide any guidance on which algorithm to use, even though more than a dozen exist. Furthermore, the Prior Knowledge (PK) that the user has on the data is very often not used in the computation whereas it is precious information in respect of the analysis the social scientist is conducting. The aim of PK-clustering is to address this problem, by allowing the users to build interactively a partition based on their knowledge of the data, the consensus of the algorithms, and the exploration of the graph in a mixed-initiative setting, i.e. where both the system and the user can initiate actions until converging to a satisfactory solution [1].

PK-clustering can be divided into 5 steps: 1) The user specifies the prior knowledge in the form of partial clusters, 2) several clustering algorithms are run on the data, 3) the algorithms are ranked by how well they match the prior knowledge, with a parsimony criterion, and an overview of their results is displayed, 4) the user-selected algorithms are evaluated through their consensus, and 5) the user builds a consolidated partition based on the knowledge on the data and the consensus of the algorithms results, in an interactive framework allowing the exploration of the graph. We implemented a functional prototype of PK-Clustering using the PAOHVis visualization [2], providing a readable visualization of the (hyper-)graph.

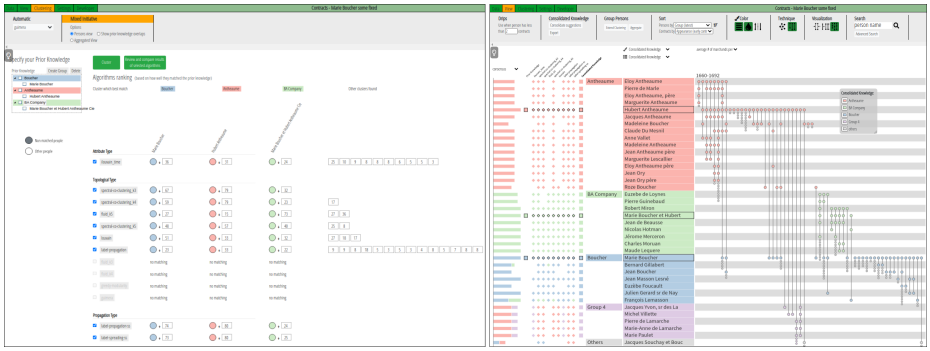

Fig. 1. Two main steps of PK-Clustering: on the left the user specified the PK (top-left), and the algorithms are ranked by how well they match it; on the right, the user created a consolidated partition based on the selected algorithms and the graph view.

2 The process

We describe each step of PK-clustering more in-depth in this section. First, users specify their PK in the form of partial groups. Typically, an expert would assign a few items to a few groups, creating a set of partial clusters (top-left of Figure 1). Then, all implemented clustering algorithms are run on the data. The list of algorithms is shown ranked by how well they match the prior knowledge and a parsimony criterion (left of Figure 1). General information about the results is shown, such as the sizes of the clusters matched with the partial groups of the PK (we call pk-groups), along with the number of other clusters found. Users can then select a set of convincing algorithms.

The second main phase of PK-Clustering is the consolidation phase. Once users selected a set of algorithms, the graph is visualized along with the results of every algorithm in a grid fashion (right side of Figure 1). Each node is represented by a line and each algorithm by a column. A color is attributed to each pk-group. In this setting, users can consolidate the pk-groups in a mixed-initiative setting, using their knowledge of the data, the explorable graph, and the consensus of the algorithms. We call *consolidated knowledge* the final partition which is being constructed and is represented as another column next to the algorithms' results on the right of Figure 1. The system can give suggestions on the labels (represented as circles) based on the consensus of the algorithms and a threshold given by the user with a consensus slider. Moreover, the results of a specific algorithm on one, several, or all nodes can be copied into the consolidated partition by clicking or dragging on the algorithm column. The results must be validated by clicking on the circles on the consolidated knowledge column, turning the circles into squares. Finally, users can at anytime override the cluster of a node by right-clicking on it to manually select a cluster.

Once the pk-groups are consolidated, it is possible that some of the nodes are still unlabeled, meaning other interesting groups may exist in the data. The user can then explore the clusters of any algorithm not corresponding to any pk-group. In that scenario, these other clusters are ranked by their level of consensus among all algorithms. For example, if all the algorithms found a cluster almost identical, it will appear first. Users can validate any of these new clusters and give it a name. A color is then at-

tributed and users are able to consolidate it with the same techniques as with the pk-groups.

Once analysts are satisfied with the consolidated partition, a summary table and a report are produced, giving the results of the algorithms, the consolidated groups and the decision of the user for each node, and statistics on the provenance of the results in the report. This allows analysts to report their results in a transparent manner.

3 Case study

We report here one of the 2 case studies we carried, corresponding to real problems [3]. We gave the system to one of our historian colleagues, who study a dataset of merchants of the XVIIth century, centered around Marie Boucher. She had a hypothesis that the data was composed of 3 groups: the Boucher family, the Antheaume family, and the Boucher & Antheaume company. She tested this hypothesis with PK-Clustering, entering these 3 groups as PK. Nine algorithms produced a perfect match out of the 13 executed. This was a promising result since it means that the majority of the results were consensual with the hypothesis. In the PAOH visualization, she consolidated the 3 pk-groups, mostly using the consensus of the algorithms and the shape of the graph. After this step, there were still several persons not labeled. She decided to review more in-depth the results of the "*ilouvain_time*" algorithm, as this algorithm uses the time attribute in its computation and she thinks it plays an important role in the formation of the groups. She validated a cluster she found relevant and that was consensual among other algorithms results. In the end, the final validated partition was constituted of 4 groups and was satisfactory to the user in regards to her knowledge of the data and the consensus of the algorithms.

4 Discussion and Conclusion

PK-clustering is a new approach for clustering, which combines aspects of automatic, ensemble, and semi-supervised methods in a mixed-initiative setting. Our proposed approach gives a lot of power to the users, as they can override the labels and have to validate the results. Also, if the PK is wrong the system will still try to match it and this could introduce some bias. However, traditional automatic methods often work as black boxes and do not include the users in the analysis loop. This induces situations where analysts must choose an algorithm almost randomly, and compare the results with their PK on their own. We believe PK-clustering solves some of these problems by creating a framework balancing the algorithmic power and the user's knowledge to find meaningful and consensual partitions.

References

1. E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.
2. P. Valdivia, P. Buono, C. Plaisant, N. Dufournaud, and J.-D. Fekete, "Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization," *IEEE TVCG*, vol. 27, pp. 1–13, Jan. 2021.
3. A. Pister, P. Buono, J.-D. Fekete, C. Plaisant, and P. Valdivia, "Integrating Prior Knowledge in Mixed Initiative Social Network Clustering," *IEEE TVCG*, Jan. 2021. to appear.