

A simple and fast exact clustering algorithm defined for complex networks and based on the properties of primes

Nicola Serra
Institute of Radiology, Faculty of Medicine and Surgery, Second University of Naples, Italy
E-mail of the corresponding author: nicola.serra@unina2.it

Abstract

In this paper a new clustering method based on primes is proposed. This method define a nodes cluster of any complex network, considering the nodes with same input/output number and same number of paths with equal length, so all the network nodes with analogous functions will be possible to identify. The clustering algorithm proposed, results very efficient because it is defined on simple computations with primes. For example, with our algorithm the analysis of a network with 500 nodes and 124750 connections is performed in 80 seconds on Pentium 4 with CPU 2Ghz and 1Gb ram.

Keywords: Complex network, clustering method**,** graph theory, bidirectional network, complete path.

1. Introduction

In the last years the networks applications are frequently been used to study the complex systems. There are many types of complex network: Information Networks, Technological Networks, Biological Networks, Social Network etc, (M. E. J. Newman 2003, S. Brohée et al. 2008) but essentially they are all characterized by nodes and connections among the nodes. The nodes or vertexes are objects, with one or more input and one or more output, instead the connections define the information flow among the nodes, which can be bidirectional or unidirectional. The graph theory (F. Harary 1995, Bornholdt & Schuster 2002, Bondi J.A. & Murty U.S.R 2008) is the appropriate tool in the study and representation of complex networks. A bidirectional (unidirectional) network is represented with an *undirected* (*directed*) *graph* $G = (V, E)$ that consists of two sets V and E, such that $V \neq \varnothing$ and E is a set of unordered (ordered) pairs of elements of V. The elements of $V \equiv \{N_1, N_2, \ldots, N_h\}$ are the *nodes* (or *vertices*, or *points*) of the graph G, instead the elements of $E \equiv \{e_1, e_2, \ldots, e_k\}$ are its *connections* (or *links*, or *lines*, *or edges*). According with graph theory we consider a *path* from node i to node j is an alternating sequence of adjacent nodes and edges that begins with i and ends with j, in which each node is considered only once. Since the network nodes may represent proteins, cells, computers, web pages, individuals or animals etc, the individuation of nodes with analogous functions is important, so a simple methodology to define a exact clustering method for any bidirectional network, is proposed. This paper is organized as follows: in Section 2 we introduce the definitions of complete path and complete paths set, that are fundamentals to define the clustering algorithm. In Section 3, the new clustering method is described. In Section 4 there are the conclusions for this paper.

2. Complete Paths

We consider in Figure 1, the bidirectional complex network represented by the graph $G = (V,E)$, with $V \equiv \{A, B, C, D\}$ and unordered set $E \equiv \{AB, AD, BC, BD, CD\}$. We give the following definitions:

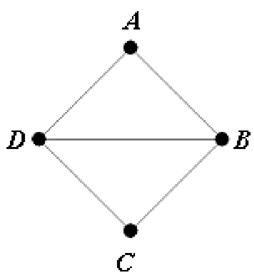

Figure 1. *Bidirectional Complex Network*

Definition 2.1. A *complete path* (*CP*) for a network node is a possible path which starts from the node and include all the possible nodes network only once in a direction.

Definition 2.2. Two *complete paths* are *equal* if they will have same length or number of nodes, same nodes and same consecutive nodes without considering the verse.

Definition 2.3. Two *complete paths* are *similar* if they will have same length or nodes number, same nodes but they haven't same consecutive nodes, without considering the verse.

Example 2.1. The complete path *ABCD* (Figure 1) is equal to: *BCDA*, *CDAB*, *DABC*, and *DCBA*, *CBAD*, *BADC*, ADCB, instead it is similar to *BACD*.

Definition 2.4. The *complete paths set* (*CPS*) of a node *V* is the set $\wp_V = \{p_1, p_2, \ldots\}$ of all *complete paths* p_i of a network, which start from *V*.

Remark 2.1. We want underline that a maximal path is complete path too, but no vice versa is true. For example we consider the path that start from *B* and ended to *C* (Figure 1). The maximal path is *BADC* instead the complete path are: *BADC*, *BDC*.

Example 2.2. In Figure 2 we show the complete paths set $\wp_B$={*BADC*, *BCDA*, *BDC*, *BDA*} of the network of Figure 1,

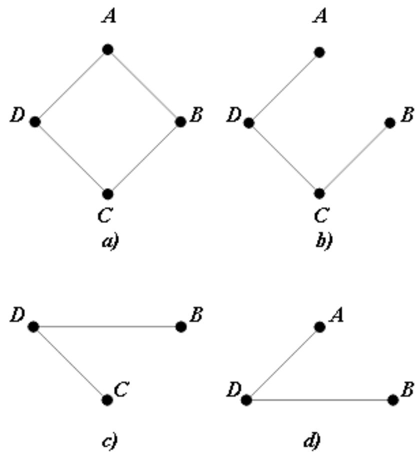

Figure 2: *Complete paths set of the node B of the bidirectional complex network in Figure* 1.

Considered $\wp_H$ and $\wp_K$ two *CPS* for *H* and *K*, if we indicate with $|\wp_H|$ and $|\wp_K|$ the number of *complete paths* of $\wp_H$ and $\wp_K$ respectively, we have the following definition:

Definition 2.5. $\wp_H$ and $\wp_K$ will be said *equal*, if $|\wp_H| = |\wp_K|$ and $\forall p_i$ (*CP*), if $p_i \in \wp_H$, than there is a $p_j \in \wp_K$ such a that p_j and p_i are equal and vice versa.

Example 2.3. We consider the *CPS* $\wp_B$ and $\wp_D$ of the nodes *B* and *D* (Figure 1),

$$\wp_B = \{BCDA, BACD, BDA, BDC\}$$

$$\wp_D = \{DABC, DCBA, DBC, DCA\}$$

for definition 2.5, they are *equal*.

Definition 2.6. Considered $\wp_H$ and $\wp_K$ the *CPS* of *H* and *K*, they will be said *similar*, if $|\wp_H| = |\wp_K|$ and $\forall p_i$ (*CP*), if $p_i \in \wp_H$, than there is $p_j \in \wp_K$ such a that p_j and p_i are similar and vice versa.

Remark 2.2. Two equal *CPS* are similar too, but two similar *CPS* are not equal.

Definition 2.7. Considered $\wp_H$ and $\wp_K$ the *CPS* of *H* and *K*, they will be said *almost-similar*, if they will have in common only some *CP* equal/similar.

Example 2.4. We consider the network in Figure 1 and the *CPS* $\wp_A$ and $\wp_B$ of the nodes *A* and *B* respectively:

$$\wp_A = \{ABCD, ADCB, ABDC, ADBC\}$$

$$\wp_B = \{BCDA, BACD, BDA, BDC\}$$

in this case, we observe that the first and second $\wp_A$-*CP* with the first $\wp_B$-*CP* are equals and the second $\wp_B$-*CP* is similar with the first and second $\wp_A$-*CP*, besides the set $\{ABDC, ADBC\} \not\subset \wp_B$ and the set $\{BDC, BDA\} \not\subset \wp_A$, therefore $\wp_A$ and $\wp_B$ are *almost-similar*.

3. Clustering Method

The *clustering* is a method that defines a number *k* of subsets separated of a initial set, according with an opportune measure of likelihood or similarity. Such operation can be used for classifying the kind, to reassume results or still to analyze images, etc. Therefore the clustering represents in many cases an essential operation of statistic analysis of the experimental data. Further advances in this field are given by Anderberg, M. R. 1973, Hartigan, J. 1975, Dubes, R. C. et al. 1988, White D.R. 2001, Brades U. et al. 2003, Shamir R et al. 2004, Schaeffer S.E. 2007, Xing B. et al. 2007, Xu X. et al. 2007, Kim C. et al. 2008, Tan, L. et al. 2009, Gorke R. et al. 2010, Ahmed A. et al. 2012. The mathematical language that we have introduced, allows us to define a clustering algorithm. This algorithm divide in classes the nodes of a bidirectional complex network, according their *CPS*. To proceed to the clustering of the network nodes, for each *CPS* of a node will be encoded with an univocal integer. We will call this integer, *CPS-code*. Therefore it is possible to define the clusters for the network nodes, with equal *CPS-code* (or *CPS-code* of the same order of greatness), codifying the length of every *CP* of a node with a prime number (*CP-code)*, according to the associations in Table 1.

Table 1. *CP-code for bidirectional complex network*.

$N_1N_2 = 2$;	*CP-code* with two nodes
$N_1N_2N_3 = 3$;	*CP-code* with three nodes
$N_1N_2N_3N_4 = 5$	*CP-code* with four nodes
$N_1N_2N_3N_4N_5 = 7$	…………………………………..
$N_1N_2N_3N_4N_5N_6 = 11$;	
$N_1N_2N_3N_4N_5N_6N_7 = 13$;	
$N_1N_2N_3N_4N_5N_6N_7N_8 = 17$;	
$N_1N_2N_3N_4N_5N_6N_7N_8N_9 = 19$;	
$N_1N_2N_3N_4N_5N_6N_7N_8N_9N_{10} = 23$	*CP-code* with ten nodes
………………………………..	…………………………………..

with N_1, N_2, … the nodes of the generic *CP*. For example codifying all the *CP* of the node *A* (Figure 1), we will have:

$$\wp_A = \{ABCD, ADCB, ABDC, ADBC\}$$

$CP-code(ABCD) = 5$, $CP-code(ADCB) = 5$, $CP-code(ABDC) = 5$, $CP-code(ADBC) = 5$, i.e. we have four common *CP* of same length. so the $\wp_A$ *-code* will be defined by:

$$\wp_A\text{-}code = \prod_{i=1}^{n=4} CP-code(p_i) = 5\times 5\times 5\times 5 = 5^4 \qquad 3.1$$

where n is the number of the *CP* associated to the node *A*. The integer $\wp_A$ *-code* will individualize a nodes cluster. Analogous for the nodes *B*, *C* and *D*, we have: $\wp_B = \{BCDA, BACD, BDA, BDC\}$, $\wp_C = \{CBAD, CBDA, CDAB, CDBA\}$, $\wp_D = \{DABC, DCBA, DBC, DCA\}$ with $\wp_B$ -code = $3^2 \times 5^2$, $\wp_C$ -code = 5^4, $\wp_D$ -code = $3^2 \times 5^2$.

Remark 3.1. We have codified the *CP* with a prime number, because this defines an univocal *CPS-code* for a generic network node.

Definition 3.1. We will define a *node cluster* ς_A the nodes set of a network, with same $\wp_A$ *-code*, so $\varsigma_A = \{A, C\}$.

Remark 3.2. We observe that, if two nodes belong to the same complex network (Figure 1), it doesn't implicate that they belong to the same class, because each node class is defined by the *CPS-code*. Particularly two similar *CPS* define two different clusters in a unidirectional networks and same clusters in a bidirectional network.

If we consider two *CPS*, $\wp_H$ and $\wp_K$ with $\wp_K \subset \wp_H$ i.e. if $\wp_K$ -code is a integer divisor of $\wp_H$ -code, than ζ_K is a subcluster of ζ_H, therefore it is possible to individualize all *subclusters* of the node cluster *H*, considering all the nodes with *CPS-code* integer divisor of ζ_H -code. We observe that, if the Greatest Common Divisor (*GCD*) between two *CPS-codes* associated to ζ_H and ζ_K, is equal to *CPS-code* of ζ_K, then ζ_K is a *subcluster* of ζ_H.

Remark 3.3. If we consider *GCD* between $\wp_A$ -code and $\wp_B$ -code (Figure 1): GCD(5^4, $3^2 \times 5^2$) = 5^2, then the *GCD factors*, define the common number of *complete paths* between $\wp_A$ and $\wp_B$ with *same length*. Now, if we divide $\wp_A$-code with *GCD*($\wp_A$, $\wp_B$), we have a integer rest. If we factorize this rest, the primes define the length of the *complete paths* of $\wp_A$ not belonging to $\wp_B$.

We can verified that the nodes network *A* and *B* are *almost-similar* (Figure 1, *Example* 2.4), using *CPS-code.* In fact $\wp_A$ *-code* = 5^4, $\wp_B$ *-code* = $3^2 \times 5^2$ and the *GCD*($\wp_A$ *-code*, $\wp_B$ *-code*) = 5^2, i.e. $\wp_A$ and $\wp_B$ have two common *CP* of length 5, therefore $\wp_A$ and $\wp_B$ are *almost-similar* and the nodes *A* and *B* define two different node *clusters*, ζ_A and ζ_B for the bidirectional network in Figure 1.

Finally in Table 2, we have considered 10 bidirectional networks with different number of nodes and of connections and evaluated the networks analysis times, with our clustering algorithm, using a Pentium 4 with CPU 2Ghz and 1Gb ram.

Table 2. *Algorithm performance on different bidirectional networks.*

Nr. Nodes	**Nr. Connections**	**Time (sec)**
10	45	≈ 0.05
20	190	≈ 0.05
40	780	≈ 0.05
80	3160	≈ 0.22
100	4950	≈ 0.35
150	11175	≈ 0.95
200	19900	≈ 3.00
300	44850	≈ 13.00
400	79800	≈ 36.00
500	124750	≈ 80.00

5. Conclusion

The clustering method proposed for the analysis of the bidirectional complex networks, is a fast method that could be applied in different scientific sectors. The analysis is founded on idea to gather in classes the network nodes or more networks with same number of connections and paths with equal length. Each class of nodes is characterized by a code obtained codifying each complete path (*CP*) with a prime number. The simplicity of the clustering algorithm introduced, gives a strong computational implementation and therefore of great applicability to every network type. Finally, we want underline that by generic *CPS-code,* it is possible to individualize the nodes with more connections, i.e. the nodes with more input/output and the root nodes with paths more long. To start these information, this clustering method could be used as routing algorithm in the telecommunication networks too. Finally the object of future paper will be develop the present research to describe the clustering method based on primes for unidirectional complex networks.

References

H. A. Ahmed, P. Mahanta, D. K. Bhattacharyya, J. K. Kalita, *Autotuned Multilevel Clustering of Gene Expression Data*, American Journal of Bioinformatics Research 2012, 2(5): 68-78.

Anderberg, M. R. (1973), *Cluster Analysis for Applications*. Academic Press, New York, NY.

Brades U.,Gaertler M., Andwagner, D. 2003. *Experiments on graph clustering algorithms*. In Proceedings of the 11th Annual European Symposium on Algorithms *(ESA '03)*. Springer, 568–579.

Bornholdt, S., Schuster, H. G., eds. (2002). *Handbook of Graphs and Networks – From Genome to the Internet,* Berlin: Wiley – VCH

Bondi J.A. & Murty U.S.R., Graduate Texts in Mathematics series, Graph Theory, Springer, 2008

S. Brohée, K. Faust, G. Lima-Mendez, G. Vanderstocken & J. van Helden, *Network Analysis Tools: from biological networks to clusters and pathways,* Nature Protocols ***3****, 1616 - 1629 (2008)*

Dubes, R. C., & Jain, A. K., (1988), *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ.

Gorke R., Gaertler, M., Hubner, F., AND Wagner D. 2010. Computational aspects of lucidity-driven graph clustering. *J. Graph Algor. Appl. 14*, 2, 165–197

F. Harary, *Graph Theory*, Perseus, Cambridge, MA, 1995

Hartigan, J. (1975) Clustering Algorithms. Wiley, New York, NY.

Kim, C., Cheon, M., Kang, M., & Chang, I. (2008). A simple and exact Laplacian clustering of complex networking phenomena: Application to gene expression profiles. *Proceedings of the National Academy of*

Sciences, *105*(11), 4083-4087.

M. E. J. Newman, *The Structure and Function of Complex Networks;* SIAM REVIEW Vol. 45,No . 2,pp . 167–256, 2003.

S. E. Schaeffer, *Graph clustering,* Computer Science Review, vol. 1, no. 1, pp. 27–64, 2007.

R. Shamir, R. Sharan, and D. Tsur, *Cluster graph modification problems*, Discrete Applied Mathematics, vol. 144, no. 1-2, pp. 173–182, 2004.

Tan, L., Zhang, J., & Jiang, L. (2009). An evolving model of undirected networks based on microscopic biological interaction systems. *Journal of biological physics*, *35*(2), 197-207.

White, Douglas R. and Newman, Mark, *Fast Approximation Algorithms for Finding Node-Independent Paths in Networks* (June 29, 2001). Santa Fe Institute Working Papers Series. Available at SSRN: http://ssrn.com/abstract=1831790.

Xing B, Greenwood C.M. and Bull S.B., *A hierarchical clustering method for estimating copy number variation*, Biostatistics. 2007 Jul;8(3):632-53. Epub 2006 Oct 23

X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. *Scan: a structural clustering algorithm for networks*. In KDD pages 824–833, 2007.