

Local Community Detection of High Density: an Upper Bound for the Optimal Solution

Patricia CONDE CESPEDES

LISITE-ISEP, 10 rue de Vanves, 92130 Issy-les-Moulineaux, Paris

Tel.: (+33) 01 49 54 52 41

E-mail: patricia.conde-cespedes@isep.fr

Received: 15 May 2019 /Accepted: 15 June 2019 /Published: 30 June 2019

Abstract: Community detection in complex networks has attracted so much attention in the last years. Usually, community detection is referred to the problem of partitioning an entire network. In contrast, local community detection aims to detect the community of a given node in the network. This can be useful when we do not have information concerning the entire network or when there is a specific node of interest in the network. In this paper, we focus on the problem of detection of local communities of very high density. Communities of maximal density are called complete cliques in graph theory. In real complex networks, whose degree distribution follows a power law, usually complete cliques are small sets of nodes. This led to the problem of finding quasi-cliques of maximal size. This problem is NP-hard. Some heuristics on the optimal solution were recently proposed. In this paper, we propose an algorithm to calculate an upper bound on the optimal solution in order to evaluate the existing heuristics. The proposed upper bound will be evaluated experimentally on real networks.

Keywords: Local Community detection, α -quasi-clique, Density, Maximal α -quasi-clique problem, Upper bound.

1. Introduction

Over the past decade, the study of complex networks (network-based representations of complex systems) has taken the sciences by storm. Due to various factors, such as globalization, the Internet, social networks, etc. Researchers from biology to physics, from economics to mathematics, and from computer science to sociology, are more and more involved with the collection, modeling and analysis of network-indexed data.

A network is usually described by a set of entities, called vertices or nodes, connected by links, also called edges. Real complex networks share important characteristics (degree distribution, local clustering) and often exhibit community structures. The study of the communities has attracted a lot of attention (see [1]). Detecting communities in large complex

networks is important to understand their structure and allows to extract features useful for visualization or prediction of various phenomena like the diffusion of information or for social recommendation.

A community is usually referred to as a set of strongly interconnected nodes. The density of links measures the strength of the relationships in the community. However, many community detection methods do not guarantee anything about the density of the resulting communities (for instance, when using the Newman-Girvan modularity [2], the density of the output communities can become very low due to its resolution limit [3]). A complete clique is a set of nodes where every two distinct nodes are connected to each other. One can easily deduce that complete cliques are communities of maximal density. However, the size of a clique is limited by the degree of its nodes. Most real complex networks' degree

distribution follows a power law, then cliques can be very small or even trivial, such as pair of nodes or triangles. This led to the relaxation of the concept of a complete clique to an *almost complete subgraph*, also called *quasi-clique*. Therefore, we focus on the problem of finding *quasi-cliques of maximal size*.¹

To give a concrete definition of a quasi-clique, we consider the concept of an α -quasi-clique (for a given α , such that $0 < \alpha < 1$). Then, we define, an α -quasi-clique as a group of nodes where each member is connected to more than a proportion α of the other nodes. Consequently, an α -quasi-clique has a density greater than α . By choosing α equal to 1, an α -quasi-clique becomes a complete clique. Considering an α -quasi-clique instead of a complete clique can be preferable for applications where interaction between members of the community does not need to be direct and could be successfully accomplished through intermediaries. Mining all the maximal α -quasi-cliques of a network is NP-complete (see [8] and [9]). Efficient exact methods or approximations to solve it are available. However, all these methods generally assume that the network is entirely known and they try to find all existing α -quasi-cliques.

In some particular applications, the network can be so large that we do not have access to information concerning the entire network. Furthermore, one can be only interested in the community of a particular node in the network. Moreover, detecting the local communities of specific nodes may be very important for applications dealing with huge networks, when iterating through all nodes would be impractical or when the network is not entirely known. The detection of the community of a given node of interest is called *local community detection problem*.

The problem of finding local communities of maximal size of type α -quasi-cliques was first tackled by [10] [Conde et al. (2018)] (one can also see a preliminary version on [11] [Conde et al. (2015)]). The proposed algorithm, denoted *RNN* (Rank on the Number of Neighbors). This algorithm, like most algorithms of community detection, is a heuristic. Usually heuristics are evaluated by comparing the obtained results to existing methods. However, there is no a concrete measure on how close the obtained solution is to the optimal one.

In this paper, we propose an upper bound for the optimal solution of the so-called *Maximal α -quasi-clique community of a given node*. This upper bound will be useful to evaluate the experimental results obtained by the *RNN* algorithm.

This paper is organized as follows: Section 2 presents the main definitions (notations and problem formulation). Next, we discuss about the proposed upper bound in Section 3. Then, Section 4 presents the results obtained by comparing the upper bound with the results obtained by the *RNN* algorithm. Finally, Section 5 draws some conclusions and perspectives.

2. Main Definitions

A graph, denoted $G = (V, E)$ is defined by V the set of vertices or nodes, and E the set of edges or links, formed by pairs of vertices. To simplify we consider undirected graphs, where edges are not oriented. The neighborhood $\Gamma(u)$ of a node u is the set of nodes v such that $(u, v) \in E$. The degree of a node u , denoted $d(u)$, is the number of its neighbors, i.e. $d(u) = |\Gamma(u)|$. Considering all these notations, an α -quasi-clique is defined as follows:

Definition 1: α -quasi-clique

Given an undirected graph $G(V, E)$, and a parameter α with $0 < \alpha < 1$, an α -quasi-clique is the subgraph induced by a subset of the node set $C \subseteq V$ if the following condition holds:

$$|\Gamma(n) \cap C| > \alpha(|C| - 1), \quad \forall n \in C, \quad (1)$$

where the symbols $|S|$ denotes the cardinality of the set S .

Eq. (1) implies that each node in the quasi-clique C must be connected to more than a proportion α of the other nodes. Notice that for $\alpha=1$ an α -quasi-clique is a complete clique. Then, when choosing a high value for α the resulting communities are robust, contain strongly connected nodes and have an edge-density greater than α (see [10] [Conde et al. (2018)] for the proof).

In the following, we will call Eq. (1) the rule of an α -quasi-clique. This rule constitutes a lower bound on the minimal internal connections of each node.

In the literature, one can find other definitions of the so-called α -quasi-clique. The most common variant, considered by [12-15], is just a relaxation of Definition 1 as it just is just constraints the global density of the quasi-clique to be at least α . Other variants much closer to Definition 1 are considered by [16] and [17]. However, these latter allow the equality in Eq. (1) which implies that each node might have as many connections as non-connections in the community whereas Definition 1 requires the absolute majority. For all these reasons, in this study, we consider Definition 1 as it guarantees that the detected communities are robust, contain strongly connected nodes.

2.1. Problem Formulation

The size of an α -quasi-clique is limited by the degree of its nodes. Most real complex networks' degree distribution follows a power law. Therefore, mining for the α -quasi-clique community of specific nodes

¹ This is a well-known problem in graph theory. The interested reader can see [4-7].

with low degree can lead to trivial solutions, such as pairs of nodes or triangles. Such trivial communities are not interesting for applications. Therefore, the purpose is to find quasi-cliques of *maximal size*.

In the literature, one can find several methods for detecting the maximal α -quasi-clique of an entire network. In this paper, we consider the problem of finding an α -quasi-clique of maximal cardinality containing a given node. This set of nodes will be the local community of that node with high density for α high. This problem can be formulated as follows:

Problem 1: The maximal α -quasi-clique local community problem

Given a node n_0 of a graph $G(V, E)$ and a parameter α ($0 < \alpha < 1$), the purpose is to find the largest α -quasi-clique, denoted $C(n_0)$ containing n_0 , mathematically:

$$\begin{aligned} & \text{maximize } |C| \\ & \text{subject to } n_0 \in C \\ & \text{and } |\Gamma(n_i) \cap C| > \alpha(|C| - 1), \quad \forall n_i \in C \end{aligned} \quad (2)$$

Problem 1 is NP-complete (see [8] and [9]). Therefore, only heuristics that run in reasonable amount of time can be proposed. We will discuss these methods in the Section Related works.

The Problem 1 can have multiple solutions. Indeed, a node can belong to more than one maximal α -quasi-clique community.

2.2. Related Works

The problem of finding the maximal α -quasi-clique local community was addressed by [11] and later improved by [10]. In those studies, the authors proposed a method denoted *RANK-NUM-NEIGHS* (*RNN*). The *RNN* method is a greedy and iterative algorithm. The resulting local community is denoted $C(n_0)$. At first iteration $C(n_0)$ is composed of only one node n_0 . Then, iteratively, a set of nodes from the neighborhood of the community $\Gamma(C(n_0))$ is chosen to become a member of $C(n_0)$ provided that the new nodes will satisfied the rule of an α -alpha-quasi-clique given in Eq. (1). The choice of nodes is based on the number of common neighbors between with the local community $C(n_0)$. Indeed, the *RNN* algorithm establish a rank according to the number of neighbors for all possible set of nodes in the neighborhood of $C(n_0)$. That is where the name *RNN* comes from. For more details, the interested reader can see [10].

Other heuristics exist that deal with similar problems to Problem 1. However, either they try to find quasi-cliques of maximal size subgraph mining the whole graph with no reference to a given node (see for instance the *QUICK* method de [17], the *RLS-DLS* method de [16], [Conde-Céspedes et al. 2016] or even the *Louvain* method adapted to *Zahn-Condorcet* and *Owsinski-Zadrozny* criteria, also called studied by [Conde-Céspedes et al. 2015], [23]) or they do not

constraint the output communities to be α -quasi-cliques (see for instance [18-20]).

For all these reasons, we will consider the results obtained only with the *RNN* algorithm in the experimental results. First of all, we will define the algorithm to calculate an upper bound for the optimal solution of Problem 1 and then, compare the difference with the results obtained by the *RNN* algorithm. Some preliminary results were presented in [24].

3. Calculation of the Bound

It is well known that community detection is an NP-hard problem. Problem 1 is not an exception. Then, many authors usually propose heuristics and evaluate their proposals by comparing the obtained results to those obtained by already existing heuristics. Furthermore, heuristics might lack of stability. That is, it is necessary to execute the algorithm several times to obtain reliable results. Calculating a bound on the optimal solution of an NP-hard problem allows to evaluate a heuristic in an objective way. Considering all these reasons, in this section we propose to approach, the value of the optimal solution of Problem 1 in terms of community size. Since it is a maximization problem we deduce an upper bound.

In the following let us denote $C^*(n_0)$ the community that optimizes problem 1 for a given node n_0 . Then, an optimal solution for Problem 1 will be $|C^*(n_0)|$. Then, the following theorem holds (see appendix for proof):

Theorem 1: Upper bound $B_0(n_0)$ for the maximal α -quasi-clique community $C^*(n_0)$ of n_0 .

Given a node n_0 with degree $d(n_0)$, the size of the maximal α -quasi-clique community n_0 can belong to $|C^*(n_0)|$ is bounded by $B_0(n_0)$, given by:

$$B_0(n_0) = \left\lceil \frac{d(n_0)}{\alpha} \right\rceil \quad (3)$$

The symbols $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor and the ceiling function of a real number x respectively.

Then, $B_0(n_0)$ constitutes an upper bound for the optimal solution of Problem 1:

$$C^*(n_0) \leq B_0(n_0) \quad (4)$$

The bound $B_0(n_0)$ is satisfied if and only if all the neighbors of n_0 belong to its local community. Therefore, it might be loose specially for nodes that have a high degree.

Now, consider that the optimal solution contains at least two nodes (for any α) provided that n_0 has at least one neighbor. That is, at least one of the neighbors of n_0 belongs to $C^*(n_0)$ and any node in the community also must verify the rule of an α -alpha-quasi-clique in Eq. (1). Then, we can deduce a tighter bound for the optimal solution denoted $B_1(\cdot)$:

$$B_1(n_0) = \min \left(\max_{n \in \Gamma(n_0)} B_0(n), B_0(n_0) \right) \quad (5)$$

Then, the following inequalities hold:

$$C^*(n_0) \leq B_1(n_0) \leq B_0(n_0) \quad (6)$$

Now, consider the node 0 in Fig. 1 and $\alpha=0.5$. We are interested in an upper bound of node 0 and we obtain $B_0(0)=8$ and $B_1(0)=6$, then, $C^*(0) \leq 6$. However, this bound is reached if and only if exactly three of the neighbors of node 0 are in $C^*(0)$. By calculating the upper bound B_1 for neighbors 3 and 4, we notice they cannot belong to a community of size 6. Indeed, their degrees are worth 1 then $B_0(3)=B_0(4)=2$. That is, nodes 3 and 4 maximal quasi-clique communities are of size 2. We can deduce that the upper bound $B_1(0)=6$ cannot be achieved. Then, there are only two possibilities:

- Either nodes 3 and 4 belong to $C^*(n_0)$ and in that case $C^*(0) = 2$;
- Either nodes 3 and 4 do not belong to $C^*(n_0)$ and in that case the remaining degree of 0 is 2 and as a result $C^*(0) \leq 4$ the community size of the optimal solution is at most 4.

In conclusion, the optimal solution $C^*(0)$ is bounded by 4, which is exactly the size of $C^*(0)=\{0,1,2,5\}$.

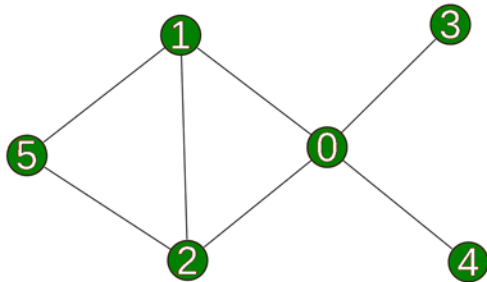


Fig. 1. The node 0 cannot belong to a community of size 6 as it needs at least 3 connections to belong to a community of this size and 2 of its neighbors, nodes 3 and 4, can belong to a community of size at most 2.

All these explanations can be summarized in an algorithm to calculate a tighter bound which will be denoted B . One can easily realize that:

$$C^*(n_0) \leq B(n_0) \leq B_1(n_0) \leq B_0(n_0)$$

In Algorithm 1 we can see all the steps to calculate the bound B .

Experimentally we found that the bound B is useful especially for nodes that have a high degree. For example, consider the node that represents the instructor in the Zachary Karate Club network [25] (node 1). We obtain $B_0(1)=32$, $B_1(1)=20$ and $B(1)=10$ whereas the optimal solution is $C^*(1)=8$.

Algorithm 1: Upper Bound B for the optimal community size $C^*(n_0)$ of a node n_0

Require: A node n_0 and a parameter α

Ensure: An upper bound B for the optimal community size $C^*(n_0)$

1: Set $B = B_0(n_0)$, the remaining degree $r=d(n_0)$ and its neighborhood $\Gamma(n_0)$.

2: **For** every node n in $\Gamma(n_0)$

Calculate $B_0(n)$

end For

3: Sort the values $B_0(n)$ by ascending order and calculate a histogram of each value. Denote the smallest value $B_0(n)[1]$, its frequency $ff[i]$ in the histogram and set $i=1$.

4: **while** $B > B_0(n)[i]$

Update $r = r - ff[i]$

Update $B = \left\lceil \frac{r}{\alpha} \right\rceil$

increase by 1 the value of i

end while

return B

4. Experimental Results

In this section, we calculated the upper bound B for all the nodes of the 3 real networks for different values of α . Then, we compared the upper bound to the community sizes experimentally obtained by the RNN algorithm. The Fig. 2 shows the histogram of the difference between the upper bound B and the results of the RNN method for the following networks:

- "The Zachary Karate Club network" (karate) [25] [Zachary 1977], 34 nodes and 78 edges.
- "Books about US politics" (polbooks) [26] [Krebs 2004], 104 nodes and 441 edges.
- "Political blogosphere" (polblogs) [27] [Adamic and Glance, 2005], 1490 nodes and 16715 edges.

The Fig. 2 shows that the difference is concentrated in small values for the 3 networks because the frequency is decreasing. Especially for the *karate* dataset, the bound was very close to the optimal solution for all but two nodes. The smallest is the difference, the tighter is the bound. A value of the difference equal to zero means that the bound is equal to the optimal solution as it can be reached by the RNN algorithm. For the three real networks, we obtained small differences for most of the nodes. For the bound to be reached at least one node in the community must be saturated, that means the number of its internal connections in the community must be equal to its degree. That is not always the case, and in that case the bound might not be tight enough.

5. Conclusions and Perspectives

In this paper, we tackled what we called the maximal α -quasi-clique local community problem. This problem being NP-hard, we proposed an upper bound on the optimal solution. We started by

proposing a loose bound based on the degree of the starting node and then we deduce a tighter bound. Experimentally in real networks, we saw that in most cases the bound was equal to the optimal solution.

The proposed bound is satisfied if and only if at least one node in the local community is saturated, so it may be loose in case of a complete clique. This drawback can lead to perspectives on improving the

bound. We can also consider structural properties of the optimal solution, one can mention, for instance, the fact that if $\alpha > 0.5$ the optimal solution contains only nodes of the first and the second neighborhood of the starting node as it was demonstrated in [10]. This must simplify the calculation of the node as well as give as ideas for a new algorithm.

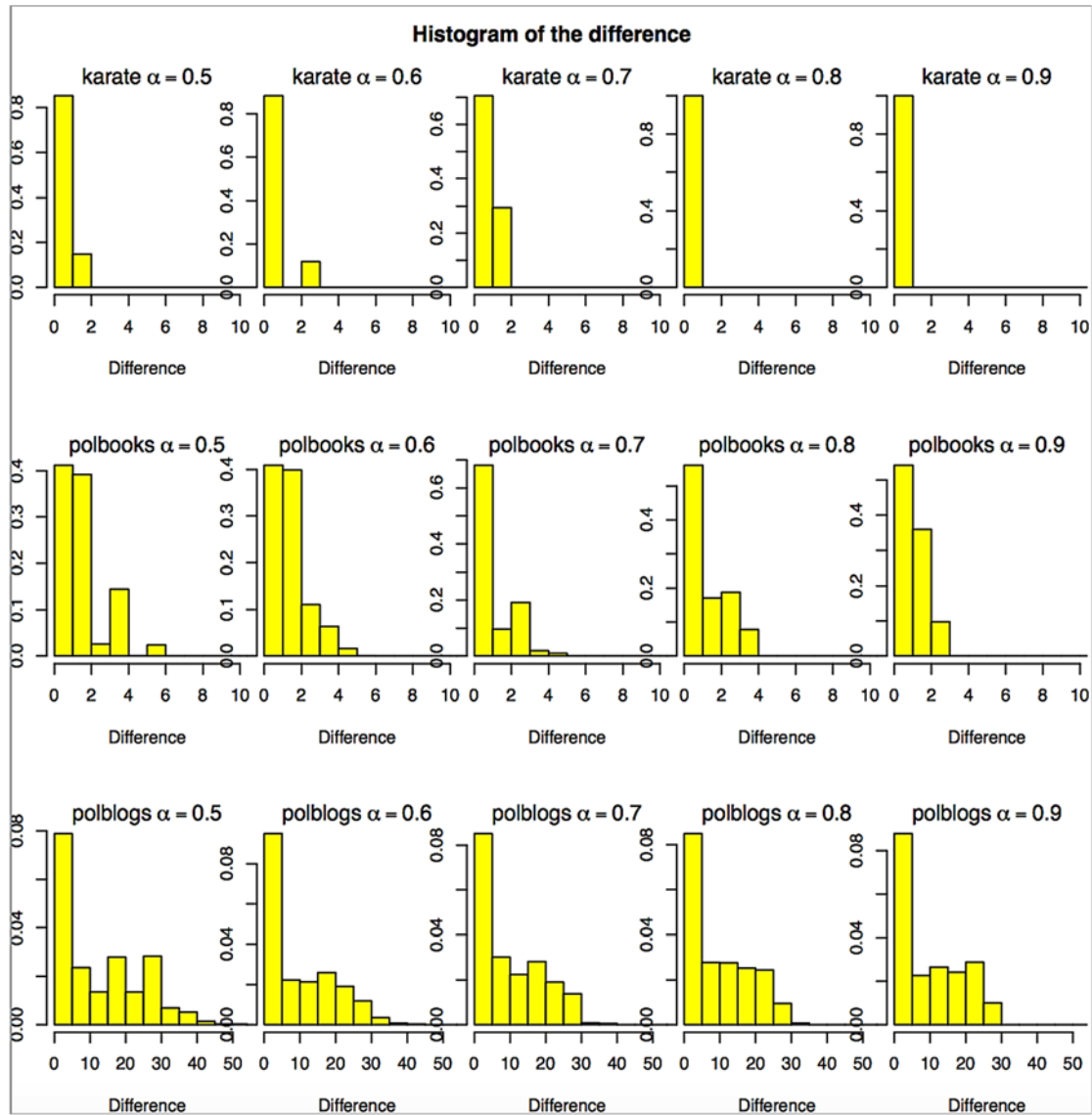


Fig. 2. Histogram of the difference between the upper bound B and the results of RNN algorithm for real networks.

Acknowledgements

The author would like to thank the anonymous reviewers for their constructive remarks and comments.

References

- [1]. Fortunato S., Community detection in graphs, *Physics Reports*, Vol. 486, 2010, pp. 75–174.
- [2]. Newman M., Girvan M., Finding and evaluating community structure in networks, *Physical Review E*, Vol. 69, Issue 2, 2004.
- [3]. Fortunato S., Barthelemy M., Resolution limit in community detection, in *Proceedings of the National Academy of Sciences of the United States of America*, 2006.
- [4]. Bomze I. M., Budinich M., Pardalos P. M., Pelillo M., The maximum clique problem, in *Handbook of Combinatorial Optimization*, Du D.-Z. and Pardalos P. M. (Eds.), *Kluwer Academic Publishers*, 1999, pp. 1–74.

- [5]. Lee V. E., Ruan N., Jin R., Aggarwal C. C., A survey of algorithms for dense subgraph discovery, in Managing and Mining Graph Data, C.C. Aggarwal, H. Wang (Eds.), *Advances in Database Systems, Springer US*, Vol. 40, 2010, pp. 303-336.
- [6]. Pattillo J., Youssef N., Butenko S., On clique relaxation models in network analysis, *European Journal of Operational Research*, Vol. 226, Issue 1, 2013, pp. 9-18.
- [7]. Wu Q., Hao J. K., A review on algorithms for maximum clique problems, *European Journal of Operational Research*, Vol. 242, Issue 3, 2015, pp. 693-709.
- [8]. Karp R. M., Reducibility among combinatorial problems, in Complexity of Computer Computations, R. E. Miller, J. W. Thatcher (Eds.), The IBM Research Symposium Series, *Plenum Press*, New York, 1972, pp. 85-103.
- [9]. Asahiro Y., Hassin R., Iwama K., Complexity of finding dense subgraphs, *Discrete Applied Mathematics*, Vol. 121, Issue 1-3, 2002, pp. 15-26.
- [10]. Conde-Céspedes P., Ngonmang B., Viennet E., An efficient method for mining the maximal α -quasi-clique- community of a given node in complex networks, *Social Network Analysis and Mining*, Vol. 8, Issue 1, 2018, pp. 1-20.
- [11]. Conde-Céspedes P., Ngonmang B., Viennet E., Approximation of the maximal α -consensus local community detection problem in complex networks, in *Proceedings of the IEEE 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Bangkok, Thailand, 2015.
- [12]. Abello J., Resende M. G. C., Sudarsky S., Massive quasiclique detection, in *Proceedings of the 5th Latin American Symposium on Theoretical Informatics, (LATIN'02)*, Springer-Verlag, London, UK, 2002, pp. 598-612.
- [13]. Chen J., Saad Y., Dense subgraph extraction with application to community detection, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, Issue 7, 2012, pp. 1216-1230.
- [14]. Pattillo J., Veremyev A., Butenko S., Boginski V., On the maximum quasi-clique problem, *Discrete Applied Mathematics*, Vol. 161, 2013, pp. 244-257.
- [15]. Tsourakakis C., Bonchi F., Gionis A., Gullo F., Tsiarli M., Denser than the densest subgraph: Extracting optimal quasi-cliques with quality guarantees, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*, ACM, New York, NY, USA, 2013, pp. 104-112.
- [16]. Brunato M., Hoos H. H., Battiti R., On effectively finding maximal quasi-cliques in graphs, in *LION, Lecture Notes in Computer Science*, V. Maniezzo, R. Battiti, J. P. Watson (Eds.), *Springer*, Vol. 5313, 2007, pp. 41-55.
- [17]. Liu G., Wong L., Effective pruning techniques for mining quasi-cliques, in Machine Learning and Knowledge Discovery in Databases, in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2008, Lecture Notes in Computer Science*, W. Daelemans, B. Goethals, K. Morik (Eds.), *Springer*, Berlin, Heidelberg, Vol. 5212, 2008, pp. 33-49.
- [18]. Bagrow J. P., Evaluating local community methods in networks, *Journal of Statistical Mechanics*, 2008, p. 05001.
- [19]. Clauset A., Finding local community structure in networks, *Physical Review*, Vol. 72, 2005, p. 026132.
- [20]. Luo F., Wang J. Z., Promislow E., Exploring local community structure in large networks, in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WT'06)*, 2006, pp. 233-239.
- [21]. Conde-Céspedes P., Marcotorchino J., Viennet E., Comparison of linear modularization criteria using the relational formalism, an approach to easily identify resolution limit, *Revue des Nouvelles Technologies de l'Information Extraction et Gestion des Connaissances, RNTI-E-28*, 2015, pp. 203-214.
- [22]. Conde-Céspedes P., Marcotorchino J., Viennet E., Comparison of linear modularization criteria using the relational formalism, an approach to easily identify resolution limit (extensions), *Advances in Knowledge Discovery and Management*, Vol. 6, (AKDM-6), 2017, pp. 101-120.
- [23]. Campigotto R., Conde-Céspedes P., Guillaume J., A generalized and adaptive method for community detection, *CoRR* abs/1406.2518, 2014. <http://arxiv.org/abs/1406.2518>.
- [24]. Conde-Céspedes P., An Upper Bound for the Maximal α -quasi-clique community Problem, in *Proceedings of the 1st International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAP'19)*, Barcelona, Spain, 2019, pp. 148-149.
- [25]. Zachary W. W., An information flow model for conflict and fission in small groups, *Journal of Anthropological Research*, Vol. 33, Issue 4, 1977, pp. 452-473.
- [26]. Krebs V., Books about US politics, 2004. <http://www.orgnet.com/>
- [27]. Adamic L. A., Glance N., The political blogosphere and the 2004 u.s. election: Divided They Blog, in *Proceedings of the WWW2005 Workshop on the Weblogging Ecosystem*, ACM, NY, USA, 2005, pp. 36-43.

Appendix

Proof of Theorem 1

n_0 must respect the rule of an α -alpha-quasi-clique (see Eq. (1)) as it is part of $C^*(n_0)$. Let us denote $d^{in}(n_0)$ the number of internal connections of n_0 in C^* . That is $d^{in}(n_0) = |E(n_0) \cap C^*(n_0)|$. Then, we have:

$$d^{in}(n_0) > \alpha(|C^*(n_0)| - 1).$$

This inequality implies:

$$\frac{d^{in}(n_0)}{\alpha} + 1 > |C^*(n_0)|$$

The left-hand side of this last expression gives an upper bound for $|C^*(n_0)|$, which we denote $B_0(n_0)$. This inequality is equivalent to:

$$B_0(n_0) = \begin{cases} \frac{d^{in}(n_0)}{\alpha} & \text{if } \left(\frac{d^{in}(n_0)}{\alpha} + 1\right) \text{ is integer} \\ \left\lfloor \frac{d^{in}(n_0)}{\alpha} \right\rfloor + 1 & \text{otherwise} \end{cases}$$

which is equivalent to:

$$B_0(n_0) = \left\lceil \frac{d^{in}(n_0)}{\alpha} \right\rceil$$

$$B_0(n_0) = \left\lceil \frac{d(n_0)}{\alpha} \right\rceil$$

qed.

Since $d^{in}(n_0)$ is upper bounded by $d(n_0)$ we obtain:



Published by International Frequency Sensor Association (IFSA) Publishing, S. L., 2019
(<http://www.sensorsportal.com>).