Encyclopedia of Information Science and Technology, Third Edition

Mehdi Khosrow-Pour
Information Resources Management Association, USA

A volume in the

Information Science REFERENCE
An Imprint of IGI Global

Managing Director: Lindsay Johnston
Production Editor: Jennifer Yoder & Christina Henning
Development Editor: Austin DeMarco & Jan Travers
Acquisitions Editor: Kayla Wolfe
Typesetter: Mike Brehm, John Crodian, Lisandro Gonzalez, Deanna Zombro
Cover Design: Jason Mull

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of information science and technology / Mehdi Khosrow-Pour, editor.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4666-5888-2 (hardcover) -- ISBN 978-1-4666-5889-9 (ebook) -- ISBN 978-1-4666-5891-2 (print & perpetual access) 1. Information science--Encyclopedias. 2. Information technology--Encyclopedias. I. Khosrow-Pour, Mehdi, 1951-
Z1006.E566 2015
020.3--dc23

2014017131

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Clustering Methods for Detecting Communities in Networks

H

Ademir Cristiano Gabardo
Universidade Tecnológica Federal do Paraná - UTFPR, Brazil

Heitor S. Lopes
Universidade Tecnológica Federal do Paraná - UTFPR, Brazil

INTRODUCTION

Real-world networks, such as social networks, enterprise relationships, and the Internet itself, present large amounts of data that can be represented as networks and organized according to some criteria. Such criteria can be, for instance, a measure of similarity, connectivity or a physical distance. In the last years, many efforts have been spent in graph clustering, so as to develop and apply efficient computational methods to group massive data and find communities in networks (Frank, 1996).

As an example we can address social networks. Social networks are groups of individuals or entities that share one or more types of relationships, these relationships can be of various types, common interests, degrees of kinship, shared services, etc. With the popularization of the Internet, there is an increasing number of connected devices, and even more people and organizations are sharing information. Consequently, social networks are becoming ubiquitous (Kumar, Novak, & Tomkins, 2010). Popular social networks such Twitter, Facebook, Google, etc. are widely known by the general public. There is also a large amount of other social related data among the Internet and other networks forming implicit social networks. For instance, citation networks, e-mail traffic, phone users, coworkers, classmates, etc.

Social networks can reveal several aspects of the social behavior of their users, providing relevant information about relationships, identification of influential groups, spread of information, political behavior or even epidemic diseases. The analysis of complex networks has arisen in many areas, such as sociology, communications, computer science, physics and biology.

In this sense, it is relevant to identify clusters, structural communities where a large number of edges join vertices as a cohesive group, a strongly related group of members which can be described as an independent portion of the network or a subgraph.

Usually, methods for detecting communities in large networks are computationally intensive, demanding high processing power. To achieve good clustering results, efficient methods to discover communities in complex networks are needed.

There are several approaches to group the subjects in complex networks. e.g.: Graph Degree Linkage (Zhang et al., 2012), Hierarchical Clustering Algorithms (Murtagh, 1983), Nearest Neighbor Clustering (Ertoz et al., 2002), Partition Algorithms (Fortunato, 2010), etc. Some clustering methods are mathematically formulated to evaluate the connections between vertices of a graph, instead of being focused on similarity measures. The choice of methods depends on which kind of information the social network analyst is pursuing.

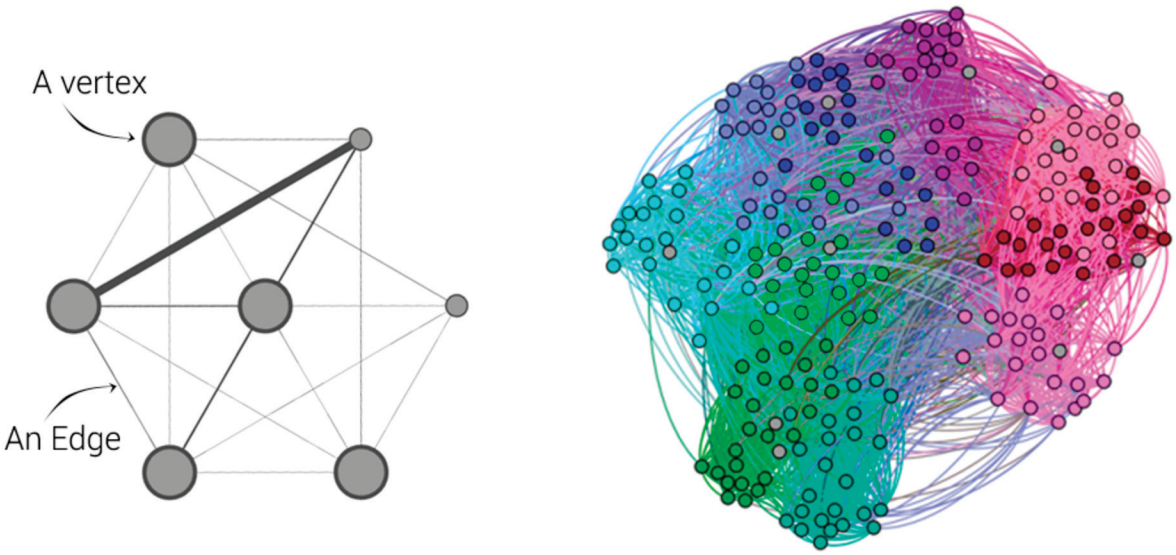
An example is the Girvan and Newman (2002) approach for community detection that focuses on *betweenness*, by removing edges with largest centrality (Freeman, 1977). Another example comprises the Modularity Optimization Methods (Newman, 2006) that uses node degree (how many connections relate to a vertex) as part of the procedure to detect communities.

K-means and its variants, on the other hand, is focused on vertex characteristics. It is more related to data-mining than to community detection, but still can be a powerful tool to group clusters of individuals with high similarity.

This article presents some of the main properties of social networks and complex networks, how the communities and clusters are characterized and the ways used to identify clusters in networks using the

DOI: 10.4018/978-1-4666-5888-2.ch344

Figure 1. (a) An example of a graph, (b) a complex graph



K-means algorithm and its main variants, the fuzzy *c*-means and the weighted *K*-means.

This article is intended for Social Networks analysts, students and researchers in the field of data mining, and for those seeking for agile methods of data arrays in complex networks. Although the K-means algorithm and its variants do not cover the completeness of community detection in complex networks it can be a powerful tool for discovering groups with high similarity. We also show an extended version that weights the data dimensions to be grouped.

BACKGROUND

It is possible to represent a variety of structures by means of complex networks and graphs. A graph is represented as a set of points (vertices) connected by links (edges). More formally, a graph G is an ordered pair of vertices $G = (V,E)$ where: V is a set of nodes (vertices) and E is a set of links (edges). An example of graph is shown in figure 1(a).

Graphs are an abstract mathematical representation of a network. Social networks follow the patterns of complex networks with similar properties. Evolving from purely mathematical models of graphs through the Random Graph of Erdos and Renyi (Erdos & Renyi, 1960) to The Small-World Model of Watts and Strogatz (Watts & Sstrogatz, 1998), complex network analysis have encompassed graph theory and gone so much further to represent real world networks.

A more complex example is shown in the network of figure 1(b). It is from a weighted network of face-to-face proximity between students and teachers. The dataset represents relations of children and teachers from the first to the fifth grade, and it is already grouped in ten clusters. Each cluster represents a particular class of students. This is a good example of how communities can be displayed by a network (Stehlé et al., 2011).

To analyze complex networks, it is mandatory some knowledge about the basic metrics and attributes regarding how the vertices are connected. One of those characteristics is the weight. When connections (edges) between nodes of a graph have weights, it is said a weighted graph. Such weights can represent the strength of a connection or its cost. For instance, to represent a network of cities, the weights could be the distances between them. Both vertices and edges can be weighted. There are also non weighted graphs, in which all connections or vertices receive the same unity value or cost.

Another important property of the connections is the direction; edges can be directed or undirected. In directed graphs, edges have a specific direction, and the relations between pairs of vertices are asymmetric.

In undirected graphs, edges do not have a specific direction, and the relations between pairs of vertices are symmetric.

There are many other metrics related to networks and network components, going deep on mathematical properties and the physical representation of complex networks. However, such subject falls outside the scope of this article and further understanding of the metrics for complex networks is provided by the references cited in the additional readings section.

Characterization of Clusters and Communities

Into the social context we can address a network as groups of people who are related to each other in some way. In social networks people tend to organize themselves according to the same criteria, as in the ordinary life. A network can present several distinct relationships regarding its structure of individuals and the ties between them.

For graph clustering, besides other metrics, to measure the similarity of the network elements is very important. The term "similarity" should be understood as mathematical similarity. These similarity measures will denote how near or far each element of the network is from the others. The concepts of similarity and distance are the roots for clustering algorithms grouping individuals or network vertices.

The key idea of cluster analysis is grouping similar objects. In simple words we can say that a graph cluster is a group of individuals (vertices) with a strong relation (edges) between them and a week relation with the remaining. A graph cluster also can be a subgraph, which is connected or not to other clusters or to the graph itself. In the sense of a complex network, a community is a group of vertices with similarity between the vertices. This measure depends on which attribute is used to build the edges of the network.

The goal of clustering is, given a set of objects, assign them to groups based on their mutual similarity. In social networks we can address clusters as a collection of individuals with dense friendship patterns internally and sparse friendships externally. Vertices assigned to same cluster should be highly similar. Vertices assigned to different clusters should be highly dissimilar.

Network clustering analysis is based on two major levels, macroscopic and microscopic. At a macroscopic level, there are global properties, such the network distance, graph diameter, longest path and shortest path. At the microscopic level, there are properties related to the nodes, mainly degree distribution and clustering coefficient.

Clustering coefficient is a metric used to evaluate the degree to which vertices tend to cluster together. There are two clustering coefficient metrics: global clustering coefficient and local clustering coefficient. The former is based on triplets of nodes and measures the number of closed triplets or triangles. The latter is related to the number of connections to a particular vertex. The proportion between the number of connections to a vertex and the total number of possible connections between the vertex and his neighbors define the clustering coefficient for a particular vertice. It is known that real world networks such as social networks tend to have a clustering coefficient higher than those of random networks (Holland & Leinhardt, 1971).

Community detecting algorithms are strongly attached to the theories of graph partitioning and hierarchical clustering. Many techniques have been employed to clustering data, e.g.: Graph Degree Linkage, Hierarchical Clustering Algorithms, Nearest Neighbor Clustering, Partition Algorithms and many others.

Graph Degree Linkage uses the indegree and outdegree measures to denote clusters in graphs. (Zhang et al., 2012). The vertex degree represents the number of connections to a vertex. In Graph Degree Linkage with directed edges, indegree and outdegree are used to measure the affinity between vertices or between a vertex and a cluster. The communities are identified by iteratively agglomerating new members to an existing cluster based on his similarity with the cluster.

Hierarchical clustering is a method for cluster analysis which builds a hierarchy of clusters. Such as *K*-means, hierarchical clustering is a popular clustering method. Hierarchical clustering generates a binary tree in which the original data items are the leaves, and internal nodes represent clusters of items (Wills, 1998).

Nearest Neighbor Clustering is another approach to find communities in networks. This approach uses the same concept of similarity between vertices such that the vertices are assigned to a "near" cluster. Also, the clusters can "grow," involving a labeled vertex. Similar to *K*-means, the process will occur iteratively until no additional labeling can be done (Jain, Murty, & Flynn, 1999).

Graph partition algorithms can also be used to find communities in networks. A partition is a network division in clusters, where each vertex belongs to one cluster (Fortunato, 2010). A definition for graph partitioning is: given a graph, divide the vertices in sets of equal size, such that the number of the edges between these parts are minimized. This technique is also known as edgecut (Bondy & Murty, 1976). *K*-means is the most popular and the simplest graph partition algorithm.

Each technique or algorithm has strengths and drawbacks, using one or another depends on several factors such as dataset dimension, available hardware, required accuracy, and processing time. We choose *K*-means due to its simplicity, ease to implement, extensive literature, and also because it meets the requirements of a *n*-dimensional data clustering using Euclidean multidimensional metrics with the weight variation.

Historical Background and Applications

One of most famous social network experiment was conducted by the psychologist Stanley Milgram, known as small-world experiment which consists in an attempt to prove that everyone and everything is six or fewer steps away. In this context any ordinary people can reach whoever they want in the world if knows the right six connections, from famous movie actors, politicians or billionaires (Milgram, 1967).

Although Milgram´s experiment is not directly related to clustering analysis or to detect communities in graphs, it drew the attention to social network studies and later inspired many other researches. As an area with a vast field of applications is easy to find related work to social network analysis, such as the study of cohesive subgroups of professionals and the influence exercised from the group one example, addressed by Kenneth Frank (Frank, 1996). Also, (Girvan & Newman, 2002) have conducted experiments with a collaboration network of scientists at the Santa Fe Institute to denote how the members interact and how they contribute as a team. They also conducted experiments with a food web of marine organisms. These are few examples on how community detecting can be applied to unveil a wide range of interactions.

Social Network analysis is often used in marketing. Statistical studies demonstrate that if a group of people has similarities, the overall group shall behavior similarly to a portion of the group. In this sense discovering the communities became relevant.

The *K*-Means Method

The *K*-Means method is numerical, unsupervised, non-deterministic and iterative. In *K*-means, clusters are usually represented as groups of data points with similarities around a center, which is called centroid. It uses the Euclidean distance between vertices to measure their similarity and determine to which cluster a vertex must be assigned. *K*-Means iteratively allocates the partitions of a dataset into *K* clusters, locally minimizing the distance between the vertices to the centroids. (MacQueen, 1967).

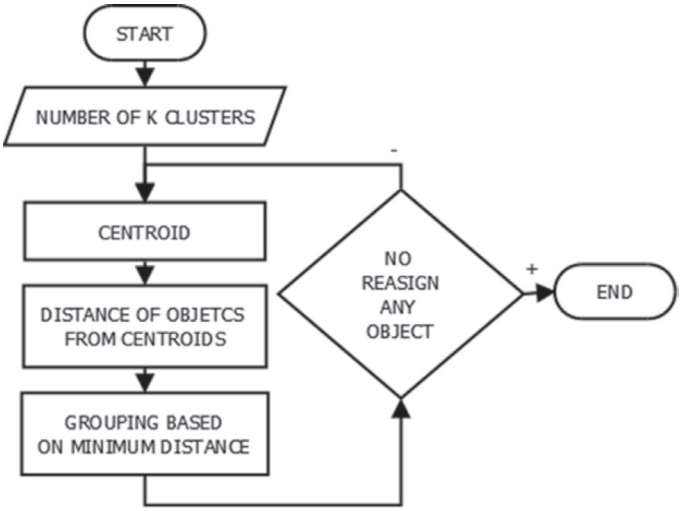
The following steps detail the operation of the algorithm, and a flowchart of the algorithm is illustrated in Figure 2.

1. Given a dataset and the *K* number of desired clusters, the algorithm randomly initializes *K* centroids .
2. Next, the Euclidean distance between each vertex to the centroids is computed. Such a distance is simply the geometric distance in the *n*-dimensional space. Based on this metric the algorithm assigns the vertices to nearest centroid.
3. In the next iteration the centroids are repositioned based on the average distance of all vertices assigned to that centroid.
4. Then, all vertices will be reevaluated again and reassigned to the nearest centroid.
5. Steps 2 to 4 are repeated until the centroids do not change anymore.

Usually, the *K*-means algorithm takes few interactions to discover the position of the *K* centroids and to assign the vertices to each *K* centroid. To compute the *K*-means centroids, it is also possible to use other Minkowiski's metrics such as the Manhattan distance or the Chebychev distance. Equation 1 shows the Euclidean distance formula.

$$d_{xy} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

Figure 2. Flowchart of the K-means algorithm



The objective of K-means is to minimize total intra-cluster variance, or the squared error function, as shown in equation 2, where there are K clusters S_i such that $i = 1, 2, ..., k$, and μ_i is the centroid or mean point of all the points x_j in S_i. Equation 2 shows the K-means squared error function.

$$V = \sum_{i=1}^{k} \sum_{x_j \in s_i} (x_j - \mu_i)^2 \quad (2)$$

In the *K*-means algorithm, every graph vertex has equal importance in locating the centroid of the cluster, this characteristic makes *K*-means very sensitive to outliers, that is, vertices that have values dramatically far from centroids and tend attract the centroid to themselves. The algorithm is also sensitive to the initialization of the centroids especially with very heterogeneous cluster sizes and noisy data. Aiming at improving such drawbacks, some variations of the algorithm have been proposed. These variations will be detailed in next sections.

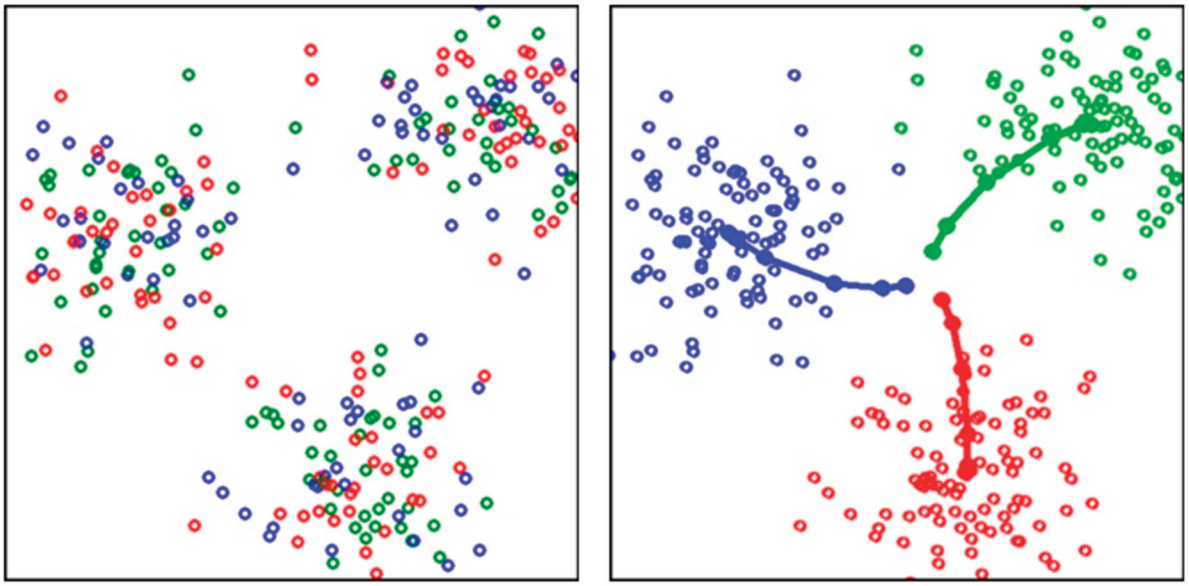
Fuzzy *C*-Means (FCM)

In 1965 Lotfi Zadeh introduced the fuzzy sets, in order to model mathematically the imprecision inherent to the physical world. Fuzzy logic, derived from fuzzy sets, admits ranges of values between the crisp true or false Boolean values. Fuzzy clustering methods followed the creation of fuzzy sets. Such methods allow vertices to be assigned to different clusters in different degrees, consisting of partial memberships. The proposition is that vertices with a high degree of similarity are closer to a cluster, rather than vertices with a low or close to zero degree of similarity to that cluster. Thus, every vertex in the network belongs to all clusters with a distinct degree of membership. Degrees of membership vary from zero to one (Bezdek, Ehrlich, & Full, 1984). When a vertex coincides with the center of the cluster, it is assigned to that vertex the maximum degree of membership. It is possible to blur the boarders of the clusters by using a fuzzyfication constant. Figure 3 (left) shows a sample of points with unclustered data. In the right side of the figure, the same data points are clustered using the FCM algorithm into three distinct clusters.

Different from the usual *K*-means that has a single parameter *K* (number of desired clusters), in order to use FCM the following parameters must be provided: the number of clusters, *c*, the "fuzziness" exponent, *M*, the termination tolerance, and the norm-inducing matrix, *A*. Also, the fuzzy partition matrix, *U*, must be initialized. The number of clusters, *c*, similar to the crisp *K*-means, represents the number of desired partitions and, usually, it is an empirical value.

FCM algorithm is a powerful unsupervised method for the analysis of data and, in several situations, it produces better results than the traditional *K*-means approach, avoiding the local minima. FCM is also less sensitive to noisy data than the *K*-means.

Figure 3. Example of unclustered data sample (left), clustered data with three clusters (right)



Weighted *K*-Means

The analysis of data from interactions commonly occurs in different dimensions, it is reasonable to assume that in certain cases some dimensions may be more relevant (Modha & Spangler, 2003). However, even weak correlations still can have significant value for data analysis. Thus, it is convenient to have an algorithm able to measure attributes under different weights (Huang et al., 2005).

Weighted *K*-means is also called Minkowski Weighted *K*-means (de Amorim & Mirkin, 2012a), since it automatically calculates feature weights for each cluster using the Minkowski metric. Different from the Euclidean distance, Minkowski space also has only one time-like dimension. Weighted *K*-means will output the *K* centroids to a given set of *n* data points considering weights when computing the centroids.

To automate the discovery of weights, different characteristics (dimensions) should have clearly distinct weights. The weights must be non-negative values between zero and one. Thus, the dimensions, uniformly distributed across the clusters, will be assigned with a small weight, while those agglutinated near the centroids will be assigned with a large weight (de Amorim & Mirkin, 2012b). These adjustments result in a balance between the dimensions of the network. This balanced characteristic of weighted *K*-means results into more homogeneous divisions, since no one of the network dimensions will lead the entire network to a specific direction.

We propose a supervised approach to the *K*-means that provides a weight to each dimension of data. With different weights assigned to each dimension of data, it is possible to change the balance of the equation and get an overall modified result. If an attribute has less importance we will counterbalance the others assigning more importance to them. The modified version of the weighted *K*-means with a Euclidean distance subject to weights is shown in equation 3.

$$D = \sqrt{\frac{(p_1 - q_1)2}{W1} + \frac{(p_2 - q_2)2}{W2} + \ldots + \frac{(p_n - q_n)2}{Wn}} \tag{3}$$

Equation 4 shows the Euclidean distance submit to weights in sigma notation.

$$D = \sum_{k=1}^{n} \sqrt{\frac{(pk + qk)^2}{W_k}} \tag{4}$$

The weights will only influence the clustering results if distinct values have been set to at least two

dimensions of data, otherwise the algorithm will behave as the traditional *K*-means.

If there is sufficient change in at least one centroid to reassign at least one vertex to a different cluster, both centroids (which lost the vertex and which receive the vertex) will be recalculated. This procedure changes centroids' positions and may cause new changes to other vertices. The process is repeated until the centroids are mathematically in the center of each cluster.

Therefore, it is possible to balance the dimensions without losses. Even the weak relations can be added to the network with a small weight.

This approach is a complementary way to find communities in graphs with a wide dimensions of data, if distinct dimensions has more or less relevance to the network. This approach minimizes distortions caused by outliers.

As per the previous presented versions the main goal of weighted *K*-means is to identify similarity between vertices or vertices to clusters.

We have successfully combined the weighted *K*-means procedures shown on this article to obtain partitions of social networks and post process the network with Graph Layout algorithms such as Frutchermand and Reingold (1991) and Hu (2005). The *K*-means produced satisfactory partitions of the original data and the direct graph layout algorithms helped to obtain a clear aesthetics network community separation.

Figure 4. shows examples of communities where an index of success was created through the weighted *K*-means algorithm. This network shows the success rate of companies which have participated on public bids. In this case, we have $K=3$ clusters, which correspond to the discretized index of success, in which we have: always win, average win, always lose. There are three distinct groups to be clustered according to several dimensions of data, such as: number of participation in public bids, number of victories, financial value measured, and so on. The most aesthetic result is the Frutchermand and Reingold approach.

FUTURE RESEARCH DIRECTIONS

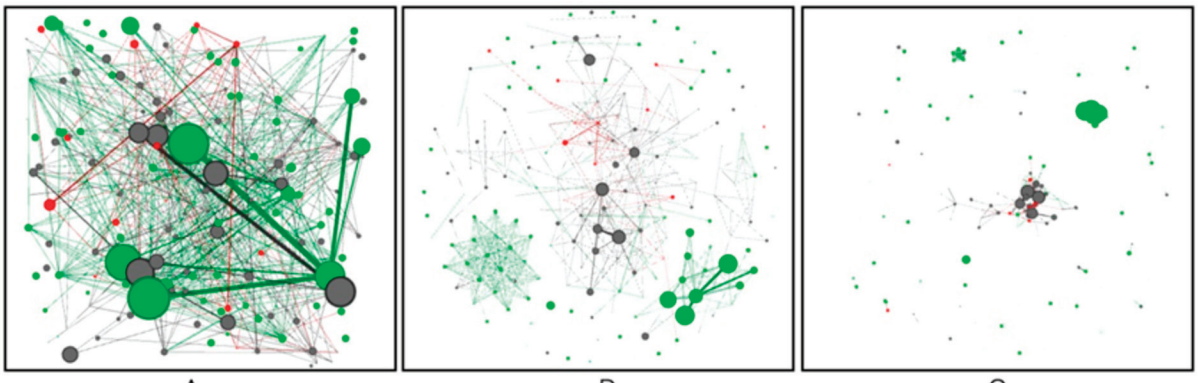
The present work is an attempt to present simple but yet powerful techniques for clustering data, the major intent is to provide a method capable to measure the similarity between elements into a complex network. As future work we can highlight the addition of new metrics to our proposed variation of weighted K-means.

We also intend to do comparison to several other clustering methods by means of benchmarks with well-known complex networks. Thus, expanding the presented work and his results.

CONCLUSION

Network analysis and Sociometrics have received growing attention as the networks become larger and ubiquitous. Therefore, more and more, software, techniques, algorithms and knowledge about them have been developed for its analysis and the extraction of useful information.

Figure 4. Random positioned vertices, Fruchterman and Reingold and Yifan Hu layout examples



In this article we briefly reviewed some known clustering methods for grouping data in complex networks, with focus on the *K*-means and his major variants.

We showed some approaches to find communities in graphs and complex networks using distinct algorithms with their particular strengths and drawbacks. Combining the ability of the *K*-means algorithm to group objects with high similarity and combining the results of the partitions found with algorithms visualization of complex networks we can arrive to a satisfactory performance for real-world massive data.

REFERENCES

Bezdek, J., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191–203. doi:10.1016/0098-3004(84)90020-7

Bondy, J. A., & Murty, U. S. R. (1976). *Graph Theory with Applications*. North-Holland: Elsevier.

de Amorim, R. C., & Komisarczuk, P. (2012). On initializations for the Minkowski weighted k-means. [Springer Berlin Heidelberg]. *Proceedings of the Advances in Intelligent Data Analysis*, *XI*, 45–55.

de Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern Recognition*, *45*(3), 1061–1075. doi:10.1016/j.patcog.2011.08.012

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, *5*, 17–61.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174. doi:10.1016/j.physrep.2009.11.002

Frank, K. A. (1996). Mapping interactions within and between cohesive subgroups. *Social Networks*, *18*(2), 93–119. doi:10.1016/0378-8733(95)00257-X

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41. doi:10.2307/3033543

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software, Practice & Experience*, *21*(11), 1129–1164. doi:10.1002/spe.4380211102

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 7821–7826. doi:10.1073/pnas.122653799 PMID:12060727

Holland, P. W., & Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*, *2*(2), 107–124.

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, *10*(1), 37–71.

Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 657–668. doi:10.1109/TPAMI.2005.95 PMID:15875789

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*, 264–323. doi:10.1145/331499.331504

Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. Link Mining: models, algorithms, and applications(pp. 337-357). New York: Springer.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In L. M. Le Cam & J. Neyman (Eds.). In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*(pp.281-297).

Milgram, S. (1967). The Small World Problem. *Psychology Today*, *1*(1), 61–67.

Modha, D. S., & Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine Learning*, *52*(3), 217–237. doi:10.1023/A:1024016609528

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, *26*(4), 354–359. doi:10.1093/comjnl/26.4.354

Newman, M. E. J. (2006). Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences*(pp.8577-8582).

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., & Pinton, J.-F. et al. (2011). High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE*, *6*(8), e23176. doi:10.1371/journal.pone.0023176 PMID:21858018

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. doi:10.1038/30918 PMID:9623998

Wills, G. J. 1998). An interactive view for hierarchical clustering. In Proceedings of Information Visualization(pp. 26-31).

Zhang, W., Wang, X., Zhao, D., & Tang, X. (2012). Graph Degree Linkage: Agglomerative Clustering on a Directed Graph. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato & C. Schmid (Eds.). In *Proceedings of European Conference on Computer Vision*(pp. 428-441).

ADDITIONAL READING

Adamic, L. A., & Adar, E. (2005). How to Search a Social Network. *Social Networks*, *27*, 2005. doi:10.1016/j.socnet.2005.01.007

Alba, R. D. (1973). A graph-theoretic definition of a sociometric clique. *The Journal of Mathematical Sociology*, *3*, 113–126. doi:10.1080/0022250X.1973.9989826

Barabási, A.-L. (2002). *Linked: The New Science of Networks*. Perseus Publishing.

Barabási, A.-L. (2007). Network Medicine — From Obesity to the "Diseasome. *The New England Journal of Medicine*, *357*(4), 404–407. doi:10.1056/NEJMe078114 PMID:17652657

Costa, L. F., Oliveira Jr, O. N., Travieso, G., Rodrigues, F. A., Villas Boas, P. R., & Antiqueira, L. et al. (2011). Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, *60*(3), 329–412. doi:10.1080/00018732.2011.572452

de Sola Pool, I., & Kochen, M. (1978). Contacts and Influence. *Social Networks*, *1*, 5–51. doi:10.1016/0378-8733(78)90011-4

Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. BookSurge Publishing.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., & Airoldi, E. M. (2010). A Survey of Statistical Network Models. Found. *Trends in Machine Learning*, *2*(2), 129–233. doi:10.1561/2200000005

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, *1*, 201–233. doi:10.2307/202051

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409. doi:10.1073/pnas.98.2.404 PMID:11149952

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256. doi:10.1137/S003614450342480

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *74*(3), 036104. doi:10.1103/PhysRevE.74.036104 PMID:17025705

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *69*(026113).

Newman, M. E. J., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *68*(3), 036122. doi:10.1103/PhysRevE.68.036122 PMID:14524847

Rapoport, A. (1953). Spread of information through a population with socio-structural basis. *The Bulletin of Mathematical Biophysics*, *15*, 523–543. doi:10.1007/BF02476440

Reka, A. & Barabási. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97. doi:10.1103/RevModPhys.74.47

Spearman, C. (1904). The Proof and Measurement of Association Between Two Things. *The American Journal of Psychology*, *15*, 88–103. doi:10.2307/1412159 PMID:3322052

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*(6825), 268–276. doi:10.1038/35065725 PMID:11258382

Travers, J., Milgram, S., Travers, J., & Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*, *32*, 425–443. doi:10.2307/2786545

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press. doi:10.1017/CBO9780511815478

Watts, D., Dodds, P., & Newman, M. (2002). Identity and Search in Social Networks. *Science*, *296*, 1302. doi:10.1126/science.1070120 PMID:12016312

White, H. C., Boorman, S. A., & Breiger, R. L. (1976). Social Structure from Multiple Networks. *American Journal of Sociology*, *81*, 730–780. doi:10.1086/226141

Williams, K., & Durrance, J. C. (2008). Social Networks and Social Capital: Rethinking Theory in Community Informatics. *J. Community Informatics*, *4*(3).

KEY TERMS AND DEFINITIONS

Centroid: The point that inside a geometric shape defines its geometric center.

Cluster: A group of equal or similar elements that occurring closely or together. A sub network divided in groups of nodes with dense connections internally and sparser connections between groups.

Community: A group or set of individuals with similarities into a network.

Complex Networks: Graphs with nontrivial features.

Data Mining: The process of exploring large amounts of data in search of consistent patterns.

Discretization: The act of discretizing, turn something off in several continuous (or discrete) parts.

Fuzzy: An approach based on "intermediate degrees" rather than the "true or false" Boolean logic.

Graph: A graph is represented as a set of points (vertices) connected by lines (edges).

Indegree: The number of the edges directed into a vertex in a directed graph.

K-Means: A clustering algorithm to partition n observations into K clusters, which each K cluster has a centroid and each observation is assigned to a cluster.

Outdegree: The number of the edges directed out a vertex in a directed graph.

Social Networks: A social structure such individuals or organizations and the relationship between them.